UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

DISSERTAÇÃO

Sumarização Automática de Textos Jurídicos: apoio tecnológico no enfrentamento à sobrecarga de informação jurídica

Bruno de Menezes Perdigão

2023

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

Sumarização Automática de Textos Jurídicos: apoio das Humanidades Digitais ao enfrentamento à sobrecarga de informação jurídica

Bruno de Menezes Perdigão

Dissertação submetida como requisito para a obtenção do grau de Mestre em Humanidades Digitais, no Programa de Pós-Graduação em Humanidades Digitais.

Orientador: Prof. Dr. Rodrigo de Souza Tavares

Nova Iguaçu/RJ
novembro/2023

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
com os dados fornecidos pelo(a) autor(a)

Perdigão, Bruno de Menezes , 1978-

P433s SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS JURÍDICOS: APOIO DAS HUMANIDADES DIGITAIS AO ENFRENTAMENTO À SOBRECARGA

DE INFORMAÇÃO JURÍDICA / Bruno de Menezes Perdigão. Rio de Janeiro, 2023.

143 f.: il.

Orientador: Rodrigo de Souza Tavares .

Dissertação(Mestrado). -- Universidade Federal Rural do Rio de Janeiro, Universidade Federal Rural do Rio de Janeiro, Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais/PPGIHD, 2023.

1. Humanidades Digitais. 2. Mineração de Texto. 3. Direito. 4. Sumarização de Texto Jurídico. 5. Sobrecarga de informação. I. Tavares , Rodrigo de Souza , 1978-, orient. II Universidade Federal Rural do Rio de Janeiro. Universidade Federal Rural do Rio de Janeiro, Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais/PPGIHD III. Título.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Universidade Federal Rural do Rio de Janeiro

PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

ATA Nº 24

Aos 12 dias do mês de DEZEMBRO do ano de dois mil e vinte e três, às 10:00 horas, instalou-se a banca examinadora de mestrado do aluno BRUNO DE MENEZES PERDIGÃO, sob a orientação do professor doutor RODRIGO DE SOUZA TAVARES, que foi composta pelos professores/pesquisadores: LEANDRO GUIMARÃES MARQUES ALVIM e LUÍS CLÁUDIO MARTINS DE ARAÚJO. Deu-se inicio às 10 horas 05 minutos, e teve a duração de 1 hora.

O Candidato, após avaliado pela banca examinadora obteve o resultado:

(x) APROVADO, devendo entregar a versão da DISSERTAÇÃO até 60 dias à coordenação do seu curso.
( ) APROVADO (a) COM RESSALVA, devendo o (a) Candidato (a) satisfazer, no prazo estipulado pela banca, as exigências constantes da Folha de Modificações de Qualificação da Dissertação de Mestrado anexa à presente ata.
( ) REPROVADO (a).

Parecer da banca:

Após deliberação, a banca decidiu pela aprovação da dissertação "Sumarização Automática de Textos Jurídicos: apoio tecnológico no enfrentamento à sobrecarga de informação jurídica", defendida pelo aluno Bruno de Menezes Perdigão. A banca recomendando apenas revisão final do texto antes do seu depósito.

Nova Iguaçu, 12 de dezembro de 2023.

Dr. LUIS CLAUDIO MARTINS DE ARAUJO, UCAM
Examinador Externo à Instituição

Dr. LEANDRO GUIMARAES MARQUES ALVIM, UFRRJ
Examinador Interno

Dr. RODRIGO DE SOUZA TAVARES, UFRRJ
Presidente

BRUNO DE MENEZES PERDIGAO
Mestrando

OBSERVAÇÃO: Esta ata é documento administrativo de uso exclusivo da Pró-Reitoria de Pesquisa e Pós-Graduação e NÃO pode ser utilizada a título de comprovação de Grau pelo candidato, que deve seguir o trâmite institucional para emissão de Diploma, Histórico Escolar e demais declarações.

Universidade Federal Rural do Rio de Janeiro

PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

FOLHA DE CORREÇÕES

ATA Nº 24

Autor: **BRUNO DE MENEZES PERDIGAO**

Título: **Sumarização Automática de Texto Jurídico na Experiência do LegalSumm**

A condição para a aprovação do (a) Candidato (a) é cumprir as seguintes exigências:

O prazo para o cumprimento das exigências é de ____ dias, sendo responsável(eis) pela verificação o(s)

Prof. LUIS CLAUDIO MARTINS DE ARAUJO — Examinador Externo à Instituição

Prof. LEANDRO GUIMARAES MARQUES ALVIM — Examinador Interno

Prof. RODRIGO DE SOUZA TAVARES — Presidente

Atesto que as alterações exigidas ( ) foram / ( ) não foram cumpridas, sendo o candidato considerado (a):
( ) Aprovado (a) / ( ) Reprovado (a).

Seropédica, ___ de _____ de _____.

Prof. RODRIGO DE SOUZA TAVARES
Orientador

BRUNO DE MENEZES PERDIGAO
Mestrando

ATA DE DEFESA DE TESE Nº 378/2023 - PPGIHD (11.39.00.16)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 16/03/2024 13:25)
LEANDRO GUIMARAES MARQUES ALVIM
PROFESSOR DO MAGISTERIO SUPERIOR
DeptCC/IM (12.28.01.00.00.83)
Matrícula: ####008#2

(Assinado digitalmente em 29/03/2024 17:15)
RODRIGO DE SOUZA TAVARES
PROFESSOR DO MAGISTERIO SUPERIOR
DeptCJ/IM (12.28.01.00.00.85)
Matrícula: ####922#0

(Assinado digitalmente em 27/03/2024 10:43)
BRUNO DE MENEZES PERDIGAO
DISCENTE
Matrícula: 2021#######2

(Assinado digitalmente em 27/03/2024 20:54)
LUIS CLAUDIO MARTINS DE ARAUJO
ASSINANTE EXTERNO
CPF: ###.###.617-##

*Visualize o documento original em https://sipac.ufrrj.br/documentos/ informando seu número: **378**, ano: **2023**, tipo: **ATA DE DEFESA DE TESE**, data de emissão: **16/03/2024** e o código de verificação: **d73fbae6bf***

RESUMO

PERDIGÃO, B. M. **Sumarização Automática de Textos Jurídicos: apoio das Humanidades Digitais ao enfrentamento à sobrecarga de informação jurídica**. Dissertação (Mestre em Humanidades Digitais). 143 f. Universidade Federal Rural do Rio de Janeiro, Instituto Multidisciplinar, Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais/PPGIHD, Nova Iguaçu, RJ, 2023.

A grande integração representada pela globalização foi capaz de elevar o nível de informação disponível, gerando uma grande massa de dados, a *big data*, com enorme oportunidade para a pesquisa. Todo esse movimento também provocou sobrecarga da informação na área jurídica, por meio da inserção dos novos conceitos e preocupações, com contaminação da relação entre trabalho e tempo disponível, que inclui não apenas o descanso, mas as oportunidades para aperfeiçoamento profissional. No campo das Humanidades Digitais, o Direito utiliza as ferramentas tecnológicas da Computação para solução de fenômeno social complexo. A análise do grande manancial de dados jurídicos, diante da necessidade de apresentação de razões escritas como forma de justificar decisões judiciais, administrativas ou legislativas, no ambiente do Estado de Direito, lança luz sobre ferramentas tecnológicas voltadas para a mineração dos textos, em especial, a sumarização automática de texto, por meio da utilização do Processamento de Linguagem Natural, que possibilita o resumo de um ou mais textos, com geração de um sumário, com informações condensadas, inclusive na língua portuguesa. Essa dissertação investiga a sumarização automática de texto jurídico como ferramenta tecnológica de apoio aos operadores do Direito no enfrentamento da sobrecarga de informação jurídica, no campo das Humanidades Digitais, por meio da experiência com sumarizadores automáticos na língua portuguesa. O estudo discorre sobre a cultura jurídica e as mudanças necessárias para acolhimento das ferramentas tecnológicas, o fenômeno da *big data* e o processo de descoberta de conhecimento em base de dados, com ênfase na mineração de textos. Aborda a sumarização automática de textos por meio da comparação de trabalhos com sumarizadores automáticos em língua portuguesa, especialmente, o LegalSumm, sumarizador de texto jurídico. Apresenta a experiência com sumarização de texto jurídico por meio de ferramenta preordenada, de acesso online e gratuito, bem como via Python, linguagem de programação modular, comparando o desempenho do ponto de vista da qualidade em sentido amplo dos sumários gerados.

Palavras-chave: Humanidades Digitais; Direito; Mineração de Dados; Mineração de Texto; Sumarização de Texto.

ABSTRACT

PERDIGÃO, B. M. **Automatic Summarization of Legal Texts: support from Digital Humanities to combat the overload of legal information**. Dissertation (Master in Digital Humanities). 143 p. Federal Rural University of Rio de Janeiro, Multidisciplinary Institute, Interdisciplinary Postgraduate Program in Digital Humanities/PPGIHD, Nova Iguaçu, RJ, 2023.

The great integration represented by globalization was able to increase the level of available information, generating a large mass of data, big data, with enormous opportunities for research. This entire movement also caused information overload in the legal area, through the insertion of new concepts and concerns, contaminating the relationship between work and available time, which includes not only rest, but opportunities for professional improvement. In the field of Digital Humanities, Law uses the technological tools of Computing to solve complex social phenomena. The analysis of the large source of legal data, given the need to present written reasons as a way of justifying judicial, administrative or legislative decisions, in the rule of law environment, sheds light on technological tools aimed at mining texts, in particular, automatic text summarization, through the use of Natural Language Processing, which makes it possible to summarize one or more texts, generating a summary, with condensed information, including in Portuguese. This dissertation investigates the automatic summarization of legal text as a technological tool to support legal operators in dealing with the overload of legal information, in the field of Digital Humanities, through experience with automatic summarizers in the Portuguese language. The study discusses legal culture and the changes necessary to embrace technological tools, the phenomenon of big data and the process of discovering knowledge in databases, with an emphasis on text mining. It addresses the automatic summarization of texts by comparing works with automatic summarizers in Portuguese, especially LegalSumm, a legal text summarizer. It presents the experience with legal text summarization through a pre-ordered tool, with free online access, as well as via Python, a modular programming language, comparing the performance from the point of view of quality in the broadest sense of the summaries generated.

Keywords: Digital Humanities; Law; Data Mining; Text Mining; Text Summarization.

LISTA DE ABREVIATURAS

ADA	Análise Discursiva Automática
AMR	*Abstract Meaning Representation*
BDTD	Biblioteca Digital de Teses e Dissertações
BERT	*Bidirectional Encoder Representations*
Bert	*Bidirectional Encoder Representations*
Capes	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
ChatGPT	*Generative Pre-trained Transformer*
CST	*Cross-document Structure Theory*
DIKW	*Data Information Knowledge Wisdom*
DMSumm	*Discourse Modeling Summarizer*
Enem	Exame Nacional do Ensino Médio
IA	Inteligência Artificial
JDP	Jornada da Descrição da Língua Portuguesa
KDD	*Knowledge Discovery in Databases*
kNNSumm	*K-nearest neighboors Summarizer*
NILC	Núcleo Interinstitucional de Linguística Computacional
NPL	*Natural Language Processing*
PDF	*Portable Document Format*
PLN	Processamento da Linguagem Natural
RE	Recurso extraordinário
REsp	Recurso especial
RIT	Reconhecimento de Implicação Textual
Rouge	*Recall-Oriented Understudy for Gisting Evaluation*
RST	Rhetorical Structure Theory
SA	sumarização automática
SciELO	*Scientific Eletronic Library Online*
SOM	*Self-Organizing Map*
STF	Supremo Tribunal Federal
STJ	Superior Tribunal de Justiça
TAC	*Text Analysis Conference*
TF-IDF	*Term Frequency – Inverse Document Frequency*
TF-ISF	*Term Frequency-Inverse Sentence Frequency*
TF-ISF-Summ	*Term Frequency Inverse Sentence Frequency*
UIT	União Internacional de Telecomunicações
UNL	*Universal Networking Language*
USP	Universidade de São Paulo

LISTA DE QUADROS E FIGURAS

SUMÁRIO

INTRODUÇÃO

O tempo é um elemento constante em toda existência humana. Ele serve de medida daquilo que já passou, do que está acontecendo ou ainda do que será. Interage com as relações, às vezes acelerando, em outras, diminuindo o ritmo. Delimita os jogos e competições, as agendas, as circunstâncias, o início e o fim.

Ao longo do caminho, talvez se torne um adversário contra o qual a vida lute incessantemente, porque o tempo de existência se desenvolve em direção ao fim. Em alguns pontos, é amigo, com impacto nas lembranças prazerosas, transformando o afoito em experiente, o inábil em virtuoso, a tristeza em alegria, a ausência em descoberta e cura.

Ocorre que, de tão presente, o tempo acaba caindo no esquecimento, no cotidiano, passando a ter sua existência simplesmente ignorada. Mas logo é relembrado, no momento de morte, quando se reflete sobre o tempo vivido e o quanto pode ter sido perdido. Assim, o tempo instrumentalizado deve ser objeto de profunda reflexão (Hartog, 2006), sobremaneira, a influência que exerce nas relações humanas.

O mundo tem experimentado um movimento cada vez maior de mudanças, já não é novidade, sendo perceptíveis alterações nos mais variados ramos de atuação humana, especialmente, a partir do século XX e o início do século XXI. A grande integração representada pela globalização foi impulsionada pelo desenvolvimento tecnológico, repaginando os antigos conhecimentos, potencializando os resultados, contribuindo para o aparecimento de novas fontes de atuação, com influência no tempo disponível (Saber, 2006).

Houve mudança de perspectiva na sociedade, conforme advertido por Bauman, passando da modernidade sólida, caracterizada pela rigidez e solidificação das relações humanas e institucionais, com construção de longo prazo, para modernidade líquida, em que as relações econômicas sobrepõem as relações humanas e institucionais, com a substituição da lógica da moral pela lógica do consumo, dotada de superficialidade suficiente para desfazer as conexões humanas, e de agilidade, para acompanhar os pensamentos do momento (Bauman, 2001).

Uma nova era, batizada de Revolução 4.0 ou Quarta Revolução Industrial, por Klaus Schwab, Fundador do Fórum Econômico Mundial, foi atingida. A sociedade será submetida a um processo de mudança em escala, tamanho e densidade jamais verificados, com alteração no

modo de viver, pensar e debater (Schwab, 2016), com diversidade tecnológica, sistemas inteligentes conectados e interação dos domínios físicos, digitais e biológicos.

O notável desenvolvimento científico, com o surgimento de instrumentos tecnológicos, métodos e processos automatizados, a partir dos ensaios de Turing (meados de 1930) e Neumann (década de 1940), forneceu grande oxigênio para o fogo das mudanças, em quantidade e qualidade, com franca expansão do acesso ao conhecimento (Eifler Saraiva; De Lima Argimon, 2007).

O diálogo de diversas fontes de conhecimento, de caráter interdisciplinar, partindo da Filosofia para Matemática, com passagens pela Engenharia e Economia, com contribuição das ciências humanas e sociais aplicadas, propiciou abertura para discussão dos novos campos (Saracevic, 1996).

O nosso mundo, outrora físico, foi se tornando cada vez mais digital, e a capacidade, qualidade, diversidade do que vai sendo produzido e armazenado criou uma massa de dados, de origem pública ou privada, chamada de *big data*, de possível análise, com grande oportunidade para a pesquisa e melhoramento dos processos.

Todo esse movimento também impactou no processo cultural, com reflexos positivos, como melhoramento dos mecanismos produtivos humanos, e negativos, com contaminação da relação trabalho e tempo disponível, que inclui não apenas o descanso, mas oportunidade de aperfeiçoamento da formação profissional.

O fluxo gerado pela utilização das novas tecnologias vai, aos poucos, sendo inserido nos conceitos e preocupações jurídicas, com ingerência no desempenho das atividades dos operadores do Direito: advogados privados e públicos, membros do Poder Judiciário, membros do Ministério Público, membros das Defensorias Públicas, membros de forças policiais e demais atores que utilizam o Direito como profissão. O tempo de trabalho vai se tornando escasso enquanto o volume de dados tem crescimento contínuo, seja no setor público ou privado.

A partir da percepção do aumento vertiginoso de produção de conteúdo e da dificuldade imposta aos profissionais na tomada de decisão, diante do manancial de dados disponíveis, alertando sobre a necessidade de adaptação aos novos tempos, em um processo de aprendizado contínuo e sistémico, Alvin Toffler foi um dos primeiros a notar o impacto do problema do volume de dados sobre as relações humanas, tendo criado a expressão sobrecarga de informação (Fernandes; Meirinhos, 2021).

As novas fronteiras impõem um olhar crítico por parte dos pesquisadores, tanto no que toca aos instrumentos quanto aos dados utilizados, a necessidade de discutir as posturas éticas, além de incorporar aspectos de sustentabilidade ao processo científico, pois as abordagens metodológicas potencializam resultados quando estão dispostas a entender os processos e aperfeiçoar as ferramentas.

O Poder Judiciário do Brasil, um dos três Poderes da República, independentes e harmônicos, consagrados na Constituição da República Federativa do Brasil de 1988, com função de intérprete máximo da lei, é composto por uma enorme estrutura, com cinco segmentos, vale dizer, a Justiça Estadual, a Justiça Federal, a Justiça do Trabalho, a Justiça Eleitoral e a Justiça Militar. No topo do sistema está o Supremo Tribunal Federal e, logo abaixo, os tribunais superiores: o Superior Tribunal de Justiça, o Superior Tribunal Militar, o Tribunal Superior Eleitoral e o Tribunal Superior do Trabalho. Toda essa estrutura demanda custos elevados.

Segundo informações do Conselho Nacional de Justiça (2021), constantes do relatório Justiça em Números 2021, o Poder Judiciário chegou ao fim do ano de 2020 com um estoque de 75,4 milhões de processos aguardando uma solução definitiva, sendo correto afirmar que ingressaram, no mesmo ano, 25,8 milhões de feitos novos com a baixa de 27,9 milhões de processos, ficando demonstrada a dificuldade na prestação da tutela jurisdicional no tempo razoável, e do volume crescente de dados, materializados principalmente pela adoção do processo eletrônico, uma vez que é possível demandar a partir do acesso à *internet*.

Em complemento, a crescente disponibilidade de dados vai gerando um problema para os operadores do Direito, uma vez que necessitam, para o bom desempenho de suas funções, manter um nível renovado de informações relevantes, de teorias e jurisprudência, passando pela leitura de textos de estrutura longa e complexa, pertencente ao universo jurídico.

Diante das premissas traçadas, o Direito deve participar do momento histórico, compartilhando dos recursos oriundos da Computação, possibilitando o alcance de novos patamares do conhecimento em atuação interdisciplinar por meio de ferramentas tecnológicas, viabilizando a investigação de fenômenos sociais complexos. Nesse passo, os recursos tecnológicos passam a auxiliar na manipulação de grande massa de dados, o que seria impossível pela atuação singular de cada ser humano. Aqui está o campo das Humanidades Digitais.

Diversas iniciativas já servem de inspiração interdisciplinar. Por exemplo, por meio da mineração de dados de entidade pública foi possível identificar e classificar o perfil de empresas com maior potencial de se comportarem de maneira irregular, em relação ao trato com os impostos estaduais (Nascimento *et al.*, 2018).

Em estudo voltado para o combate de práticas inidôneas, instrumento de apoio das auditorias no controle interno da Administração Pública, a mineração de dados foi utilizada para mapear, em processos licitatórios, cartel de empresas (Silva *et al.*, 2020).

O tempo dispendido na construção de consultas e no treinamento de servidores, motivou o desenvolvimento de estudo voltado para Avaliação de Técnicas de Similaridade Textual na Uniformização de Jurisprudência, no âmbito do Superior Tribunal de Justiça (Gomes, 2020).

As tarefas rotineiras no âmbito da Procuradoria Geral do Distrito Federal (PGDF), que impunham grandes gastos em recursos humanos, para agrupar e classificar processos repetitivos, isto é, aqueles que exigem muitas ações manuais e menos ações de natureza intelectual (Souza, 2021), passam a contar com assessoria tecnológica.

Ao mesmo tempo, no âmbito nacional, surge a preocupação com a privacidade dos dados pessoais, em razão do crescimento exponencial das atividades em meio digital, dando azo ao aparecimento de um marco regulatório de proteção, materializado por meio da Lei 13.709/2018, Lei Geral de Proteção de Dados.

Portanto, a atualidade demonstra existência de diversos processos e iniciativas tecnológicas que oferecem mecanismos necessários e oportunidades aos operadores do Direito: a manipulação de tipos de processos, quantidade e fluxo de trabalho, compreensão de estruturas sintáticas ou semânticas das decisões, previsão de votos, argumentação, análise de sentimentos, identificação de entidades ou autores, jurimetria e outros.

A linha de pesquisa baseada na Mineração de Dados Digitais, na área de Análise Qualitativa e Quantitativa de Dinâmicas Sociais, oportunizou estudo da área de sumarização de texto em iniciativa das Humanidades Digitais como ferramenta tecnológica capazes de colaborar no enfrentamento do problema da sobrecarga de informações jurídicas.

A prática da simplificação encontra campo na própria essência do ser humano, que tem necessidade de reduzir os fatos para formação da memória e transmissão do conhecimento. Diuturnamente, são apresentadas sínteses de notícias, resumo de trabalhos de natureza científica, sinopse de romances, filmes e outros elementos apresentados de forma simplificada, incorporados na rotina.

Nessa esteira, a análise do grande manancial de dados jurídicos produzidos, mormente, pelo registro de informações na forma escrita, lança luz sobre o desenvolvimento e utilização de ferramentas tecnológicas voltadas para a mineração dos textos. Em especial, a sumarização automática de texto surge como uma alternativa viável porque possibilita a geração de um resumo a partir de um ou mais documentos, com a separação das informações desnecessárias, ocasionando ganho de tempo e produtividade pela redução do volume de informação disponível.

De outro lado a solidificação do sistema jurídico, que adota formas sacramentais, usualmente com a linguagem escrita, gera reflexão sobre a mudança cultural a ser empreendida, para acolher as ferramentas tecnológicas necessárias aos novos tempos.

Em adição, a estrutura dos documentos produzidos em processos judiciais, tanto no que tange à extensão, em razão do vasto número de páginas, quanto ao estilo da produção textual adequada para validade do ato jurídico, bem como a complexidade da língua portuguesa, lançam desafios adicionais a serem enfrentados na identificação das informações relevantes, espalhadas por diversas seções e formatos, de tormentosa compreensão e síntese, pelo ser humano ou pela máquina.

Em outro campo de debate, em um mundo cada vez mais marcado pelo aumento da produtividade e economia de custos, buscando mais lucros, a substituição gradual das atividades desempenhadas por seres humanos, devido ao progressivo avanço das tecnologias, impõe aprimoramento contínuo dos profissionais com aquisição de conhecimentos interdisciplinares.

O desenvolvimento tecnológico tem favorecido a criação de novos campos de conhecimento humano. Ao mesmo tempo que a tecnologia possibilitou o aumento de produtividade, pela eficácia e agilidade propiciadas, também gerou um crescimento massivo de dados, incentivado pelas iniciativas digitais, produzindo sobrecarga de informação.

As preocupações narradas também vão sendo inseridas no mundo jurídico, que precisa passar por um movimento de acomodação cultural para acolher os novos processos necessários para equalização da relação quantidade de trabalho e tempo disponível, para qualificação e lazer.

Como enfrentar o problema da sobrecarga de informação na área jurídica? Há iniciativas tecnológicas disponíveis para otimizar o trabalho dos operadores do Direito? As iniciativas são acessíveis aos operadores do Direito ou demandam uma formação técnica especializada? As ferramentas atendem aos padrões da língua portuguesa?

Assim, recursos oriundos da Computação passam a prestar relevante suporte aos profissionais de áreas humanas, para pesquisa e solução de dinâmicas sociais complexas, de natureza interdisciplinar, no campo das Humanidades Digitais, como é o caso do crescimento exponencial do volume de dados jurídicos, que não pode mais ser tratado de forma individualizada pelo ser humano.

Uma solução possível para enfrentar o problema do crescimento de dados jurídicos é a sumarização automática de texto, técnica do Processamento de Linguagem Natural (PLN), que poderá ser utilizada como ferramenta tecnológica no auxílio do trabalho dos operadores do Direito, com aumento de produtividade e ganho na relação trabalho e tempo disponível, ficando demonstrada a relevância do tema objeto da pesquisa.

O ato de sumarizar pertence ao comportamento humano, que não tem por tradição armazenar todas as informações de que toma conhecimento, mas apenas as mais importantes. Em complemento, a sumarização humana é uma atividade complexa, que contempla entendimento do conteúdo e formulação de síntese. Atentando para tais aspectos, de forma similar, foi construído um caminho para que o ato de sumarizar pudesse ser realizado de forma automática, a partir de um ou mais textos de entrada, por meio de mecanismos de Processamento de Linguagem Natural.

Não obstante, qual o estado da arte dos sumários gerados atualmente? Há experiência com sumarização automática especialmente desenvolvida para área jurídica? Qual o nível de qualidade dos sumários automáticos gerados?

A adaptação da técnica ao conteúdo da língua portuguesa, com poucos estudos e *corpora* disponibilizada, em comparação com as iniciativas de língua inglesa, somadas às características distintas dos textos jurídicos, de estrutura interna e vocabulário bastante técnico, com citações de doutrina, jurisprudência e arcabouço legal, com caracteres numéricos, impõem limitações adicionais ao desafio a ser enfrentado.

Importante pontuar, tendo em vista as frequentes atualizações científicas e consequente superação de tecnologias pelo advento de instrumentos mais modernos, que a pesquisa foi realizada entre março de 2021 e março de 2023.

Diante do exposto, o objetivo geral projetado é verificar a viabilidade da aplicação da sumarização automática de texto para atenuar ou resolver o problema da sobrecarga de informação jurídica, como ferramenta tecnológica de apoio aos operadores do Direito.

Como forma de identificar o problema e apresentar os marcos teóricos para compreensão e avaliação da qualidade em sentido amplo, é necessário debater a cultura jurídica e as mudanças para acolhimento das ferramentas tecnológicas; analisar o fenômeno da *big data* e o processo de Descoberta de Conhecimento em Base de Dados; o processo de mineração de textos; abordar a sumarização automática de textos por meio da comparação de trabalhos com sumarizadores automáticos em língua portuguesa; apresentando, ainda, experiência em sumarização de texto jurídico por meio de ferramenta preordenada, de acesso online e gratuito, bem como via Python, linguagem de programação modular.

A abordagem de pesquisa resultou na seleção de 191 textos direcionados para dois blocos de apoio: (A) Cultura, Informação, Direito, Humanidades Digitais, Descoberta de Dados e Mineração de Textos, com 113 trabalhos; e (B) Sumarização de Texto e Linguística, com 78 trabalhos. Os estudos foram triados, mediante leitura do conteúdo de sumários, índices e lista de referências, com identificação dos temas dos dois blocos de apoio supraexpostos. Em adição, foi possível realizar experimento com sumarizadores automáticos e julgar a qualidade dos sumários gerados por meio de avaliadores humanos, operadores do Direito.

Em razão dos limites impostos pela Pandemia da Covid-19, a maioria da pesquisa foi realizada em repositórios virtuais. Devem ser destacadas as seguintes plataformas de apoio à pesquisa: *Scientific Eletronic Library Online* (SciELO), Google Acadêmico, o Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), o Portal de Dissertações e Teses Defendidas na Universidade de São Paulo (USP), Biblioteca Digital de Teses e Dissertações (BDTD).

Esta dissertação é composta por sete blocos, sendo uma introdução, cinco capítulos propriamente ditos, uma conclusão e novas pesquisas.

Na introdução, por meio de breve contextualização, com perspectiva histórica da evolução da produção de conhecimento, são expostos os motivos ensejadores da pesquisa objeto da dissertação, sendo esclarecido o fenômeno do crescimento exponencial dos dados, com a geração de sobrecarga de informação, o impacto causado no mundo jurídico e as dificuldades enfrentadas pelos operadores do Direito, passando pela reflexão acerca de conhecimentos interdisciplinares necessários para solução de fenômenos sociais complexos, no âmbito das Humanidades Digitais, apresentando a sumarização de texto jurídico como uma alternativa para resolver ou mitigar o problema destacado.

No Capítulo 2 está inserida a compreensão do universo jurídico, com surgimento do Direito a partir do processo de acomodação cultural, bem como as preocupações atuais, que passam a ser inseridas na atuação profissional dos operadores do Direito: as novas tecnologias, o crescimento exponencial dos dados jurídicos, a sobrecarga de informação e a mudança cultural necessária para acolher ferramentas tecnológicas de auxílio; o tratamento de fenômenos sociais complexos, em atuação interdisciplinar. Em adição, é destacado o fenômeno da *big data*, massa de dados com características especiais, resultado do crescimento vertiginoso do conhecimento humano, que requerem auxílio de ferramentas tecnológicas para manipulação, impondo desafio no campo da pesquisa, em especial, o papel das Humanidades Digitais, apresentada sumarização automática de texto como alternativa a ser considerada para lidar com o problema da sobrecarga de informação jurídica.

No Capítulo 3 é analisado o processo científico de Descoberta do Conhecimento em Base de Dados, cujo núcleo é a mineração de dados, por meio de breve descrição do sistema, suas principais tarefas e métodos. Em complemento, é abordada a mineração de textos com interesse na sumarização automática, por meio da utilização do Processamento de Linguagem Natural, para extração do conhecimento a partir de informação na forma de texto, parte indissociável do Direito, diante da necessidade de apresentação de razões escritas como forma de justificar decisões judiciais, administrativas ou legislativas no ambiente do Estado de Direito. Em movimento preparatório para experiência de sumarização, é apresentada a importância do desenvolvimento de *corpus* de apoio para o Processamento de Linguagem Natural, com iniciativas em língua portuguesa, descrevendo, na última seção, o RulinBR, *corpus* formado a partir de decisões judiciais voltado para sumarização automática de textos jurídicos.

No Capítulo 4 está descrita a sumarização automática de textos, iniciando a trajetória via ato natural do ser humano até alcançar o espectro automático, com apresentação de revisão bibliográfica, com exposição dos métodos mais relevantes, com apontamentos acerca das vantagens e desvantagens de cada um, seguida pela taxonomia, com os principais sistemas de sumarização. Também são prestados esclarecimentos acerca dos sistemas de avaliação de desempenho e pontos necessários para aferição do sumário sob o prisma da qualidade em sentido amplo.

No Capítulo 5 são abordados trabalhos de desenvolvimento de sumarizadores automáticos de texto, em especial, pela utilização da língua portuguesa. Em adição, é realizada a comparação dos resultados das ferramentas sob o ponto de vista da qualidade em sentido

amplo. No ato derradeiro, como ato preparatório do experimento de sumarização, é analisado o LegalSumm, sumarizador de texto jurídico, voltado para decisões judiciais em língua portuguesa.

No Capítulo 6 é apresentado o experimento de sumarização automática de texto jurídico, a partir do resumo de três acórdãos: Acórdão de Julgamento em REsp 1842613-SP, perante o STJ; Acórdão de Julgamento em Recurso Extraordinário (RE 654833-AC), perante o STF; Acórdão de Julgamento em REsp 1.846.649-MA perante o STJ, via ferramenta de sumarização preconcebida e, ainda, por meio do Python, linguagem de programação modular, de modo a viabilizar respostas aos questionamentos acerca da existência de iniciativa com sumarização automática, especialmente desenvolvida para área jurídica, e da utilização pelos operadores do Direito das ferramentas computacionais vocacionadas para sumarização.

Por fim, são apresentadas as principais conclusões do estudo realizado e discutidas perspectivas de trabalho futuro.

1 SOBRECARGA DE INFORMAÇÃO E O PROCESSO DE TRANSFORMAÇÃO DA CULTURA JURÍDICA

No estágio atual, em proporções estratosféricas, as relações humanas passam a ser cada vez mais capturadas por sensores digitais, a partir dos mais variados dispositivos, devidamente conectados. Os dados coletados, os sistemas de vigilância virtuais, o uso indevido de informações, impõem debates sobre lesão aos direitos e garantias individuais: intimidade, liberdade de expressão, direitos patrimoniais e outros tantos.

Todo esse movimento é capaz de gerar uma grande quantidade de fatos, objeto de análise e extração de conhecimento para tomada de decisão, outrossim, aumentam a demanda de tempo de trabalho por parte dos profissionais em razão da sobrecarga de informação.

1.1 Expansão do conhecimento humano e a geração de grande volume de dados: processo de construção da memória

No campo das funções psíquicas, pela memória, há recordação de acontecimentos, consciência daquilo que foi experimentado, princípio de anterioridade, algo que remete ao passado. São cheiros, cores, lugares, sons, conversas, refeições e festividades, lembranças, prazerosas ou desconfortáveis, que auxiliam na construção do acervo humano.

Sob outro ângulo, o homem não é capaz de recordar todos os acontecimentos, mas apenas aqueles que possuem a cicatriz da associação, aquilo que marca a vida, o importante, o que se faz uso. É necessidade de conservação, de manutenção do que compõe o contexto individual ou coletivo, familiar ou social, sendo elemento de consolidação da própria identidade (Goff, 1990).

Inicialmente, a passagem das lembranças ocorre em ambiente familiar ou em pequenos grupos. Posteriormente, o movimento ganha força e conquista um caráter transcendente, de modo a percorrer toda a coletividade, indistintamente, pois as recordações ganham as cores da cidade, país ou nação, um corpo de grupo (Córdula; Nascimento, 2018).

Embora se possa notar um traço cultural distinto entre povos com linguagem escrita e não escrita (ágrafo), isso não impediu a transmissão do conhecimento ao longo do tempo. No exercício da memória, na organização do conhecimento na Antiguidade, os povos sem

linguagem escrita cultivavam suas tradições por meio de narrativas do contexto sócio-histórico, transmitidas de geração em geração, via oralidade (Fernandes *et al.*, 2021).

Os idosos, por exemplo, foram durante muito tempo fonte das memórias produzidas via oralidade, entendidos como homens memória ou memórias vivas, capazes da transmissão intergeracional de conhecimentos enraizados (Simson, 2003). Eram tutores, professores, mentores e conselheiros que ensinavam, influenciavam e formavam a memória. Na linha do desenvolvimento humano, com o surgimento da linguagem escrita e o volume das informações, as narrativas começaram a ficar gravadas, documentadas, incentivando o exercício do armazenamento dos acontecimentos. Era necessidade de guardar conhecimento para comprovar fatos.

No que tange ao processo de evolução da organização jurídica, paralelamente, os costumes foram fonte essencial das normas dos povos ágrafos, pois a forma tradicional de viver determinava a regra a ser seguida, assim, os julgamentos eram resultado da dinâmica prevalente na sociedade do momento. O aparecimento do Direito como forma escrita, por seu turno, foi indício da necessidade de comprovação dos fatos relevantes, organização e fiscalização da atuação nas relações públicas e privadas, em grupos sociais cada vez maiores, bem como estabelecimento de regras prévias e transparentes, de modo a limitar abusos dos dominantes, garantir diversidade de pensamento e pacificar os conflitos (Reis, 2019).

Na Idade Média, a difusão e consolidação do cristianismo fez parte do processo progressivo de construção da memória agregada ao contexto da fé, citando a memória dos mortos, além da prática de veneração aos santos e a incorporação de tradições nos calendários religiosos. Enquanto os clérigos desenvolviam a memória escrita, os fiéis exercitavam a oralidade das recordações com cânticos e missas. Posteriormente, com a utilização cada vez maior da forma escrita, o aparecimento da imprensa e dos tratados científicos condicionaram a memorização dos conhecimentos. Aparecem os grandes acervos de documentos, arquivos suporte da memória (Smolka, 2000).

No que toca a evolução do próprio conceito de memória, já no século XIX, é enfatizada a posição da memória coletiva, bibliotecas de saberes, uma vez que a memória individual demonstra incapacidade de retenção e assimilação de todo conhecimento, que deve ser construído por meio da interação social, em uma experiência vivida por meio da comunidade, formadora da identidade. Com o espetacular crescimento do conhecimento humano, no século XX, identifica-se a necessidade de memorizar o maior número de informações possíveis, com

a utilização de imagens ou símbolos como instrumento de síntese dos acontecimentos (Rodrigues, 2015).

Modernamente, as ferramentas tecnológicas e o compartilhamento global de informações geraram uma grande massa de dados, produto da expansão do conhecimento humano e do intercâmbio de relações: são cadastros de clientes, compras realizadas por meio de cartões, produtos e *sites* acessados, localização geográfica, jornais, bibliotecas, aplicativos. Dispositivos como computadores, celulares e aparelhos de televisão *smart*, mesmo veículos e geladeiras, devidamente conectados, são capazes de coletar dados que, armazenados, possibilitam estudo e extração de conhecimento.

No plano jurídico, o registro de dados na forma escrita, especialmente em texto, como modo de fundamentar as decisões judiciais, administrativas ou legislativas, no ambiente do Estado de Direito, foi se consolidando como parte indissociável da Ciência Jurídica, seguindo a concepção de que o governo de leis seria mais benéfico de que um governo de homens, sendo a lei escrita uma garantia contra o arbítrio (Dallari, 2007). É necessário detalhamento dos atos, um registro oficial como forma de proteção contra os abusos dos governantes.

No plano brasileiro, a forma de memorização dos atos exerceu influência na construção de preceitos da Constituição da República, ao impor ordem escrita e fundamentada da autoridade judiciária competente para determinação de prisão, conforme preceitua o art. 5º, LXI, além de impor, no art. 93, IX, a obrigatoriedade de fundamentação de todas as decisões judiciais, sob pena de nulidade.

O crescimento vertiginoso dos dados impactou nos conceitos e preocupações jurídicas, com ingerência no desempenho das atividades dos operadores do Direito, razão pela qual demandam estudo e reflexão, constituindo grande oportunidade de pesquisa no âmbito das Humanidades Digitais: são contratos celebrados, como compra e venda, mútuo, aluguel, prestações de serviço; milhares de decisões judiciais com alimentação dos diversos sistemas do Poder Judiciário do Brasil e do mundo; questões relacionadas à proteção dos dados e outras.

1.2 A *Big Data*

Atentando para o fenômeno da tomada de decisão, que requer interpretação humana, foi idealizada uma hierarquia de conceitos para análise dos dados, cunhada como Pirâmide do Conhecimento ou DIKW (*Data Information Knowledge Wisdom*), que percorre o fluxo de dados, informação, conhecimento e sabedoria, conforme estudado por Russel Ackoff, ainda na

década de 1980 (Queiroz; Fialho; Remor, 2017). A tomada de decisão, como produto da reflexão humana, é a materialização do que foi encontrado, apreendido, interpretado, é o poder de síntese sobre o todo processado.

O dado resulta primeiramente da simples observação de fatos, pois traduz a ideia de registro, evento, item elementar. Portanto, é uma espécie de representação, uma observação sem valor agregado, que se encontra na base da pirâmide hierárquica que leva ao conhecimento. No segundo ponto, já no nível intermediário da pirâmide hierárquica, a informação aparece como um dado que já foi cultivado, construído, com qualidade. É um dado com valor agregado, com contexto definido, com conexões e relações. Rumo ao topo da pirâmide idealizada está o conhecimento, que resulta no conjunto de informações adquiridas, um refinamento em relação ao estágio anterior, que possibilitará um processo de análise da informação, um estado de interpretação (compreensão). Por fim, a sabedoria representa o ápice do processo humano, em que há soma das diversas habilidades ou conhecimentos, com decisão sobre o momento da utilização ou tomada de decisão, pois já é possível entender as consequências do ato (Valentim, 2002).

No panorama enfrentado, o objeto de distinção, na atualidade, não é a existência dos dados, mas o manancial, o grande volume de fatos gerados pelo desenvolvimento, que impede o processamento natural, via ser humano, individualmente considerado. Esse grande volume de dados recebeu a denominação de *big data*, cunhado em língua inglesa que, em tradução livre, significa dados em grande quantidade ou dados enormes. São dados constantemente coletados, razão pela qual os pesquisadores podem lidar com a alterabilidade da dinâmica social, com a produção de estimativas em tempo real, com ganhos para os formuladores de diretrizes políticas e econômicas, bem como estudo de eventos inesperados, fora do padrão. É um conceito em formação, relativamente novo, que se refere à informação, tecnologia, métodos e impacto (De Mauro; Greco; Grimaldi, 2014).

Historicamente, começou a ser trilhado por volta dos anos 2000, quando foi verificada a existência de uma grande quantidade de dados, com taxa de atualização em tempo real, a partir da difusão do uso de dispositivos e plataformas virtuais, sendo constatada a impossibilidade de gerenciamento das informações por meio das ferramentas tradicionais, como observado por Rogers Magoulas (Miranda, 2018).

Coube a Doug Laney, em pesquisa sobre gerenciamento de dados, definir os pilares clássicos dos grandes dados, sob o prisma do volume, variedade e velocidade. Não obstante, em razão do novel conceito, diante da constante atualização inerente à complexidade da *big*

data, foram reconhecidas demais características, como o valor e a veracidade, por exemplo (Barbosa, 2017).

O volume está ligado à quantidade ou crescimento exponencial dos dados, característica mais marcante da *big data.* É preciso estar diante de quantidade extensa de dados, cuja análise demande o uso de instrumentos tecnológicos, porque não são apreciáveis pela atuação individual de um ser humano. Diariamente, ao redor do mundo, milhares de operações são capazes de gerar conhecimento, como interações em redes sociais, aplicativos de chamadas, troca de mensagens eletrônicas, transações bancárias e outras. De acordo com o atributo variedade, os dados podem ser de diferentes tipos, tanto na origem quanto na forma de comunicação, ou ainda, no que toca ao conteúdo. Os dados podem ser coletados a partir de mensagens eletrônicas, redes sociais, áudios, cartões de crédito, fotografias, de origem públicas ou privadas, em comunicação oral quanto escrita. Quanto ao conteúdo, o conhecimento se refere aos diversos valores cultivados no mundo. No que tange à terceira característica clássica, velocidade, os dados devem ser acessados em tempo real, possibilitando decisões mais precisas, em razão do dinamismo do processo de geração. Em um mundo sedento por conhecimento, é necessário agilidade no processamento, pois é relevante entender que os dados serão analisados após um processo, cuja conclusão demanda tempo. Assim, o destempo importa na desclassificação dos dados, em razão da ausência de atualidade (Ribeiro, 2014).

Complementando a visão, o valor é inserido como uma consequência das análises da *big data*, em que se torna possível adquirir informações com grande importância agregada, por exemplo, tendências de mercado, com lucros ou melhoramento de serviços. Diante disso, é de extrema relevância saber a qualidade dos dados a serem pesquisados, pois possuem uma finalidade. A veracidade surge como um requisito relacionado à precisão do processo. Para análises mais precisas é necessário que sejam utilizados dados de conteúdo autênticos, corretos, verossímeis, para que as conclusões busquem um ideal de certeza, não sejam equivocadas, díspares da realidade (Freitas Junior *et al.*, 2016).

Importante lembrar, ainda, que os dados da *big data* impõem olhar crítico sobre as perspectivas práticas e metodológicas da pesquisa (Salganik, 2018). Na estrada trilhada, dados incompletos, inacessíveis e não representativos, podem provocar erro de avaliação e resultados. Outrossim, há dados coletados com enviesamento, cujo conteúdo foi distorcido por algoritmos, dados com erros no processo de inserção e de extração. Há, também, dados sensíveis, isto é, que possuem condição afeta à privacidade ou sigilo, em razão de característica especial da pessoa ou natureza empresarial, público ou privada, fazendo com que se tornem inacessíveis

pela sua natureza, o que dificulta a análise e o resultado. Os dados coletados serão, em regra, incompletos porque provavelmente todas as informações necessárias ao pesquisador não estarão disponíveis no primeiro plano, o que demanda adotar técnicas de preenchimento de lacunas por meio da investigação de outros dados que possam ajudar na conclusão do estudo. Além disso, há possibilidade de terem sido carregados contendo erros, comumente denominados de lixo, que são inconsistências, desvios, redundâncias, o que torna necessário tratamento prévio para correta utilização.

Esse olhar diferenciado para o manancial de dados é de extrema importância para o entendimento e resolução das dinâmicas sociais complexas, que interessam ao mundo corporativo e às ciências, como armas de eficiência e aumento dos lucros, bem como para as instituições governamentais, que podem administrar melhor o emprego de verbas, com identificação das áreas prioritárias, em educação, saúde, meio ambiente, justiça e segurança.

A grande quantidade de dados já interfere no cotidiano dos profissionais e demanda mecanismo de solução, sob pena de desequilíbrio na equação quantidade de trabalho e tempo disponível para lazer e qualificação. Em complemento, a massa exponencial, de natureza privada ou governamental, representativa da dinâmica social complexa, já não pode ser analisada pela atuação singular do humano, necessitando de atuação integrada com a tecnologia.

1.3 Sobrecarga da informação e o acolhimento de ferramentas tecnológicas no âmbito da cultura jurídica

O termo cultura, que na sua origem deriva de *culturae*, está ligado ao significado de trabalho, cultivo ou colheita, encontrando terreno nas formas simples do viver, no cotidiano do homem do campo para, posteriormente, acolher as transformações que levaram ao homem urbano, com toda a sofisticação dos tempos modernos.

As formas da natureza florescem, criam espécies, morrem, reiniciam o ciclo e, da mesma maneira que eclodem espontaneamente, também podem ser cultivadas pelo homem. O processo cultural não é diferente, surge de forma natural pela própria existência do homem, mas também é cultivado por ele. Sofre influência do ambiente em que o ser humano se desenvolve e, ao mesmo tempo, influencia no cotidiano e na formação de novas ideias.

Assim, a cultura através dos tempos é descrita como um processo denso, marcado por novidades, transformações, releituras e refinamentos, é um complexo que resulta de um

processo de construção do homem, acompanhando as transformações sociais e políticas, conforme bem explicado por Terry Eagleton (Eagleton, 2005).

Por vezes, comparada a culto, no sentido religioso, algo de conteúdo espiritual, com laços de divindade e transcendência, passava ao sagrado, protegido e venerado, compondo elemento enraizado nos traços da sociedade, de forma tão latente que passa a indicar as características marcantes de um povo.

A vida em sociedade está ligada ao ideal de convivência e atuação conjugada do ser humano, em ambiente organizado. É resultado das congruências e contradições dos valores presentes, de aspecto múltiplo, caracterizado pela diversidade de ideias, gostos e práticas, campo necessário para desenvolvimento e acomodação do ser humano que, por sua natureza, é um animal social.

No período mais primitivo da vida, o homem retirava sua subsistência da natureza, por meio da caça e da pesca. Os grupos eram pequenos, com homogeneidade ou quase unanimidade de pensamento, de modo que não se tornava primordial o estabelecimento de um conjunto de regras sistematizado, pois bastavam as emoções para resolver os conflitos, predominava um direito natural, independentemente de qualquer ordem escrita.

Com o aumento dos agrupamentos humanos, aparece também a diversidade de pensamento, com choques e contradições, havendo necessidade de maior organização coletiva, com um ordenamento de regras preestabelecidas, em que o olhar individual sede ao sentimento do grupo, pois há um objetivo comum a ser atingido (Santos, 2013). Conforme asseverado por Francois Ost, o tempo e o direito se entrelaçam e exercem influência recíproca na formação da memória da coletividade, por meio da colheita dos fragmentos do passado necessários para construção da ideia de sujeito de direito (Ost, 2005).

Toda sociedade possui valores fundamentais, essenciais, que são os padrões ou condutas coletivamente eleitas. Esses valores acabam sofrendo alterações naturais que são reflexo das mudanças do próprio comportamento coletivo (Inatomi, 2019). Diante disso, o direito não pode ficar estático, ao contrário, deve evoluir com a sociedade, sob pena de se distanciar dos valores por ela acolhidos, tornando ilegítima a sua atuação. Assim, o Direito surge para garantir a segurança e a justiça, pois de nada valeria se não pudesse solucionar as divergências sociais.

Modernamente, aparecem os arquivos e processos digitais, armazenamento em nuvem, sites, plataformas, aplicativos que contribuem para aprimoramento da organização e solução de

conflitos por meio de ações mais ágeis. São tecnologias baseadas na análise de dados, sistemas de inteligência artificial e outros.

A atualização tecnológica no âmbito do Poder Judiciário, com estudo de modelos mais econômicos para promoção do acesso à Justiça, o uso da internet para prestação de serviços jurídicos e a alteração dos parâmetros de comunicação no processo judicial, é um processo gradual e contínuo, sendo diversas as iniciativas dos tribunais brasileiros que se valem de ferramentas computacionais para melhoria de processos em atendimento ao princípio da eficiência.

O Victor, tecnologia adotada pelo Supremo Tribunal Federal, por exemplo, trabalha na conversão de imagens em textos no processo digital, separação e classificação das peças e temas de repercussão geral. Já o Sócrates, no âmbito do Superior Tribunal de Justiça, automatiza as ações, proporcionando a busca de temas jurídicos dos processos, com separação de casos similares e localização de precedentes, entre outros (Silva *et al.*, 2021).

A crescente utilização de mecanismos de melhora da interação processual, como o *legal design*, que resulta da junção de elementos do *design*, da tecnologia e do Direito para aprimoramento dos serviços jurídicos, com foco na resolução das necessidades do destinatário final; bem assim, do *visual law*, parte integrante do *legal design*, que se vale de técnicas de visualização e simplificação da linguagem para facilitar a comunicação nos documentos jurídicos, conforme proposto por Hagan, demanda mudança do perfil das organizações jurídicas (Marconi, 2022).

A substituição de formas desnecessárias por mecanismos de aproximação das partes, com debates acerca do uso da tecnologia na área jurídica, com intimações virtuais, peticionamento eletrônico, oitivas e juízos totalmente digitais, com coleta das provas orais em mídia eletrônica, com incorporação dos mecanismos tecnológicos nas rotinas de trabalho e pesquisa, além da própria regulação da tecnologia digital, demonstram a conexão existente entre o tecnológico e o mundo jurídico.

Tais reflexões são necessárias, pois termos como automação, análise de dados, *starup*, aplicativos, *internet*, redes, crimes virtuais, discurso de ódio, robôs de impulsionamento, usuários falsos, mecanismos de gestão e tantos outros transpassam o mundo tecnológico para se tornarem cotidianos no mundo jurídico, com impacto no processo de acomodação cultural e revitalização da sociedade.

Em complemento, a utilização em massa dos processos tecnológicos vai gerando um volume cada vez maior de dados, com sobrecarga de informação jurídica, influenciando no aumento da demanda dos operadores do Direito e no tempo disponível para atendimento do fluxo de trabalho. Em paralelo, o tempo para aprimoramento profissional também sofre pressão cotidiana, com prejuízo para qualificação, em uma sociedade que prima pela substituição gradual do trabalho humano pela tecnologia.

Em verdadeira contradição, o enfrentamento do problema também vai requerendo o auxílio das ferramentas tecnológicas, uma vez que a manipulação de grande massa de dados, análise de um fenômeno complexo, vai se tornando impossível pela atuação singular de cada ser humano.

1.4 Contribuição das humanidades digitais na formulação de soluções tecnológicas para sobrecarga de informações jurídicas

Os operadores do Direito têm buscado cada vez mais suporte tecnológico para lidar com a enormidade de dados jurídicos disponíveis, tratando dos fenômenos sociais complexos, no campo das Humanidades Digitais.

O surgimento das Humanidades Digitais tem origem na década de 1940, em que o padre Roberto Bussa encontra auxílio no computador desenvolvido pela IBM para digitalização de textos antigos de Tomás de Aquino. O trabalho nesse período inicial foi marcado pela análise de textos, palavras, variantes gráficas, manuscritos, frequentemente caracterizado pela inexistência de aparato tecnológico avançado.

No período de 1970, até meados da década de 1980, há consolidação da utilização dos recursos tecnológicos por campos das áreas das ciências humanas, mas o foco ainda está em literatura e computação, com alguma ênfase em linguística de *corpus*. Mais recentemente, com a eclosão da era da internet, década de 1990, possibilitou-se a realização de projetos colaborativos mais amplos, com a participação de pessoas de diferentes lugares em tempo real, com avanço em relação aos métodos de trabalho anteriores, além de, gradualmente, se tornar uma fonte de quase que inesgotável e superalimentada de informação (Hockey, 2004).

O potencial das multimídias, qualidade e quantidade daquilo que pode ser armazenado, difusão da internet de alta velocidade e a introdução de programas acadêmicos, sobremaneira, em Ciência da Computação, trouxe grande estímulo para as áreas das ciências humanas e

sociais. A complexidade do mundo atual impõe o desenvolvimento de programas interdisciplinares para alcançar o ápice do conhecimento (Vilela; Mendes, 2003).

Tais práticas favorecem indústria, serviços e até mesmo o extrativismo, visto que podem planejar suas fases garantindo aumento de eficiência e produtividade, orientação de custos, fluxo e trabalho, com melhora no atendimento ao cliente, a leitura do comportamento dos consumidores, respeito ao meio ambiente, com o uso sustentável dos recursos.

As instituições financeiras passam a gerenciar melhor os riscos de crédito, limitar as fraudes e direcionar melhor os recursos. Instituições de ensino podem gerenciar o registro dos alunos com melhor gestão dos espaços, disciplinas, finanças, com aprimoramento das pesquisas.

Nos hospitais, leitos podem ser alocados de forma eficiente, com atuação na prevenção e atendimento, por urgência, áreas e doenças. A segurança pública contempla estruturação para prevenção de crimes, com sistemas de inteligência que diminuam o risco do uso da violência policial evitando lesão aos inocentes.

Todos esses acontecimentos refletem na atuação dos profissionais em diversas áreas, com crescimento na demanda de dados e ausência de tempo disponível, em ambiente próprio para auxílio das humanidades digitais.

No universo jurídico, que tem como fonte documentos em texto, de uma maneira geral, herança da necessidade de verificação dos fatos contra os arbítrios estatais, já existem iniciativas de separação ou classificação de dados, sumarização, tratamento e resolução de demandas em massa, com possibilidade de identificação de decisões conflitantes, ou, ainda, novos campos ou teses distintas que podem ser objeto de uma política de conciliação, ou mesmo de mudança de legislação, com diminuição do custo de tramitação dos feitos, em cumprimento ao princípio constitucional da razoável duração do processo e aperfeiçoamento do acesso à justiça.

Não obstante, barreiras de proteção dos algoritmos, programas e demais artefatos tecnológicos, financiamento, modularização dos sistemas e infraestruturas acabam por estabelecer a predominância e direcionamento de estilo, com lesões ou ameaças aos direitos e garantias individuais.

O respeito aos direitos autorais, a necessidade de desenvolvimento e acesso às tecnologias que se mostrem inclusivas, ferramentas e códigos que possibilitem pesquisa por

cientistas de qualquer origem, com quebra de patentes, são temas ainda em ebulição e que precisam ser arrefecidos pelo mundo jurídico.

Os temas chamam atenção para análise do conflito tempo e quantidade de trabalho, em razão do crescimento dos dados, com imediato impacto na atuação dos profissionais jurídicos. Assim, deve haver especial atenção na busca e utilização de tecnologias que favoreçam a atuação dos profissionais jurídicos, com ganhos na relação tempo e quantidade de trabalho.

Uma alternativa para enfrentar o problema do crescimento de dados jurídicos é a sumarização automática de texto, técnica do Processamento de Linguagem Natural (PLN), que poderá ser utilizada como ferramenta tecnológica no auxílio do trabalho dos operadores do Direito.

O processo de transformação cultural provocado pelo acolhimento dos recursos tecnológicos no mundo jurídico é um movimento em curso, com impacto no processamento das informações, capaz de criar métodos e formatos de trabalho, com vantagens competitivas, como a diminuição de custos, o aumento da produtividade, a qualificação profissional. Os profissionais necessitarão desenvolver competências interdisciplinares, com habilidades em tecnologia, atentando para os requisitos dos diversos setores produtivos, para acompanharem as mudanças da nova era anunciada.

2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS COM ÊNFASE EM MINERAÇÃO DE TEXTOS

Esclarecido que a *big data* se constitui em manancial de dados com características próprias e que o crescimento da quantidade de dados gerou sobrecarga de informação, necessário lançar luz para soluções que favoreçam a resolução dos problemas relacionados.

Seguindo a rota para descoberta do tesouro escondido na montanha de dados, como localização, mapeamento e identificação, descrição e estimação de valores, geração de relatórios e gráficos válidos, era necessário o desenvolvimento de tarefas e técnicas capazes de peneirar informações relevantes e transformá-las em conhecimento, desde a formulação dos objetivos, com o direcionamento e gerenciamento do projeto, passando pela seleção do algoritmo adequado, encontrando limitações na própria atuação humana.

2.1 Descoberta de conhecimento em base de dados

A busca por ferramentas adequadas para alcançar informações relevantes acabou forjando uma metodologia, celebrada como Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases*) ou simplesmente KDD, um processo não trivial de identificação de padrões novos, válidos e potencialmente úteis (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

A noção de processo está ligada ao movimento de formulação do conhecimento, uma sequência de passos para auxiliar na tomada de decisão. O conjunto é formado por atos coordenados dirigidos a um resultado, em que as etapas se mostram preparatórias e complementares para as seguintes, executadas de forma interativa, porque envolvem a participação do ser humano, detentor do conhecimento sobre determinada área. Para obtenção de resultados satisfatórios, refinados, pode haver necessidade de repetições totais ou parciais do movimento, em revisões que buscam melhoramento contínuo, o que demonstra seu caráter iterativo.

A característica da não trivialidade é o retrato de um ato complexo, com a conjugação de fatores operacionais e de controle, com as limitações de homem e máquina: existência de fatos válidos e maneira de interpretá-los, dificuldade na formulação dos objetivos, na escolha do algoritmo ou seus parâmetros, por exemplo (Boente; Goldschmidt; Estrela, 2008).

O complexo sistema comporta as etapas do pré-processamento, da mineração de dados e do pós-processamento (Galvão; Marin, 2009). O pré-processamento envolve atividade de captação, organização e tratamento dos dados, com preparação para o algoritmo da fase de mineração de dados. A qualidade dos dados importa na qualidade de conhecimento a ser extraído, há seleção ou redução do que será considerado para o processo de descoberta. O que é importante, o que será objeto de utilização? A incorporação de perguntas é essencial para planejamento do resultado.

Em seguida, os dados passam por uma fase de limpeza: tratamento realizado sobre o conjunto selecionado de modo a garantir a qualidade dos acontecimentos representados. Os erros, dados ruidosos ou divergentes, incompletudes, falta de atributo, inconsistências devem ser corrigidos para afastar contaminação do conhecimento substrato do processo.

A etapa de codificação está relacionada com a forma de transmissão do conteúdo do pensamento humano para alcance da máquina, uma mensagem que possibilite a execução e a extração, uma regra disponível para o entendimento da máquina. Nessa fase, poderá, ainda, ser necessário enriquecer os dados, vale dizer, agregar informações externas aos registros existentes, incorporando qualidade ao conjunto. Isso pode acontecer por meio de pesquisas junto às fontes originais ou mediante consulta de base da dados válidos para complementação das informações disponibilizadas.

A próxima etapa da KDD é a mineração de dados ou *data mining*, que representa o coração do processo, a essência ou núcleo, em que diversas técnicas ou métodos são efetivados para análise extração de estrutura, padrões, tendências. É a fase de aplicação de algoritmos sobre os dados em busca do conhecimento.

Finalmente, na fase de pós-processamento, o especialista na área objeto de pesquisa colhe os esforços da descoberta de dados, que envolve o tratamento do conhecimento obtido, com a geração dos relatórios, avaliação dos resultados com o direcionamento de novos caminhos a serem investigados. Enfim, foram identificados os padrões e a representação do conhecimento, de modo a facilitar a interpretação, avaliação e tomada de decisão pelo homem.

2.2 Mineração de dados

No que tange à mineração de dados propriamente dita, foi estabelecida como um processo de extração de conhecimento a partir da captura e análise de extensa quantidade de dados, tanto para encontrar respostas de acontecimentos do passado quanto para predizer as

novas tendências (Berry; Linoff, 2004). É importante especificar o que será o objeto de busca nos dados, que tipo de estrutura, regularidade ou padrão é perseguido, qual o conjunto de dados, quais os interesses do especialista, qual o objetivo da aplicação da técnica. Dentro do filtro, o problema, questão central a ser solvida, recebeu o nome de tarefa.

A mineração de dados está amplamente preocupada com a construção da técnica ou método a ser aplicado, isto é, conjunto de regras que se conecta a uma coleção de dados para descobrir os padrões de interesse (Amo, 2022). Importante lembrar, ainda, que há diversas técnicas de mineração de dados disponíveis na literatura, cabendo ao pesquisador definir a melhor no plano dos problemas detalhados, inclusive com a possibilidade de combinação entre elas. Portanto, primeiro é necessário entender o problema a ser resolvido (tarefa) para depois aplicar a técnica (método ou tecnologia) na extração do conhecimento, colhendo os resultados do empreendimento.

Sem a propensão de exaurir o tema, enfatizando a questão central a ser resolvida, a título de exemplo, as tarefas a seguir comentadas tem sido comumente apresentadas nos trabalhos sobre o tema: descoberta de associação, classificação, agrupamento, regressão, sumarização, detecção de desvios e descoberta de sequências (Goldschmidt; Passos, 2005).

A descoberta de associação é uma tarefa que descreve relação de dependência sobre fatos ou objetos que tendem a ocorrer juntos em determinado acontecimento, estabelecendo ligações entre campos, em relação de afinidade ou combinação, identificando quais atributos estão relacionados. A ideia clássica de busca por regras de associação foi introduzida em estudo no qual foram levados em conta dados no formato *basket*, que disponibilizavam informações sobre transações de compra realizadas em determinado período. A intenção era criar uma regra, frequente e válida, entre vários produtos, considerando um padrão de suporte e confiança mínimo, para incentivar a aquisição de produtos e serviços. São criadas correlações que permitem melhor disposição dos produtos em lojas ou, ainda, oferecimento dos produtos a partir da identificação de oportunidade de venda cruzada, ou combinação de pacotes de produtos e serviços (Agrawal; Imieliński; Swami, 1993).

As tarefas de classificação e agrupamento encontram similaridades, porém, não se confundem. A atividade de classificação, em geral, surge como imperativo humano, tendo em vista a quantidade de separações de grupos e seleções que se operam na sociedade: profissões, faixa etária, níveis de formação, renda, gênero, raça e outros. Assim, a classificação é uma tarefa supervisionada de mineração de dados que importa em encaminhar um objeto para uma categoria que já tenha sido predefinida, examinando as caraterísticas e selecionando para classe.

O problema ou tarefa, portanto, consiste em identificar padrões de modo a possibilitar que objetos não classificados sejam direcionados para campo especificado, associando corretamente ao rótulo. A ideia é permitir a visualização de grupos específicos, separados em categorias, dos quais são exemplos a identificação de pedidos de crédito (baixo, médio ou alto risco) ou a forma de tratamento de determinado paciente.

O agrupamento (*clustering*) é entendido como uma tarefa não supervisionada de separação, identificando similaridades e reduzindo um conjunto de dados maior para um subconjunto menor, encaminhando dados semelhantes para subgrupos, de modo que os elementos possuam propriedades em comum. Assim, há reunião dos dados por similaridade, em razão das principais características. Embora seja assemelhada à tarefa de classificação, dela difere porque o agrupamento não depende de categorias preordenadas, rótulos, uma vez que a clusterização é feita automaticamente, com os dados são organizados por semelhança (França; Amaral, 2013).

A tarefa de regressão procura estimar em valores, quantidade, possibilitando prever um valor numérico específico. Embora seja possível prever valores numéricos nos problemas de classificação, a estimativa consiste em formar um juízo aproximado de um valor, ligado ao atributo numérico, não tendo finalidade de determinar a classe, rótulo ou categoria (Amaral, 2016). Assim, serve para apreciar a probabilidade de vida, risco de determinado investimento, prever a altura de uma pessoa a partir do peso, entre outros.

De maneira geral, a sumarização consiste em descrever uma base de dados de forma simplificada, há uma síntese, um resumo dos dados mais importantes e representativos, que possibilita a apresentação de conteúdo compactado com ganho no gerenciamento do tempo dos profissionais. A tarefa de sumarização em mineração de dados, também apresentada como descrição de conceitos, é descritiva, com pesquisa e apontamento de características comuns entre conjuntos de dados (Conti, 2011), de que são exemplos os minutos utilizados, chamadas totais, ligações nacionais ou internacionais de um cliente de telefonia, porcentagem de consumidores de determinado item, ou ainda, geração de resumos de textos.

A detecção de desvios consiste na identificação de dados divergentes do contexto padrão, conjunto ou comportamento esperado, possibilitando ou descarte no processo de mineração de dados, como seria uma fraude em compra de cartão de crédito. A descoberta de sequência, por sua vez, seria uma especificação em relação à descoberta de associação, uma vez que procura identificar os itens frequentes considerando operações ocorridas em certo período, para determinar o padrão verificado na sequência (Porcaro; Lifschitz; Mcc, 2002).

De outro lado, são diversas as técnicas (métodos ou tecnologia) a serem empregadas na descoberta dos dados, cuja classificação não encontra uniformidade na doutrina. Para o objetivo do trabalho, cabe indicar, também como exemplificação, diferentes possibilidades de acordo com a tarefa planejada. Assim, as tarefas de classificação, regressão, previsão de séries temporais e clusterização podem efetivadas por meio de métodos baseados em redes neurais; métodos baseados em instâncias para tarefa de classificação; métodos estatísticos, com a utilização do classificador bayeasiano ingênuo para tarefa de classificação, *k-means*, *k-mode*, *k-prototypes*, *k-medoids* para tarefa de clusterização; métodos específicos, com a utilização do algoritmo *Apriori* para tarefa de descoberta de regras de associação; métodos baseados em indução de árvores de decisão utilizando o algoritmo C4.5 na tarefa de classificação; métodos baseados em lógica nebulosa, utilizando o algoritmo Wang-Mendel na tarefa de Previsão de Séries Temporais (Goldschmidt; Passos, 2005).

2.3 Mineração de textos

No panorama jurídico brasileiro, a análise do conteúdo dos autos impõe, não raramente, necessidade de leitura integral, apenas para identificar a finalidade do processo, de modo que a síntese das informações importaria em aumento da produtividade (Sousa; Prata, 2019).

A descoberta de conhecimento por sumarização utiliza técnicas voltadas para extração de termos, análise de conteúdo, bem como análise linguística com finalidade de produção de resumos a partir de textos (Morais; Ambrósio, 2007). As práticas de mineração de texto detêm qualificação na busca por padrões, indexação e sugestão de termos, extração de entidades, léxicos, pesquisa de jurisprudência e doutrina relevante, revisão de contratos, pareceres, identificação de sentença, acórdãos e outras decisões, sendo a sumarização automática um tema do interesse da mineração de textos (Passos Cruz; Vinícius; Leite dos Santos, 2019).

Possibilita, assim, a análise e extração de conteúdo, textos, frases ou palavras, permitindo gestar conhecimento a partir de um dado não estruturado, isto é, livre de formato ou padrão de armazenamento, em campo de especialização da mineração de dados (Machado *et al.*, 2010). As principais técnicas utilizadas são: a recuperação de informação, a extração de informação e o Processamento de Linguagem Natural.

A recuperação de informação está relacionada ao processo de busca e seleção de informações que atendem a uma questão formulada, ou localização de informação acerca de assunto indagado. Há uma procura, não uma resposta propriamente dita, com aproximação dos

dados referentes ao questionamento, razão pela qual existe um desafio no sentido de se aumentar a precisão dos resultados para que o pesquisador consiga localizar, com maior facilidade, o objeto de suas necessidades.

A extração de informação possui como característica marcante a seleção de partes importantes de um documento com o intuito de extrair conteúdo específico sobre esse objeto. Portanto, há um espaço delimitado, uma demarcação do campo de compreensão da linguagem, um lugar específico para agir.

No que tange ao Processamento de Linguagem Natural, ou *Natural Language Processing* (NPL), ligado à Inteligência Artificial (IA), a tecnologia é utilizada para compreender, alterar, sintetizar e interpretar a língua humana, diversificada e complexa. Trabalha com problemas relacionados à automação da interpretação e da geração da língua humana, com o uso contínuo da probabilidade e da estatística, em área adequada para elaboração de tradutores automáticos, ferramentas de revisão textual, análise de sentimentos, sumarização e outras.

2.4 Processamento de linguagem natural

O Processamento da Linguagem Natural (PLN), também conhecido pela denominação Linguística Computacional ou, ainda, Processamento de Línguas Naturais é uma área que tem por objeto o estudo da capacidade e limitações de uma máquina no entendimento da comunicação do ser humano. O desafio inclui reconhecimento morfológico, análise sintática, léxica e semântica, além de interpretar, aprender e extrair informações relevantes, sendo aplicada de forma contínua nos processos de sumarização automática (Benin, 2023).

A linguagem é um ato de expressão de comportamento natural, modo principal de transmissão de mensagens nas relações humanas (Matthews, 2003), caracteriza-se por ser um processo de comunicação, por meio da qual interlocutores transmitem uma mensagem, estando de um lado um transmissor e do outro um receptor. Em geral, apresenta-se sob a forma oral ou escrita, mas poderá ser viabilizada por gestos, símbolos ou sons. As diversas formas de manifestação da linguagem, complexidade e vocação para ambiguidades, demandam desafio de decodificação da mensagem.

A língua, por seu turno, é o conjunto de códigos organizados para comunicação (regras), estando relacionada ao processo cultural de cada sociedade. Outrossim, em analogia com o ser

humano, a ferramenta computacional necessita de um sistema próprio de comunicação, por meio de processos que viabilizem a compreensão das informações (Jurafsky; Martin, 2008).

Devem ser considerados na comunicação da língua: a fonologia, que está ligada ao reconhecimento dos sons das palavras; a morfologia e a sintaxe, que estão vocacionadas para estrutura das palavras ou frases, seja a unidade primitiva (morfologia) ou estrutura da frase (sintaxe); bem como a semântica ou pragmática, direcionadas ao significado. Na semântica, opera-se o significado a uma estrutura sintática, enquanto a pragmática cuida da correlação do conteúdo (significado) com a estrutura sintática. Portanto, há aspectos ligados ao som, estrutura e significado (Pereira, 2011).

2.5 Importância do *corpus* para processamento de linguagem natural

Para descoberta de conhecimento por meio da sumarização, o Processamento de Linguagem Natural tem se valido da Linguística de *Corpus*, que encontra referencial no estudo da língua e seus processos. Por meio da coleta e análise de conjunto de dados, grandes acervos de textos, denominado *corpus* ou *corpora* (plural), termos em latim que significam corpo, o trabalho se apoia em ferramentas que auxiliam no entendimento e manipulação de textos, em que se busca informação ou base de apoio, semelhantemente ao uso de um dicionário, artigos científicos, bulas e outros (Martins *et al.*, 2001).

Os sistemas computacionais buscam realizar a análise sintática de texto (*parsing*) ativando possíveis estruturas gramaticais, identificando e etiquetando partes do discurso, na busca por relacionamentos entre palavras, termos e outras informações, com auxílio do *corpus*. No processo, há categorias de palavras como preposições, artigos e conjunções que não expressam adequadamente a semântica, chamadas *stopwords*, razão pela qual tendem a fazer parte de um inventário de eliminação. Em complemento, o trabalho de segmentação textual, ou tokenização (*tokenization*), requer a delimitação de palavras ou conjunto de palavras de qualidade, conteúdos que recebem o nome de tokens, ou quebra de sentença (*sentence splitting*), quando a segmentação ocorre no nível de sentenças. As atividades descritas são um movimento necessário, porque inicialmente não estão presentes as características linguísticas adequadas para aplicação das técnicas de Processamento de Linguagem Natural (PLN). O resultado, que encontra desafio na diversidade da linguagem humana, muitas vezes é expresso em ambiguidades e incoerência.

Historicamente, o *corpus Brown University Standard Corpus of Present-Day American English*, lançado em 1964, com cerca de um milhão de palavras, é considerado o primeiro exemplo eletrônico do tipo, tendo influenciado no desenvolvimento da área de linguística e na evolução do próprio processo de sumarização.

A construção de um acervo resulta de processo, demorado e criterioso, envolvendo seleção, manipulação, limpeza e formatação de textos, em quantidade suficiente para formar o conhecimento, que inclui registro de certa quantidade de palavras, quantidade de palavras por frases, quantidade de frases ou quantidade de frases relevantes, com posterior remoção de trilhas semanticamente menos importantes no conjunto, em atividade necessária para o processamento computacional (Oliveira *et al.*, 2020).

A manipulação no ambiente do *corpus* é bastante complexa e deve atentar para inúmeros fatores: caracteres especiais, maiúsculos e minúsculos, números, datas, pontuação, problemas de representação do próprio texto, na mesma língua e entre línguas diversas. Em complemento, adversidades são constatadas no pré-processamento de conversão de determinados tipos de arquivos, como pdf para txt, demonstrando a necessidade de desenvolvimento de ferramentas capazes de tratar dos elementos indesejados oriundos de conversão de figuras, notas de rodapé e outros, o que dificulta aplicações subsequentes de Processamento de Linguagem Natural.

Por fim, deve ser entendido que ao lado da estrutura do texto existe um conteúdo subjacente, que é o objetivo de quem expressa a mensagem, razão pela qual o processo de leitura, compreensão e geração do sumário humano requer habilidade para estabelecer relações entre partes, diferenciar as ideias principais das secundárias, com estabelecimento do binômio causa e consequência entre partes e elementos textuais, bem assim entender as relações lógico-discursivas marcadas por conjunções, advérbios e outras.

De grande valia para evolução do tema, o surgimento da Teoria da Estruturação Retórica (RST), permitiu a organização, identificação e o estabelecimento de relações, a partir da segmentação de um texto em unidades elementares de significado ou preposições (Mann; Thompson, 1988), mantendo relação entre si na construção textual. Assim, é possível delinear a forma como as relações devem ser interpretadas para determinar os segmentos relevantes na composição de um sumário e, ao mesmo tempo, determinar o modo de relacionamento com o texto final, preservando a coerência.

2.6 Relevância do desenvolvimento de *corpora* em língua portuguesa para sumarização de texto jurídico no Brasil

O modo de comunicação é um valor que pertence à expressão cultural, elemento formador da identidade dos povos, não obstante, representa barreira para construção de caminhos para sumarização, uma vez que a geração do sumário encontra reflexo nas diversas línguas grafadas, limitada na diversidade de representação.

O ser humano, ao fazer um resumo, vai se aproveitando do seu conhecimento anterior, sintetizando a partir dos termos que conhece. De forma semelhante, a sumarização automática depende de conhecimento preexistente, seja para geração de um sumário a partir da extração estatística de termos, em desafio mais simples, por justaposição, seja para geração do sumário a partir de uma interpretação anterior, em que o embate será maior, havendo necessidade de treinamento prévio, com conhecimento suficiente para geração de resposta adequada (Simonassi, 2016), porque não poderá representar aquilo que não existe.

Há dificuldade de leitura, compreensão e síntese por parte dos seres humanos e, portanto, haverá dificuldades a serem enfrentadas na sumarização automática, razão pela qual o surgimento e aprimoramento de *corpora* favorecerá cada vez mais o desenvolvimento, especialização e adaptação das ferramentas computacionais, viabilizando a efetivação de sumários de melhor qualidade (Cabral, 2015).

Em complemento, os textos jurídicos apresentam natureza especializada, com fundamentação complexa, que inclui teoria, fatos e provas, arcabouço legal e jurisprudência, impondo debate acerca dos dados utilizados no aprendizado de máquina, que pode resultar na exclusão de informações vitais, ou mesmo enviesamento na formação do modelo, em desafio adicional a ser enfrentado (Maranhão; Florêncio; Almada, 2021).

No universo jurídico, a defesa do suposto interesse demanda apresentar todos os argumentos favoráveis, bem como impugnar as razões do adversário, apontando eventuais falhas, demonstrando a tese mais adequada com a lei, a analogia, os costumes ou princípios gerais do direito, com a presença de citações de diversas leis, datas, números, autores e entidades, a fim de influenciar na decisão do juízo (Engelage, 2016).

Notadamente, a construção passa pela complexidade da língua portuguesa, que é o idioma oficial do Brasil, segundo dispõe o art. 13 da Constituição da República de 1988, língua de expressão transnacional, falada em nove países, por cerca de 260 milhões de pessoas na

Europa, América, Ásia e África, produto da colonização de origem portuguesa, conforme informes da Comunidade de Países de Língua Portuguesa (PCLP). Além de ser uma língua administrativa e de trabalho de 27 organizações internacionais, é também a quinta mais utilizada na internet, segundo dados da União Internacional de Telecomunicações (UIT).

Atentando para especialidade, com o propósito de ampliar o debate e construir soluções, esforço de caráter interdisciplinar tem sido empreendido no sentido de fomentar avanços no processamento automático em língua portuguesa, como, por exemplo, a Jornada da Descrição da Língua Portuguesa (JDP), evento do Simpósio de Tecnologia da Informação e Linguagem Humana que reúne linguistas e pesquisadores da Ciência da Computação.

Entretanto, há poucas ferramentas computacionais em língua portuguesa, em comparação com o espectro da língua inglesa, o que reforça a necessidade de desenvolvimento de *corpora* especializada como parte integrante da evolução da sumarização automática de textos no âmbito nacional, sendo válido, sem a propensão de descrever o estado da arte ou mesmo a revisão bibliográfica de *corpus* em língua portuguesa, percorrer iniciativas sobre o tema, finalizando com o RulingBR, um *corpus* formado a partir de decisões judiciais voltado para sumarização automática de textos jurídicos.

2.7 *Corpus* em língua portuguesa para sumarização de automática de texto

O TeMário é formado por textos jornalísticos, criado no projeto EXPLOSA do Núcleo Interinstitucional de Linguística Computacional (NILC), idealizado a partir da experiência de um sumarizador profissional, um consultor e um professor de língua portuguesa, contando com cerca de 100 textos jornalísticos (Pardo; Rino, 2003). Houve atualização, em 2006, com o acréscimo de mais 150 textos (Maziero *et al.*, 2007).

Com origem no ideal de geração de um analisador discursivo multidocumento automático para o português do Brasil (Aleixo; Pardo, 2008), o CSTNEws foi construído a partir de 50 coleções de textos de assuntos variados, com média de 4 documentos anotados, segundo a teoria discursiva multidocumento CST (*Cross-document Structure Theory*), utilizada para descrever conexões semânticas entre tópicos relacionados.

O CM2News é apresentado para conteúdo bilíngue, utilizado em multidocumento, com origem em textos jornalísticos em português e inglês. A primeira versão partiu de 20 clusteres distribuídos pelas categorias mundo, política, saúde, ciência, entretenimento e meio ambiente, totalizando 19.983 palavras, em uma primeira abordagem. Em um movimento de expansão, foi

possível ampliar o universo para 27.270 palavras, em auxílio de pesquisas multidocumento envolvendo aplicação da língua portuguesa (Camargo; Di-Felippo, 2019). O CM3News é uma ampliação do CM2News, com a inclusão do alemão para compor o *corpus* multilíngue (português, inglês e alemão) de notícias jornalísticas (Nascimento, 2020).

Em estudo atual visando a sumarização abstrativa multilíngue, a WikiLingua incorporou uma base de dados com artigos em 18 idiomas, incluindo a 881.695 em língua portuguesa, correspondentes a textos em inglês, cobrindo diversas categorias do conhecimento humano, como artes, saúde, entretenimento e cuidados pessoais (Ladhak *et al.*, 2020).

O XLSum foi idealizado para trabalho de sumarização abstrativa multilíngue abrangente e diversificado, compreendendo cerca de 1 milhão de artigos, de 44 idiomas, extraídos da *British Broadcasting Corporation – BBC*. Inicialmente, contava com uma base de 23.521 amostras em língua portuguesa que, mais tarde, incluiria o acréscimo de 71.725 (Hasan *et al.*, 2021).

O Summ-it foi formado com 50 textos jornalísticos do caderno de Ciências da *Folha de S.Paulo*, retirados do *corpus* Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em português do Brasil – PLN-BR. A construção do *corpus* Summ-it++ recebeu acréscimo, em 2016, de duas camadas semânticas: a entidades nomeadas e seus relacionamentos (Antonitisch *et al.*, 2016). Nasceu a partir do Projeto ProCaCoSA (Processamento de Cadeias de C00-referênncia para a Sumarização Automática de textos em português), que tinha a finalidade de viabilizar os avanços acerca do discurso e da sumarização automática de textos em português (Collovini *et al.*, 2007), considerando o uso de informação de correferência e o emprego da relação retórica desenhada na Rhetorical Structure Theory – RST.

No âmbito do Projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital), com o intuito de gerar um sistema para simplificação automática de textos em língua portuguesa, é noticiada a utilização de um *corpus* de 187 textos, obtidos do jornal *Folha de S.Paulo* (Maziero; Pardo; Aluísio, 2009).

O CoMET – *Corpus* Multilíngue para Ensino e Tradução, desenvolvido junto ao Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, apresenta uma lista corpora de português. Um dos componentes é o *Corpus* do Português, compilação de natureza histórica, extraído de textos do Brasil, Portugal, Moçambique e Angola, com cerca de 45 milhões de palavras do século XIII ao XX. Outra

iniciativa é o *Corpus* Brasileiro, uma coletânea de aproximadamente um bilhão de palavras em português brasileiro (Projeto AC/DC).

No estado atual, o TeMário e o CSTNews tem sido referenciados como as bases de apoio mais tradicionais nas pesquisas de sumarização em língua portuguesa, além do uso das bases WikiLingua e XL-Sum, multilíngues, que incluem também grande quantidade de textos em português (Paiola, 2022).

2.8 RulingBR

O RulingBR (Feijó; Moreira, 2018) foi apresentado como um *corpus* formado por cerca de 10 mil decisões (acórdãos) oriundas do Supremo Tribunal Federal, órgão que se encontra no ápice da pirâmide do Poder Judiciário brasileiro, responsável pelas decisões em única e última instância relativas às matérias de natureza constitucional, com influência no âmbito administrativo, civil, processual civil, consumidor, penal, processual penal, econômico, eleitoral, ambiental, financeiro, direitos fundamentais, trabalhista, notarial, público internacional, previdenciário, tributário e outros.

Tradicionalmente, o acórdão é uma decisão judicial proferida por um órgão colegiado de magistrados: juízes, desembargadores ou ministros, em julgamentos de competência originária, em segunda instância recursal ou tribunais superiores. De acordo com os artigos 458 e 563 do Código de Processo Civil, o acórdão é composto por elementos essenciais, como a ementa (síntese do próprio acórdão, com os pontos fundamentais da decisão), o relatório (composto pelos fatos narrados, o direito aplicado), a motivação ou fundamentação (razões de convencimento com as justificativas que impõem uma certa decisão) e, finalmente, o dispositivo (parte final que caracteriza o poder decisório, a manifestação do Poder Judiciário). Nesse sentido, há complexidade estrutural e gramatical a ser considerada na seleção e manipulação do tipo de documentação para formação do *corpus*.

A construção passou pela utilização da biblioteca *Scrapy* para capturar o inteiro teor das decisões de 2010 até 2018, com a identificação de alguns pontos indesejados como cabeçalhos, rodapés, número de páginas, com seleção do texto das seções com auxílio do *JavaScript Object Notation* (JSON), com arquivo final disponibilizado no endereço https://github.com/diego-feijo/rulingbr/.

O *corpus*, formado a partir do conjunto de dados públicos de decisões judiciais, é um importante instrumento de auxílio avanços na sumarização automática de textos jurídicos em

língua portuguesa, uma vez que a especialização do conjunto tende a ser mais efetiva na produção de resultados (Spirling; Rodriguez, 2022), criando expectativas positivas.

3 SUMARIZAÇÃO AUTOMÁTICA DE TEXTO

Para fazer frente ao crescimento e armazenamento de dados em grande escala, foi idealizado um processo de síntese, facilitando busca e seleção do conteúdo, esbarrando inicialmente na ausência de tecnologia aperfeiçoada, formatos de difícil acesso e baixa qualidade dos sumários produzidos. O processamento de grande volume de dados virtuais impõe capacidade de assimilação do conhecimento, com enorme disponibilidade de tempo, que seria otimizada se houvessem sumários disponíveis (Rino, 1996).

Assim, a sumarização automática (SA) se colocou como um processo de destilação da informação, com fonte em um ou mais documentos, para produzir uma versão simplificada. A máquina passou a condensar um texto-fonte, de forma automática, gerando um sumário (resumo), propondo ganho de produtividade na relação tempo e quantidade de trabalho.

3.1 Sumarização humana

O ato de sumarizar faz parte da essência do ser humano, que não tem capacidade de armazenar todos os conhecimentos, apenas o importante. Há simplificação por símbolos, gestos, expressões úteis para acessar volume de informação de maneira mais eficiente. Partindo dessa premissa, é um ato natural, separando o que é descartável do que se reputa fundamental para agregar outros valores que irão compor o conhecimento total.

O trabalho humano envolve compreensão (interpretação), seleção de informações hierárquicas (relevância) e junção dos conteúdos, com a produção de um sumário final. Nos tempos mais remotos, a delimitação de conhecimento era marcada por meio de resumos mentais com a transmissão oral das informações, em grupos menores. Posteriormente, o próprio crescimento das informações acaba por induzir a criação de sumários escritos como forma de materialização da memória. Na atualidade, as informações, sobremaneira, na forma de texto, aparecem para apreciação das atividades profissionais, como instrumento principal ou auxiliar, em volume cada vez maior, o que demanda uma nova etapa no ciclo de síntese.

A sumarização humana é complexa, influenciada por fatores subjetivos, como grau de escolaridade, educação, domínio do tema, costumes e crenças, nível de detalhe e linguagem, razão pela qual apresenta resultados distintos, a partir do mesmo texto-fonte. Além disso, características como o ponto de vista do autor ou interesses a serem atendidos, foco do leitor, a

natureza e tamanho da publicação, ditam regras de construção, resultando em sumários variados, de que são exemplos os literários, jornalísticos, artísticos, científicos e outros.

Do ponto de vista da comunicação, o emissor da mensagem utiliza suas habilidades para transmitir um conhecimento, com um ponto médio de informação, considerada satisfatória quando permite recuperação da ideia central pelo receptor. Ao atentar para sumarização, a transmissão da informação, para ser satisfatória, deve ocorrer também de forma concisa.

No processo de transmissão da informação se estabeleceu distinção entre microestrutura, relação entre as sentenças do texto-fonte, e macroestrutura, relações entre as sequências de blocos de sentenças e o conteúdo geral do texto (Hutchins, 1987). Assim, o sumário humano expressaria a macroestrutura, que atribui maior importância para a sequência de informações em conjunto com a ideia principal, implícita a preocupação com a coerência, o que revela uma das maiores barreiras a serem enfrentadas e implementadas no âmbito da sumarização automática de texto.

3.2 Sumarização automática

O sistema de sumarização automática está estruturado nas etapas de análise, transformação e síntese. A análise consiste na verificação do texto-fonte possibilitando uma interpretação inicial, com a extração do conteúdo que será processado automaticamente: são aplicadas ferramentas de pesquisa da morfologia, sintática, do conteúdo semântico e do discurso (Sparck Jones, 1993). A transformação é o núcleo do processo, com seleção do conteúdo relevante da etapa anterior e produção de uma estrutura interna do sumário. Na fase de síntese, há produção do sumário propriamente dito, com elementos de linguagem natural, com fusão, justaposição, condensação ou ordenação das sentenças mais importantes. Notadamente, o conteúdo extraído deverá representar informações relevantes ou a ideia central.

Os sumários automáticos podem apresentar uma vasta gama de propriedades, com diversas classificações propostas (Santos, 2012). Não obstante, de forma a situar o leitor no objeto do trabalho, serão dispostas as mais correntes, que servirão de norte para comparação de desempenho dos sumarizadores automáticos.

Inicialmente, é possível estabelecer diferença entre geração, cujo resultado considera o inteiro conteúdo informativo, dito da melhor forma possível, enquanto o termo sumarização é empregado para tarefa de diminuição do texto-fonte, sem perda significativa do seu conteúdo (Rino, 2005) .

Quanto à forma de gerar ou construir, o resultado coincide com um extrato ou *extract*, em inglês, texto mais curto, que resulta das partes mais importantes do texto-fonte, por justaposição; ou um sumário ou *abstract*, em inglês, que surge da fusão ou reedição do texto-fonte, formado por sentenças novas, com estrutura diferenciada do inicial (Camargo, 2013). Para o objetivo do trabalho, as ideias expostas serão tratadas como sumário.

A sumarização automática extrativa, ou estatística, resulta da seleção das sentenças mais relevantes ou representativas do conteúdo do texto-fonte, apresentando-se como uma atividade menos complexa, de justaposição ou junção de partes selecionadas, sem maior preocupação com eventual coerência, embora mereça crítica na análise da qualidade. A sumarização automática abstrativa, ou fundamental, por sua vez, acaba por produzir alterações, com novas sentenças, que têm por objetivo melhorar a coerência resultado do texto reduzido, com partes reescritas, buscando eliminar ambiguidades e redundâncias (Rino, 2005).

Conforme a abordagem, com foco no nível de conhecimento linguístico (De Luca, 2019), o sumário poderá ser superficial ou profundo, ou ainda de modelagem híbrida. A ideia está em percorrer o caminho da menor complexidade para maior, em termos de compreensão de conteúdo, partindo da morfologia para sintaxe, da semântica até alcançar a pragmática e o discurso, com interação entre os níveis, visto que não é um processo isolado. Em aderência, a complexidade da atividade também aumentará no percurso descrito, com maior uso de capacidade computacional.

Na abordagem superficial há pouca preocupação com a complexidade do conteúdo, com conhecimento potencial da língua, sendo fundamentada em dados estatísticos, pela seleção e junção das sentenças do texto-fonte que refletem os termos mais frequentes, não considerando o arsenal linguístico disponível na mente humana, razão pela qual recebem a denominação de métodos cegos. Assim, restaria mais próxima de uma sumarização extrativa, de atividade computação menos complexa.

A abordagem referenciada como profunda procura modelos e teorias dos recursos formais da língua, com preocupação na seleção das partes mais importantes do texto-fonte que refletem o discurso e compreensão efetiva do conteúdo, com resumos de qualidade superior, se comparado ao superficial, de modo que estariam relacionadas com a sumarização abstrativa. A abordagem profunda está conectada com o conhecimento linguístico, processos cognitivos e inferência lógica, habilidades próprias da mente humana, com a utilização de atividade computacional e algoritmos mais complexos para obter um processamento eficiente da língua. Por fim, a abordagem híbrida é um processo que combina aspectos superficiais e profundos.

No plano da quantidade de línguas objeto de exploração, o sumário poderá ser gerado no mesmo idioma dos documentos de entrada, realizada em determinada língua (monolíngue); em abordagem entre línguas (*cross-lingual summarization*), quando o documento de entrada está em um único idioma, mas utiliza informações de outro idioma; ou ainda a sumarização multilíngue (*multilingual summarization*), com documentos de entrada de diferentes idiomas, viabilizando a exploração de resumos em diversas línguas, especialmente, quando houver ausência de *corpus* para treinamento, em demanda ainda desafiadora (Lloret; Palomar, 2012).

Atentando-se para quantidade de material analisado, a sumarização automática é classificada em monodocumental, quando parte de apenas um único texto-fonte. De outro lado, a chamada sumarização multidocumental está direcionada para seleção de vários textos-fontes ou coleção, com crescente destaque, tanto do ponto de vista das informações interdisciplinares, quanto pela complexidade de tratamento de arquivos de tecnologia diversas, com equalização do conteúdo semântico (De Luca, 2019).

A sumarização intrasentencial é definida como aquela que ocorre no interior da mesma sentença (dentro dos limites da sentença ou da frase), por meio da inclusão de conteúdo mínimo, de maneira que o resultado diminuto mantenha a informação principal. Envolve a tarefa de segmentação de grupos linguísticos como entidades nomeadas, verbos e nomes, em um processo de organização e agrupamento de unidades de informação individual que irão influenciar na informação coletiva. A sumarização intersentencial ocorre entre sentenças diferentes, ou mesmo parágrafos, envolvendo agrupamento em uma mesma sentença ou parágrafo (Leitão, 2015).

Em outra classificação, são abordagens supervisionadas aquelas em que se faz a construção de sumários a partir de rótulos anteriormente definidos pelo ser humano, sumários construídos manualmente para um conjunto de textos, enquanto os métodos não supervisionados prescindem de pré-seleção humana para decidir quais são os atributos mais importantes de um documento, com vistas à geração de um sumário, razão pela qual necessitam de algoritmos mais sofisticados (Paiola, 2022).

No campo da função, os sumários podem ser divididos em críticos, informativos e indicativos. Os críticos, apresentam conteúdo avaliativo, representativo de opiniões, por exemplo, as resenhas de livros. Os informativos são caracterizados por conteúdos coerentes, na gramática e no tema, de tal forma que poderiam dispensar o próprio texto-fonte. Em ponto derradeiro, os indicativos são comumente conhecidos como índices, isto é, organizam as partes do texto-fonte sem substituir seu conteúdo (Mani e Maybury, 1999).

No que toca aos receptores da mensagem ou audiência, os sumários são considerados genéricos quando traduzem informações gerais, sem preocupação com o interesse específico de um grupo. Na outra ponta, os sumários podem ser customizados para demanda especial, com conteúdo para um grupo especializado, focados nos interesses típicos de certos leitores (De Luca, 2019).

Quadro 1 – Principais classificações na Sumarização Automática de Texto

Critério	Classificação	Referência
Forma	Extrato ou abstract	(Camargo, 2013)
Método baseado na forma	Extrativo (estatístico) ou abstrativo (fundamental)	(Rino, 2005)
Abordagem quanto ao conhecimento linguístico	Superficial, profunda ou híbrida	(De Luca, 2019)
Quantidade de línguas	Monolíngue, multilíngue ou entre línguas	(Lloret; Palomar, 2012).
Número de textos-fontes	Monodocumento ou multidocumento	(De Luca, 2019)
Partes objeto do sumário	Intrasentencial ou intersentencial	(Leitão, 2015)
Supervisão	Supervisionados, não supervisionados	(Paiola, 2022)
Função	Indicativo, informativo ou crítico	(Mani e Maybury, 1999)
Audiência	Genérico ou focado no interesse do leitor ou usuário	(De Luca, 2019)

Fonte: Elaborado pelo autor (2022).

3.3 Revisão bibliográfica da sumarização automática de texto

Os estudos delineadores da sumarização automática apareceram no final da década de 1950, com a geração de resumos baseados em palavras-chave, que refletiam maior incidência dos termos capazes de reproduzir a suposta ideia principal de um texto-fonte (Luhn, 1958). Naquele início, sem maiores preocupações com o nível semântico, a síntese era gerada pelo agrupamento das sentenças com palavras de maior frequência.

Críticas foram dirigidas no sentido de que sentenças construídas apenas pela frequência das palavras necessitavam de maior coerência, uma vez que algumas elegidas (artigos, preposições, pronomes) não eram capazes de representar a essência do conteúdo do texto-fonte, razão pela qual foi demandado criar um inventário de determinados termos, classificados como *stopwords*, que deveriam ser retirados da tabela de importância, como critério de refinamento. Por outro lado, o modelo apresentava vantagem econômica em razão do baixo custo computacional para geração de extratos superficiais.

Em nova abordagem, que optava pela utilização da localização da sentença no texto-fonte como critério de relevância, foi sugerida seleção do texto da primeira e da última sentença, ou ainda, expressões chaves e palavras tendência, ou previstas no título, consideradas importantes no processo de extração (Baxendale, 1958). Entretanto, atentando que os títulos, as primeiras e últimas sentenças podem não apresentar conteúdo significativo, o método mostrou-se incapaz de produzir sumários consistentes.

Cerca de uma década depois, em linha semelhante, era aplicado um método de sumarização em que as palavras-chave estariam inseridas no título ou cabeçalho, ou palavras do título, que possuíam maior probabilidade de serem representativas do tópico fundamental do texto-fonte, contribuindo positivamente para formação de um sumário final mais coerente (Edmundson, 1969).

Contribuíram para evolução, métodos que indicavam escolha pelo tamanho das sentenças, caso em que as curtas e longas restavam excluídas, por supostamente não representarem conteúdo relevante. Assim, as sentenças de tamanho médio justificavam manutenção, seleção e incorporação, em que pese a complexidade do ajuste de métricas para extração. Outros modelos, utilizavam a combinação de características como a localização ou existência de palavras específicas, sinalizadoras, com maior potencial de representação da substância do texto-fonte, podendo ser citada sumarização implementada pelo cruzamento de sentenças relevantes com o próprio título do texto-fonte, com restrição de domínios ou assunto (Pollock e Zamora,1975).

Note que os experimentos iniciais atentaram para seleção pelo critério da relevância de palavras, sentenças ou partes do texto-fonte como processo fundamental de extração. São métodos extrativos, superficiais, fulcrados na justaposição de partes para construção da síntese. Embora verificados avanços, houve estagnação da evolução do tema, sobremaneira, pela ausência de insumos adequados para implementação de novos desafios (hardware, software, dicionários e repositórios linguísticos em quantidade e conteúdo que pudesse ser processado pela máquina). A retomada ocorreria na metade dos anos 1980, com os impulsos da inteligência artificial, a produção de material em meio digital, mecanismos de Processamento de Linguagem Natural, com maior preocupação com a coerência do resultado, trilhando avanços na área abstrativa (Martins *et al.*, 2001).

De relevo citar os estudos sobre competência linguística, processos de compreensão e apreensão de Chomsky, que contribuíram na formação da própria linguagem computacional e no entendimento sobre a estrutura textual (Chomsky, 2002) que passou a ser objeto de análise,

atentando-se para o gênero e tipo de texto (Mckeown, 1985), além de possibilitar o desenvolvimento de uma abordagem profunda, em emulação do processamento da mente humana. Com fundamento no classificador proposto por Naive-Bayes, foi introduzida a utilização do aprendizado de máquina para tarefas de sumarização automática, em trabalho que se tornou referência (Kupiec; Pedersen; Chen, *et al.*, 1995). A combinação de várias características textuais, a presença de *corpus* de auxílio, representou um avanço. De acordo com o classificador proposto, era possível aferir que a probabilidade de cada sentença do texto-fonte fosse incluída no sumário, a partir do reconhecimento de características textuais fundamentais, que demonstrassem o juízo de importância na formação do extrato.

Sugiram experimentos fulcrados tanto na frequência de termos (técnica estatística) quando na importância específica de um evento (técnica simbólica), sumarização de um ou vários documentos, na aplicação do SUMMARIST, em que as técnicas são manipuladas para identificar o tópico de um documento, para compreensão e geração do resumo (Hovy; Lin, 1997).

Na busca de sumários que se aproximassem das sínteses humanas, em continuidade, o campo da linguística auxiliava no desenvolvimento de *corpus* necessários para treinamento, em movimento de refinamento da sumarização por meio da textualidade, com técnicas capazes de gerar resumos cada vez mais claros e atraentes, refletindo o conteúdo do texto-fonte (Moens; Teufel, 2001).

Um dos primeiros experimentos de sumarização em língua portuguesa foi o GistSumm, método baseado na existência de uma sentença *gist*, também conhecido como identificação de *gist*, composto pelas palavras mais frequentes do texto-fonte que, paralelamente, recebe informações de sentenças complementares, com palavras em comum ou compatíveis, agregando informações relevantes (Pardo; Rino 2002; Pardo, 2005; Pardo; Rino; Nunes, 2003). Em complemento, a evolução da sumarização extrativa incorpora experimentos com métodos baseados em grafos (Mihalcea, 2004), estrutura utilizada para representar um modelo em que existem relações entre objetos de determinados conjuntos, vértices ou nós, interconectados com arestas ou arcos. Podem ser referidos, ainda, métodos baseado em *Term Frequency – Inverse Document Frequency* (TF-IDF), que mede a relevância de um termo dentro de um documento em relação a um *corpus*, tem sido amplamente utilizado no Processamento de Linguagem Natural, bem como técnicas baseadas em *Bidirectional Encoder Representations* (BERT), capazes de compreender padrões humanos da língua humana, sobremaneira, a relação das

palavras dentro de uma frase (Devlin *et al.*, 2019), a partir de uma fase de pré-treinamento, com auxílio de um *corpus* que colabora para o entendimento do contexto de aplicação das palavras.

Assim, no campo extrativo, houve organização do seguintes conhecimentos: (a) atributos ao nível da palavra: palavras-chave, palavras do título, expressões-chave, palavras de tendência, palavras em caixa alta; (b) atributos ao nível da sentença: localização da sentença, comprimento da sentença, localização do parágrafo, coesão entre sentenças (Paiola, 2022).

A abordagem abstrativa surge com a necessidade de melhorar a coesão e coerência dos sumários gerados pela abordagem extrativa. Passam a ser necessários não apenas a justaposição das partes importantes do texto-fonte, mas efetivamente compreender o conteúdo do texto-fonte, condensar o conhecimento e reescrever o tema de forma sintetizada (Rino, 2005).

Os modelos baseados na Teoria da Estruturação Retórica (RST), teoria descritiva que estuda a organização de um texto estabelecendo as relações existentes entre suas diversas partes, foram de grande valia para evolução do tema. Assim, permitindo delinear como as relações devem ser interpretadas para determinar os segmentos relevantes na composição de um sumário e, ao mesmo tempo, determinar o modo de relacionamento com o texto final, é possível gerar sumários mais coerentes (Mann; Thompson, 1988). É preciso analisar a retórica do texto-fonte para, posteriormente, fixar o conteúdo do sumário subsequente, de forma que foi desenvolvido estudo para segmentação do discurso para identificar o tópico de importância e, então, estabelecer o relacionamento de informação necessário para formação do sumário adequado (Marcu, 1997).

Em linha mais recente, a evolução do Processamento de Linguagem Natural deu impulso a sumarização abstrativa permitindo o desenvolvimento de modelagem discursiva em profundidade, com aplicação de redes neurais, método inspirado no sistema nervoso do ser humano, possibilitando a realização de aprendizado e reconhecimento de padrões. Há referência à métodos de compressão e fusão de sentenças (Condori, 2015).

Por meio da identificação de métodos relevantes, a abordagem abstrativa se desenvolveu no seguinte sentido: (a) baseadas em estrutura, que buscam as informações mais importantes em um documento e as representam por meio de estruturas predefinidas, como grafos, árvores ou ontologias, para criar sumários abstrativos; (b) baseadas em semântica, isto é, obter uma representação semântica do texto e, a partir disso, gerar um sumário; (c) baseadas em aprendizado em profundidade, uma estrutura elaborada de ponta a ponta, que permite treinamento e aprendizado para concatenar as principais informações do texto-fonte e gerar o

sumário correspondente. A maioria dos métodos desta abordagem utilizam o modelo *encoder-decoder*, útil para entrada (sequência de entrada) e saída (sequência de saída), em sequências de palavras, problema conhecido como Seq2Seq, possibilitando a entrada de uma sequência maior e a saída de uma sequência de tamanho menor (Paiola, 2022).

Por fim, conforme pontuado nas linhas introdutórias do trabalho, as frequentes atualizações científicas são capazes de superar as tecnologias pelo advento de instrumentos mais modernos, razão pela qual os estudos no trabalho estão limitados pelo limite temporal compreendido entre março de 2021 e março de 2023.

Assim, atualmente, já se discute a utilização da tecnologia de inteligência artificial ChatGPT (*Generative Pre-trained Transformer*), desenvolvida pela OpenAI, possibilitando a geração de textos originais, o que inclui a sumarização de texto, a partir do reconhecimento formado com o trabalho de treinamento. Notadamente, já foi possível a geração de decisão judicial no Poder Judiciário da Colômbia, por meio da utilização do ChatGPT como instrumento para auxiliar no rascunho da decisão, conforme noticiado no excelente trabalho direcionado para utilização de inteligência artificial generativa no Direito (Tavares, 2023).

3.4 Métodos de avaliação dos sistemas de sumarização

A avaliação dos sistemas de sumarização é tema de relevo pois desafios são trilhados no sentido de obter resultados que possam refletir o conteúdo do texto-fonte, pois saber o que será objeto de avaliação é tão importante quanto a avaliação em si (Estival; Sparck Jones; Galliers, 1997). Surgem ponderações sobre as partes e métricas que melhor preservam o conteúdo: algoritmo, taxa de compressão, sistemas disponíveis, usabilidade e qualidade.

Anualmente, há encontros na conferência *Text Analysis Conference* (TAC), incentivando pesquisas em Processamento de Linguagem Natural e tecnologias correlatas, fornecendo coleção de testes, procedimentos e foro para divulgação dos resultados acerca dos métodos de avaliação, de extrema relevância para o progresso, essencial para comparação e apresentação dos resultados, estimulando novas iniciativas.

Tratada no plano extrínseco, a avaliação é voltada para análise de quanto um sistema influencia ou impacta em outra tarefa, de modo que um sumário será considerado adequado quando utilizado com sucesso em outro contexto. Assim, um sumário pode não ser adequado para leitura do humano, mas perfeitamente aprovado para uso de ferramenta computacional para categorização de textos, recuperação de informação ou tarefa de perguntas e respostas, por

exemplo. No plano intrínseco, que tem por objeto o desempenho do sistema em si, são avaliados os resumos por meio da aferição de sua qualidade, isto é, grau de coesão, parâmetros gramaticais e informatividade (Estival; Sparck Jones; Galliers, 1997).

Há métodos identificados como caixa preta *(black-box)*, que avaliam o resultado do sistema como um todo, enquanto a avaliação transparente *(glass-box)* é medida por parâmetros, em cada etapa interna, um módulo do sistema, por exemplo. Na outra extremidade, quanto ao conteúdo dos sumários, a avaliação é apresentada como *off-line*, quando automatizada, enquanto se reserva o termo *on-line* para avaliações realizadas pelos humanos. Por fim, a avaliação autônoma é aquela realizada no sistema em si, separadamente, enquanto a avaliação comparativa coteja métricas comuns em diversos sistemas (Pardo, 2008).

Os pesquisadores têm utilizado com maior frequência ferramentas de avaliação automática ou semiautomáticas, ainda que demandem maior evolução quanto ao grau de confiabilidade nos quesitos de julgamento. As automáticas buscam saber quais as informações devem restar mantidas, o reconhecimento ou identificação de certas partes no resumo ou avaliação da gramaticalidade (Gambhir; Gupta, 2017) encontrando maiores dificuldades quando aplicadas aos métodos abstrativos e profundos.

A literatura apresenta diversos mecanismos de avaliação, com base na (i) similaridade do cosseno, (ii) sobreposição de unidades lexicais (unigrama ou bigrama) e (iii) sobreposição da maior subsequência de unidades lexicais (Saggion; Poibeau, 2013), ou ainda o método Pyramid, que tem foco na semântica (Nenkova e Passonneau, 2004).

Não obstante, a Rouge (*Recall-Oriented Understudy for Gisting Evaluation*) (Hovy; Lin, 2003) tem sido difundida como o caminho mais utilizado na avaliação de sumários automáticos, além de, no plano da informatividade, apresentar desempenho semelhante ao humano para ranquear sumários. O método se vale das métricas *precision* (P), *recall* (C) e *f-measure* (F), respectivamente, precisão, cobertura e medida-f, para realização de comparação com sumários de referência, produzidos manualmente por especialistas humanos, determinando o cômputo automático da quantidade de palavras em sequência (n-gramas) em comum, no sumário de referência e no gerado automaticamente, de modo que quanto mais n-gramas houver, maior será a nota atribuída ao sumário gerado (Nascimento; Guelpeli, 2011).

Há diversas métricas disponíveis no pacote: Rouge-1, baseia-se em 1-grama (avalia quantas vezes cada palavra estará sendo representada no texto); Rouge-2, fundamentado em 2-grama (avalia quantas vezes um par de palavras estará sendo representada no texto); Rouge-3

(3-grama) e Rouge-4 (4-grama), avaliam quantas vezes um conjunto de repetição de 3 ou 4 palavras aparece representado; enquanto o Rouge-L procura as maiores subcadeias comuns entre dois textos (Uzêda, 2007).

3.5 Aferição da qualidade do texto de um sumário

Avaliar todos os aspectos de um sumário é tarefa complexa. Um texto é formado por um conjunto de elementos que expressam um significado, composto por unidades menores: palavras, frases, sentenças e sinais de pontuação, utilizados para comunicação. Cada elemento isoladamente pode não ter sentido para o conjunto, razão pela qual deve ser analisado sistematicamente, em função do todo.

No campo automatizado, a área de Análise Discursiva Automática (ADA) compreende técnicas computacionais de aferição de discursos, auxiliando na correção ou padronização de avaliação de textos. Por meio da tecnologia, por exemplo, foi possível identificar aumento consistente de correções conflitantes (40%), no ano de 2013 e 2014, em relação aos anos anteriores, no Exame Nacional do Ensino Médio – Enem, o que demonstra a dificuldade da padronização da avaliação textual, especialmente pela presença do elemento subjetivo, resultado de diferentes aspectos da formação humana. Considerando a quantidade de ferramentas disponíveis para língua portuguesa, em comparação com avaliadores de língua inglesa, o tema é parco de recursos no senário nacional (Cândido; Webber, 2018).

Quanto à avaliação quantitativa, o critério referido para avaliação de informações do texto é a apreensibilidade (*readability*), conforme definida por Klare, que é a facilidade de entendimento a partir da forma de escrita, consistindo em contagem de frases, palavras e sílabas, com produção de um resultado estatístico, buscando referência nas características do vocabulário e no tamanho dos elementos textuais, para perquirir o nível de dificuldade, determinando o grau de escolaridade necessário para entendimento do conteúdo (Falavigna *et al.*, 2020).

Os índices mais frequentemente referenciados são: o *Flesch Reading Ease*, o *Flesch Kincaid Grade Level* e o *Gunning-Fog Index* (Goldim, 2006). O *Flesch Reading Ease*, de Rudolf Flesch, encontra fundamento na fórmula que parte do número de sílabas do texto e o número de frases para cada amostra de 100 palavras, numa escala de 0 (difícil) a 100 (fácil), em que o resultado de nível mais elevado indica maior facilidade para entendimento do conteúdo. As métricas do índice *Flesch* também são referidas como forma de calcular o índice

de legibilidade (Martins *et al.*, 1996). O *Flesch Kincaid Grade Level (Flesh-Kincaid)*, considerado o mais utilizado, é uma releitura de *Flesch* para medição do grau de escolaridade, aplicadas ao treinamento da Marinha dos Estados Unidos, partindo de uma fórmula que inclui a divisão do número total de palavras pelo número total de frases, e o número total de sílabas pelo número total de palavras, em que o resultado representa o nível de escolaridade exigido para entendimento do conteúdo.

O *Gunning-Fog Index*, de Robert Gunning, teste de legibilidade encontrou fundamento na percepção de que algumas publicações eram expressas em escrita desnecessária, com palavras complexas, criando dificuldade de entendimento, uma neblina (razão do termo *'fog'*). Assim, descreveu uma fórmula que incluía a divisão do número de palavras e o número de frases, bem assim o número de palavras complexas (polissílabas) pelo número de palavras, em um trecho de texto com 100 palavras, em que o resultado expressava o nível de escolaridade necessário para o entendimento do conteúdo, em que um índice menor que 8 indica facilidade, e um maior que 12, dificuldade (Carvalho *et al.*, 2023).

No que tange à avaliação qualitativa, tema de maior interesse no trabalho, o termo qualidade encontra origem no vocábulo latino *qualitate*, que designa algo virtuoso, característica particular do objeto ou indivíduo em adequação com os requisitos ditados, ou nível de perfeição requerido pelo processo, por meio da observação e julgamento, baseado em parâmetros preestabelecidos, ligada ao processo cultural, com parâmetros locais e universais. É um conceito polissêmico, desencadeando diversos significados, em razão das possibilidades de interpretação de acordo com as diferentes aptidões valorativas (Oliveira; Araujo, 2005).

Exatamente pela dificuldade de compreensão e geração de sumários de qualidade, a avaliação humana, também conhecida como manual, ainda costuma ser a mais apropriada, em tarefa que demanda tempo e esforço, embora represente custo elevado, de difícil reprodução e suscetível aos erros próprios da falibilidade humana, dotada de certo grau de subjetividade (Pardo, 2008).

De acordo com o referido por Beaugrande e Dressler, o critério que mede a capacidade de comunicação ou expressão do texto, para que não seja apenas um conjunto de elementos é a textualidade, composta pela coerência e coesão, a intencionalidade, a informatividade, a aceitabilidade, a situacionalidade e a intertextualidade (Costa Val, 1991).

O autor de um texto tem objetivo comunicativo (intencionalidade): passar uma mensagem, produzir um discurso que seja entendido e apreciado pela outra parte, que pareça

coerente para o leitor, que tenha relacionamento linguístico adequado e faça sentido no ambiente social. Da mesma forma, a mensagem deve conter o conteúdo necessário (informatividade) e estar inserida em um contexto (situacionalidade). Na outra ponta, o leitor cria expectativa quanto à mensagem recebida (aceitabilidade), capaz de levá-lo a adquirir conhecimentos ou auxiliar nos objetivos do autor. Essa participação do leitor, que em grande parte demanda conhecimentos prévios, está espelhada na coerência (quer entender o conteúdo), na coesão (quer entender o relacionamento entre os elementos linguísticos) e na intertextualidade (que o conteúdo faça sentido na sua vida social) aumentando ou diminuindo seu interesse pela leitura. Há um movimento de cooperação.

Para tratar do tema qualidade, em razão do conteúdo polissêmico do termo, será utilizado o binômio qualidade textual (coerência e coesão) e informatividade (Pardo, 2008), pois constituem fatores fundamentais que influenciam no interesse dos leitores, compondo o campo da textualidade. Outrossim, a expressão qualidade em sentido amplo será destinada para se referir à avaliação da qualidade do texto de um sumário, como forma a diferenciar da expressão qualidade textual, critério de qualidade.

De relevo entender que há certo grau de subjetividade no plano da avaliação humana, uma vez que coerência, a coesão e a informatividade estão em estreita dependência dos conhecimentos partilhados pelos interlocutores, em desafio de demarcar os limites da textualidade, razão pela qual estabelecer critérios qualitativos passa por captar e sistematizar condições naturais de aceitabilidade dos discursos (Costa Val, 1991).

Acompanhando o plano de ideias, a análise e julgamento da coerência (contexto, organização, articulação), da coesão (relacionamento entre os elementos linguísticos do texto) e da informatividade (conteúdo necessário presente), deve ocorrer em função do texto global, do contexto, verificando as relações com o tema, entre suas seções com o todo, marcada sempre por um grau de subjetividade, em razão da realidade que o texto propõe.

A qualidade textual (Pardo, 2008), em que se inserem coerência e coesão, encontra características fundamentais como clareza, refletida na boa organização das ideias, sem obscuridade; ausência de redundância, que permite dizer o máximo com o mínimo de palavras; correção, identificada pela gramaticalidade, respeito às normas cultas; na harmonia ou elegância, seja pela musicalidade da frase ou estilo de escrita.

A informatividade, por seu turno, é identificada pela presença, no resumo, de conteúdo relevante contido no texto-fonte, cuja ausência influencia no interesse do leitor. Avaliar essa

característica significa verificar sucesso na empreitada em levar conhecimento para o leitor, no plano conceitual, aquilo que se espera encontrar na experiência, ou no plano da expressão (imprevisibilidade), o que tem de novidade ou inesperado no texto que torna a percepção mais envolvente.

O reconhecimento de aspectos relevantes do texto depende de conhecimento prévio, razão pela qual trabalhos de natureza científica possuem maior grau de informatividade, porque são destinados a um público específico, requerendo mais atenção e recursos de processamento; textos de natureza jornalística, por exemplo, estão incluídos no grau médio, porque necessitam de alguns conceitos técnicos para o entendimento, com ocorrência de menor trivialidade; enquanto matérias veiculadas em assuntos do senso comum, com argumentos universalmente aceitos, estariam inseridas no conceito de baixa informatividade (Fávero, 1985).

A informatividade está presente no plano da cobertura e da precisão, utilizados como métricas clássicas de avaliação. Partindo da ideia de que há um sumário de referência produzido por um ser humano, considerado adequado, a cobertura é o grau de percepção de alcance do sumário automatizado em relação ao sumário de referência, o quanto há de informação em comum entre eles. A precisão diz respeito ao grau de proximidade atingido pelo sumário automatizado no que toca à similaridade de informação do sumário de referência. A medida-*F* aparece como um média entre precisão e cobertura (Nascimento, 2020), que mede a eficiência do sistema, determinando o quão próximo do ideal foi possível chegar.

A taxa de compressão, que é métrica utilizada para medir a dimensão da condensação do conteúdo expresso no resumo, indicando a proporção entre o tamanho de um texto-fonte e o sumário produzido, poderá influenciar no grau de informatividade, uma vez que a redução significativa do conteúdo do texto-fonte pode ocasionar eliminação de informação relevante. Em linha ordinária, os sumários devem resultar em 10-20% do texto-fonte, razão pela qual as taxas de compressão tendem a ser fixadas em 80-90% (Rino; Pardo, 2003).

Outrossim, é possível medir a informatividade por meio da comparação de um sumário referência, produzido para servir de parâmetro, que facilita a instituição de sistema de avaliação automático. A referência pode surgir de sumário produzido pelo próprio autor do texto-fonte (sumário autêntico), por um profissional ou especialista (sumário profissional), ou pelo extrato das partes mais importantes do texto-fonte (sumário ideal).

4 COMPARAÇÃO DE TRABALHOS DE AVALIAÇÃO DAS FERRAMENTAS DE SUMARIZAÇÃO AUTOMÁTICA SOB O PONTO DE VISTA DA QUALIDADE DOS SUMÁRIOS PRODUZIDOS

Avaliar todos os aspectos de um sumário é tarefa árdua em razão dos diversos critérios possíveis. A comparação, sob o ponto de vista da qualidade em sentido amplo, conforme tem sido proposto, passa pelo diagnóstico sobre o estado atual na geração de sumários de interesse do leitor, bem assim, em momento mais avançado, resposta sobre a utilização pelos operadores do Direito das ferramentas computacionais vocacionadas para sumarização.

Sem a propensão de esgotar o tema, serão referidos diversos trabalhos de sumarização automática, com foco no processamento em língua portuguesa, demonstrando a relevância do tema no contexto atual de sobrecarga de informação. Outrossim, os trabalhos apresentados, cujos resultados servirão de base para comparação no campo da qualidade dos sumários gerados, já foram submetidos a métodos de avaliação.

É importante ressaltar que não há uniformidade no tratamento do tema, de modo que alguns experimentos trabalham com o conceito de textualidade, outros se referem aos componentes da qualidade textual (coerência e coesão) e informatividade.

4.1 Sumarizadores automáticos de texto que viabilizam o uso da língua portuguesa

Relacionar as ferramentas de sumarização automática, especialmente, utilizando *corpus* em língua portuguesa, permite atestar a diversidade dos experimentos: abordagens distintas, técnicas extrativas e abstrativas, mono e multidocumento, de caráter mono, multilíngue ou entre línguas. É necessário advertir, desde o início, que a maioria dos trabalhos em língua portuguesa está relacionada à abordagem extrativa, apresentando carência no que toca ao método abstrativo, talvez em razão da dificuldade do desenvolvimento de vasto material na língua nacional, o que constitui obstáculo ao aprendizado dos modelos (Paiola, 2022).

Em razão da difusão do Microsoft Word, editor de textos em português (Microsoft, 2003), uma das ferramentas de sumarização automáticas mais utilizadas foi o AutoResumo do Word, oferecendo abordagem extrativa baseada na frequência de palavras, para simplificação

de documentos, segundo a compactação desejada pelo usuário, servindo, ainda, de parâmetro de comparação do desempenho de diversos sumarizadores (Espina; Rino, 2002).

Outro sistema foi o *TextAnalyst* (TextAnalyst, 2003), que possibilita a sumarização de natureza extrativa, em diversos idiomas, por meio da utilização de redes neurais que permitem a criação de grafos com inter-relacionamentos de informações úteis, com apontamento dos elementos de maior frequência para o nível hierárquico mais alto, com a sentenças mais importantes sendo selecionadas para compor o sumário.

O sumarizador TF-ISF-Summ, ou *Term Frequency – Inverse Sentence Frequency* (Larocca Neto *et al.*, 2000) propõe ranquear sentenças em função da representatividade das palavras que a compõem. Nos métodos extrativos, em geral, há prevalência da frequência das palavras ou termos apresentados no documento. Ocorre que, em diversos casos, as palavras ou termos de maior frequência não representam as palavras ou termos de maior conteúdo semântico. Assim, o *Term Frequency – Inverse Sentence Frequency* se propõe a resolver o problema, definindo que a relevância será espelhada no produto de dois termos, um pela frequência das palavras e o outro que atribui peso maior para palavras que aparecem em menor frequência nas sentenças. As sentenças com a pontuação média, com a TF-ISF mais alta, serão selecionadas para compor o extrato (Rino *et al.*, 2004).

O GistSummarzer foi concebido inicialmente como um sumarizador automático de texto monodocumenntal, de natureza extrativa, intersentencial e genérico (Pardo; Rino; Nunes, 2003). Posteriormente, o trabalho foi aprimorado, recebendo novas funcionalidades (Pardo, 2005), sendo um dos sumarizadores mais referenciados em trabalhos científicos. A lógica funciona a partir de um texto-fonte de entrada para produção do extrato, de interesse genérico, apresentando resultado baseado na segmentação, ranqueamento, seleção de sentenças e junção das consideradas mais importantes sem, entretanto, permitir a sumarização no interior de cada sentença. O processo apresenta regras de segmentação que impõem a identificação das sentenças por meio dos sinais de pontuação como ponto final, exclamação e interrogação. Em prosseguimento, há atividade de ranqueamento (classificação) das sentenças, sendo aberta, em sequência, uma fase de seleção das sentenças ranqueadas, com a coleta das que contenham as ideias mais importantes, isto é, que apresentem coincidência com a sentença *gist* da etapa anterior, bem como as consideradas relevantes, que serão as que estejam acima da média obtida. Outrossim, o número de sentenças selecionadas estará ligado à taxa de compressão que tenha sido designada.

O NeuralSumm é um sumarizador de característica extrativa, que se vale da técnica de aprendizado de máquina em rede neural do tipo SOM *(self-organizing map*) para identificar as sentenças mais importantes (essenciais, complementares e supérfluas), classificando-as no grupo por similaridade, em processo que se aproxima da ação do cérebro humano (Pardo; Rino; Nunes, 2003). No processo, serão consideradas no sumário apenas as sentenças essenciais e complementares, descartando as supérfluas.

O DMSumm (*Discourse Modeling Summarizer)*, embora esteja no campo da sumarização automática com base no modelo de discurso de Rino, tem sido classificado como um gerador de sumários, com a elaboração de resumo a partir a seleção do conteúdo, planejamento textual e a realização linguística propriamente dita, com entrada de uma representação interna do texto-fonte (interpretação) e seleção das informações relevantes dessa entrada (Pardo; Rino, 2002). Assim, o sumário final é resultado da síntese das informações da representação elaborada internamente, com a participação de um usuário especialista nos modelos linguísticos utilizados.

Na mesma classe anterior, encontra-se o UNLSumm, como gerador de sumário, em que a codificação da representação interna é formada com o auxílio da interlíngua, *Universal Networking Language* – UNL, que possibilita a comunicação entre diversos idiomas pela exploração das informações das línguas naturais, com a finalidade de se obter uma representação neutra de significado (Martins; Rino, 2001). O processo termina com a apresentação de um sumário que é a representação do significado mais provável do texto-fonte.

O SuPor (Ambiente para Sumarização Automática de Textos em Português) é um sumarizador de natureza extrativa que se vale do aprendizado de máquina bayesiano, modelo sugerido por Kupiec, para decidir sobre a seleção e composição do sumário, com base em diversas características como posição da sentença, presença ou não de palavras específicas, cumprimento da frase e outras possibilidades. O sistema recebeu atualizações e a nova versão é denominada SuPor-2 (Leite; Rino, 2006), sendo que, em trabalho comparativo com demais sumarizadores em língua portuguesa, apresentou os melhores resultados na métrica informatividade (Leite, 2010). O ClassSumm (*Classification based Summarizer*), também baseado no modelo de Kupiec, utiliza até doze características superficiais para seleção e composição do sumário (Larocca Neto *et al.*, 2002).

O kNNSumm (*K-nearest neighboors Summarizer*) é descrito como um sumarizador automático que se vale do aprendizado de máquina baseado em instâncias, sobremaneira pela aplicação do classificador k-NN, utilizando uma base de documentos formadas por notícias

extraídas do *Wall Street Journal* da coleção Tipster. Surge a partir da necessidade de estender as funcionalidades do ClassSumm, permitindo a realização da tarefa de sumarização como uma tarefa de classificação (Silla Jr.; Kaestner, 2007).

O SATSumm (Nascimento Neto; Gomes; Neto, 2007) foi apresentado em trabalho para sumarização automática de textos jornalísticos, com abordagem superficial, com a geração de um dicionário com lista de palavras-chave e palavras menos usuais na língua portuguesa, com utilização da técnica *Term Frequency-Inverse Sentence Frequency* (TF-ISF) para pontuar as sentenças mais importantes do texto.

O TSumm é um sumarizador automático de natureza extrativa, com processamento superficial, baseado em características linguísticas clássicas dos textos-fontes em português, utilizando a distribuição da frequência dos componentes textuais para construção do extrato final (Espina; Rino, 2002).

O CSTSumm (CSTSUMMarizer) é um sumarizador automático multidocumento com base na teoria discursiva CST (*Cross-document Structure Theory*), que utiliza *corpus* em língua portuguesa, com metodologia em quatro etapas: agrupamento de textos de conteúdo similar, estruturação interna dos textos (via análise sintática, por exemplo), estabelecimento de relações CST (gerando, assim, um grafo em que os nós representam as sentenças e as arestas as relações CST), e seleção de sentenças para compor o sumário. Essa seleção constrói um ranque de sentenças em função da relevância para formar o sumário (Jorge; Pardo, 2010).

O PragmaSUM é um método de sumarização automática extrativa, que se utiliza de palavras-chave para permitir a personificação de sumários, com vocação para textos educacionais. Para tanto, foi formado um *corpus* composto apenas por trabalhos científicos na área educacional para testes e comparações com outros sumarizadores (Rocha, 2017).

O Sumex é descrito como um programa gerado em Visual C++ 6.0 para testes das estratégias de sumarização automática extrativa para textos em língua portuguesa, utilizando um programa de extração de radicais (*stemmer*) baseado no algoritmo de Porter. As palavras-chave são encontradas e transformadas nos seus radicais, e estes serão utilizados para a seleção de sentenças para o sumário. São adotadas estratégias de palavras-chave e título, palavras-chave e título e localização, palavras-chave e título e palavras sinalizadoras (Souza; Nunes, 2007).

No campo da sumarização abstrativa, é narrada experiência de sumarização abstrativa de opinião, por meio do treinamento com o corpus *OpiSums-PT*, escrito em português brasileiro (Condori; Pardo, 2017). Seguindo a linha de investigação de método voltado para opinião, em

português brasileiro, por meio da anotação e adaptação do *corpus* OpiSums-PT, foi utilizada abordagem abstrativa com representação semântica AMR (*Abstract Meaning Representation*), grandes bancos de dados (*sembanks*), com sentenças e respectivas representações semânticas (Inácio, 2021).

Foi apresentado, ainda, estudo para geração de resumos contrastivos que destacavam as diferenças entre entidades, a partir de textos opinativos extraídos do site de brasileiro Buscapé, voltado para pesquisa de produtos, sobre quatro itens: duas câmeras e dois *smartphones* (Rocha da Silva; Salgueiro Pardo, 2022). Em adição, foi elaborada pesquisa voltada para sumarização automática abstrativa, com modelo de aprendizagem profunda, para o português brasileiro, considerando a utilização de modelos pré-treinados, para produzir sumários, mesmo em cenários com poucos recursos, em termos de quantidade de dados, com auxílio do *corpus* TeMário e CSTNews, além do uso das bases WikiLingua e XL-Sum, multilíngues, que incluem textos em português (Paiola, 2022).

Por fim, para os objetivos deste estudo, deve ser mencionado o LegalSumm (Feijó, 2021), um método de sumarização de natureza abstrativa e abordagem profunda voltado para decisões judiciais, baseado em representações de codificação Bert (*Bidirectional Encoder Representations)* e implicação textual (*textual entailment*), capazes de gerar poder de decisão na seleção de frases mais adequadas para compor o resumo final (Lapata; Liu, 2019).

Partindo da premissa de que é necessário fornecer resumos mais rápidos, padronizados e eficientes das decisões proferidas pelos tribunais, atualmente gerados por seres humanos, em processo que se mostra custoso e moroso, demandando seleção e manutenção de profissionais com conhecimento bastante específico, distintos quanto ao nível de percepção, escrita e atuação, foi desenvolvido o LegalSumm. A ferramenta computacional cria fragmentos de texto e constrói um sumário-candidato para cada, com remoção de tópicos irrelevantes e geração de resumo final, com seleção via maior pontuação, com fundamento nas representações de codificação bidirecional de transformadores (Bert) e implicação textual (Feijó, 2021).

O método de Processamento de Linguagem Natural de reconhecimento de implicação textual (RIT) está ligada à possibilidade de determinar se o sentido de uma passagem ou trecho de texto está inserido ou contigo em outro trecho (Fonseca, 2018). O trabalho com Bert é direcionado para compreender padrões da língua humana, especialmente, como as palavras se relacionam dentro de uma frase.

No trabalho foram organizadas as principais técnicas aplicadas à sumarização de textos, passando pela coleta dos dados de decisões judiciais, atividades de limpeza e representação, necessárias para construção de um *corpus* jurídico especializado em língua portuguesa, o RulingBR. Os sistemas de comparação passaram por um processo de adaptação e treinamento, com concatenações das seções "relatório", "votação" e "julgamento". A ideia, que inclui a capacidade de trabalhar com documentos longos, divididos em seções predefinidas, como são os acórdãos, trata de iniciativa específica de sumarização dirigida para o campo jurídico, como o mérito de aplicar o treinamento em língua portuguesa, gerando resumos com maior grau de semelhança com o comportamento humano, adotando um viés neutro de discurso, em função dos diversos estilos de escrita dos diversos profissionais jurídicos.

Quadro 2 – Estado da arte de sumarização automática de texto que viabilizam o uso da língua portuguesa

Sumarizador específico ou técnica utilizada	**Referência**	**Critério de Forma**	**Técnica** **Monodocumento ou** **Multidocumento**	**Técnica** **mono, multilíngue ou entre línguas, com uso da língua portuguesa**
AutoResumo Microsoft Word	(Microsoft, 2003) e (Espina; Rino, 2002)	Extrativo	Monodocumento	Multilíngue
TextAnalyst	(TextAnalyst, 2003)	Extrativo	Monodocumento	Multilíngue
TF-ISF-Summ	(Larocca Neto *et al.,* 2000) e (Rino *et al.*, 2004).	Extrativo	Monodocumento	Multilíngue
GistSumm	(Pardo; Rino; Nunes, 2003) e (Rino *et al.*, 2004).	Extrativo	Monodocumento	Multilíngue
NeuralSumm	(Pardo; Rino; Nunes, 2003) e (Rino *et al.*, 2004).	Extrativo	Monodocumento	Multilíngue
DMSumm	(Pardo; Rino, 2002)	Extrativo	Monodocumento	Multilíngue
UNLSumm	(Martins e Rino, 2001)	Extrativo	Monodocumento	Multilíngue
SuPor	(Leite; Rino, 2006)	Extrativo	Monodocumento	Multilíngue

Sumarizador específico ou técnica utilizada	**Referência**	**Critério de Forma**	**Técnica Monodocumento ou Multidocumento**	**Técnica mono, multilíngue ou entre línguas, com uso da língua portuguesa**
ClassSumm	(Larocca Neto *et al.*, 2002) e (Rino *et al.*, 2004)	Extrativo	Monodocumento	Multilíngue
kNNSumm	(Silla Jr.; Kaestner, 2007).	Extrativo	Monodocumento	Multilíngue
SATSumm	(Nascimento Neto; Gomes; Neto, 2007)	Extrativo	Monodocumento	Multilíngue
TSumm	(Espina; Rino, 2002)	Extrativo	Monodocumento	Multilíngue
CSTSumm	(Jorge; Pardo, 2010)	Extrativo	Multidocumento	Multilíngue
PragmaSUM	(Rocha, 2017).	Extrativo	Monodocumento	Multilíngue
SUMEX	(Souza; Nunes, 2007)	Extrativo	Monodocumento	Multilíngue
Sumarização de opinião	(López Condori; Salgueiro Pardo, 2017)	Abstrativo	Multidocumento	Português brasileiro
Sumarização de opinião	(López Condori; Salgueiro Pardo, 2017)	Abstrativo	Multidocumento	Português brasileiro
Sumarização de opinião	(Rocha da Silva; Salgueiro Pardo, 2022)	Abstrativo	Multidocumento	Português brasileiro
Sumarização com aprendizado de máquina em profundidade	(Paiola, 2022)	Abstrativo	Multidocumento	Português brasileiro
LegalSumm texto de natureza jurídica	(Feijó, 2021)	Abstrativo	Monodocumento	Multilíngue

Fonte: Elaborado pelo autor (2022).

4.2 Comparação dos resultados das ferramentas de sumarização sob o ponto de vista da qualidade em sentido amplo

Na avaliação da performance do GistSumm, em hipótese que era indagado se, com base na *gist sentence*, é possível construir bons extratos, no que tange à textualidade, o resultado foi positivo, com validação da experiência. Os resultados obtidos apontavam que 55% dos extratos gerados pelo GistSumm estavam acima da média e 14% dos extratos estavam na média; 50% dos extratos preservaram totalmente a ideia principal e 40% preservaram parcialmente (informatividade); 50% dos extratos apresentaram textualidade total e 35% apresentaram textualidade parcial. Dessa forma, 90% dos extratos preservaram totalmente ou parcialmente a ideia principal (informatividade) e 85% dos extratos apresentaram textualidade total ou parcial, permitindo uma adequada compreensão (Rino, Pardo, 2003).

Segundo exposto na avaliação do sumarizador GistSumm (algoritmo *Gist*) em comparação com o algoritmo de Luhn (Muller; Granatyr; Lessing, 2015) ficou atestado que ambos eram capazes de gerar resultados satisfatórios que tange à manutenção da ideia principal do texto (informatividade), embora o sumário baseado no GistSumm tenha produzido maior refinamento se comparado ao de Luhn.

Trabalho desenvolvido em textos voltados para crônicas esportivas evidenciou, de maneira geral, que os extratos gerados manualmente, utilizando o método da palavra-chave, ou automaticamente, por meio do TMSumm, são satisfatórios se considerados a preservação da ideia central de um texto-fonte (informatividade), inclusive em comparação com o AutoResumo (Espina; Rino, 2002).

Em avaliação de métodos profundos em sumarização multilíngue para gerar extratos em português, a partir de ranqueamento, com seleção apenas de sentenças também em português, de pontuação mais alta, até que a taxa de compressão desejada fosse atingida, ou, ainda, pela seleção de sentenças mais bem pontuadas independentemente de sua língua-fonte, apresentaram resultados promissores com sumários de boa qualidade linguística e informatividade (Tosta, 2014).

Experimento de aferição do SATSumm, em comparação com o GistSumm, testificou que o sumarizador apresenta textualidade razoável, mantendo bom nível de conteúdo quanto à ideia central do texto-fonte (informatividade). Outrossim, a ferramenta logrou êxito em selecionar, em média, 70% das mesmas sentenças triadas manualmente por seres humanos (Nascimento Neto; Gomes, Neto, 2007).

Em relação ao DMSumm, ficou exposto que é composto dos processos da geração automática de textos, isto é, a seleção de conteúdo, o planejamento textual e a realização linguística. No processo de avalição, comparando sumário autêntico e sumário automático reproduzido, foram utilizados dez juízes, linguistas computacionais e falantes nativos do português do Brasil, que julgaram os sumários apresentados para cada texto do *corpus*. O objetivo do experimento foi medir a qualidade dos sumários, o que foi feito segundo uma série de parâmetros. Foram apresentados aos juízes 4 sumários por texto: 3 automáticos e 1 autêntico. Os sumários automáticos foram completamente gerados pelo DMSumm e correspondem, em média, a 40% do textos-fonte. Foram duas as tarefas dos juízes para cada texto do *corpus*: identificar o sumário autêntico e classificar cada sumário de acordo com pontos de decisão, que são os parâmetros do julgamento.

Quanto à textualidade, a maioria dos sumários automáticos manteve índice de 67% dos textos-fonte. Entretanto, comparados aos autênticos, dos quais 90% mantiveram a textualidade, essa taxa foi considerada baixa. Em termos da informatividade, houve proximidade dos sumários automáticos e autênticos que, pelo tamanho apresentado, representavam quase o mesmo conteúdo em relação ao texto-fonte.

Em suas conclusões, embora promissor, o autor esclareceu que o experimento carecia de aprimoramento no processo de realização linguística, demonstrando preocupação com a textualidade, com implementação da variedade de línguas naturais necessárias aos objetivos atuais de pesquisa, com o acoplamento de modelo de usuário, permitindo ao especialista direcionar a elaboração de sumário menos detalhado, com a geração de resumos mais dedicados aos seus interesses específicos, com expansão do modelo de discurso de Rino para outros objetivos comunicativos (Pardo; Rino, 2002).

Em temática envolvendo a eficiência da utilização da palavras-chave no sumarizador automático de textos, por meio da Rouge, utilizando *corpus* Educacional para testes formados por artigos científicos na língua portuguesa e com dez domínios da grande área da educação, foram avaliados os sumarizadores BLMSumm, GistSumm, e o PragmaSUM com o método original e utilizando palavras-chave com os métodos sequência, classificação e TF-ISF.

Com a métrica de precisão (informatividade), todos os métodos de sumarização com as palavras-chave mostraram vantagem sobre o método original do PragmaSUM. Em complemento, quando comparados ao GistSumm e BLMSumm, essa vantagem foi ainda maior. Com o aumento da taxa de compressão utilizada, a diferença, entre os valores obtidos pelo PragmaSUM e demais sumarizadores, foi maior. O PragmaSUM, em comparação com o

GistSumm, no domínio História da Educação, obteve os melhores resultados (Rocha, 2017). Além disso, o autor concluiu que, de maneira geral, a aplicação de taxas de compressão maiores, em textos menores, influencia na informatividade, com a geração de sumários melhores.

Estudo envolvendo aplicação do Sumex com o AutoResumo do Word, em abordagem de sumarização extrativa para a introdução de um artigo científico (Nunes; Souza, 2001), utilizou dezoito artigos científicos de computação retirados da Revista Brasileira de Informática na Educação e dos anais do Simpósio Brasileiro de Informática na Educação – 1998. Assim, foi gerado um sumário para cada estratégia de sumarização apresentada e foi feita uma análise da textualidade e da proximidade na preservação da ideia principal do texto (informatividade) por cada sumário. Para as estratégias Palavras-chave, Palavras-chave + Localização, ou Palavras-chave do autor, os sumários resultantes foram comparados com o texto original, com o sumário feito pelo autor do artigo científico e com o sumário feito pela ferramenta AutoResumo do Word. A taxa de compressão utilizada no AutoResumo foi de 25% para todos os textos.

O estudo concluiu que todos os sumários obtidos pela estratégia Palavras-chave + Localização mantiveram a ideia principal do texto (informatividade), possivelmente porque os sumários são muito extensos e praticamente iguais aos textos originais. As estratégias Palavras-chave e Palavras-chave + Sinalizadoras apresentaram resultados razoáveis; poucos textos apresentaram alto percentual de erros. Importante notar que, a maioria dos sumários resultantes do AutoResumo do Word apresenta algum erro de coesão e coerência (aferição de qualidade textual), mas poucos são os completamente incoerentes.

Na estratégia Palavras-chave do autor, 27.7% dos textos resultaram completamente incoerentes (obtendo o pior resultado entre todas as estratégias) mostrando que muitas vezes o autor não escolhe boas palavras-chave para seu artigo (aferição de qualidade). A experiência chamou atenção para problemas de textualidade (aferição de qualidade textual e informatividade) nos resultados e a necessidade de continuidade do desenvolvimento de técnicas de sumarização de textos que contemplassem o processamento da língua portuguesa.

Em trabalho envolvendo a potencialidade de sumarizadores para o idioma português do Brasil, de temática jornalística e de gênero científico, foram apresentados dois experimentos com avaliação intrínseca comparativa de quatro sistemas extrativos de sumarização: GistSumm, NeuralSumm, AutoResumo e TextAnalyst. A primeira avaliação verificou as potencialidades dos sistemas para a sumarização de textos jornalísticos com uma taxa de compressão de 70%. Na outra, para sumarizar textos do gênero científico, gerando extratos com aproximadamente

15% (compressão de 85%) das sentenças dos textos-fonte, tendo os extratos produzidos recebido notas quanto a sua informatividade e textualidade.

Uma característica importante a ser considerada na sumarização extrativa é que os textos jornalísticos geralmente apresentam o conteúdo essencial nos primeiros parágrafos, uma técnica de relatar o que há de principal nos acontecimentos logo no início da notícia, como um apelo ao leitor ou abertura, conhecido como lide ou *lead*, em inglês. Os textos científicos, por seu turno, são comumente divididos em várias seções necessárias para organizar ideias e fundamentar o trabalho. O NeuralSumm apresentou insuficiência tanto em relação aos textos jornalísticos, no quesito informatividade e textualidade (60% de seus extratos constituíam textos com deficiência), quanto aos científicos (30% dos extratos preservaram totalmente a ideia central e 30% preservaram parcialmente), gerando sumários com problemas de informatividade.

O GistSumm, por seu turno, apresentou resultados satisfatórios no tema jornalístico, quanto a textualidade e informatividade, mas insatisfatórios para textos científicos, no critério informatividade, com sumários que cobriam completamente a ideia central em apenas 30% e parcialmente em 10%. O TextAnalyst evidenciou desempenho satisfatório, na textualidade e informatividade, enquanto o AutoResumo foi colocou como a melhor opção para sumarização de textos no âmbito do estudo (Oliveira, 2008).

Para o experimento com CSTSumm (CSTSUMMarizer) foi usado um *corpus* composto de 50 coleções de textos jornalísticos escritos em português brasileiro, sendo que cada coleção tem 2 ou 3 textos sobre o mesmo tópico. Esse *corpus* contém a análise CST (realizada manualmente) de cada coleção de textos e o resumo humano (genérico, sem preferências) correspondente, cujo tamanho corresponde a 30% do tamanho do maior texto da coleção (em número de palavras). Os sumários automáticos foram produzidos considerando a mesma taxa de compressão dos sumários humanos. Foram considerados dois métodos de avaliação: o automático, para medir a informatividade dos sumários genéricos, e o humano, que é usado para avaliar fatores como a informatividade, coerência, coesão, gramaticalidade e redundância dos sumários com preferências do usuário.

Para a avaliação automática, foi usada a medida Rouge comparados os resultados com os obtidos pelo único sumarizador GistSumm e o sumarizador Mead, que é um sumarizador multilíngue superficial. Os resultados apresentados indicaram que sumários produzidos têm melhores eficiência do que o GistSumm e o Mead em termos da medida-F, bem como melhoraram os desempenhos dos sumarizadores GistSumm e Mead, com o uso das relações

CST. Já os resultados da avaliação humana mostraram um bom desempenho nos critérios elegidos, sendo que a informatividade é um dos critérios em que melhor desempenho se obteve, com baixa presença de sentenças redundantes nos sumários finais. Por fim, concluiu que o uso da CST permite explorar o conhecimento entre vários textos que versem sobre um mesmo assunto, o que ajuda na seleção de conteúdo, melhorando a informatividade e coerência (qualidade textual) nos sumários finais (Jorge; Pardo, 2010).

A experiência narrada no trabalho de sumarização com aprendizado de máquina em profundidade debateu as dificuldades a serem enfrentadas na consolidação do tema (Paiola, 2022), sobremaneira, a importância de *corpora* em língua portuguesa para geração de sumários de maior qualidade.

Com fundamento no TeMário e CSTNews, foram realizados experimentos preliminares com modelos de sumarização em inglês, com tradução dos textos do português para o inglês e, posteriormente, dos sumários candidatos em inglês para o português. Também foi utilizado o método GistSumm, em razão da relevância dos trabalhos de sumarização extrativa em português. Foram determinadas três linhas de experimentos: utilizando sumarização monolíngue, multilíngue e entre línguas. Para melhor avaliar os resultados obtidos, foi utilizada a arquitetura do modelo T5 para todos estes experimentos, uma vez que já foi utilizado para experimentos em sumarização abstrativa em inglês obtendo bons resultados, possuir modelos pré-treinados em português e em bases multilíngue, dispensando a necessidade de uma fase de pré-treinamento, o que demandaria mais tempo e recursos.

O treinamento monolíngue, com bases WikiLingua e XL-Sum, que são multilíngues, aproveitou apenas os textos em português, com ajuste nas bases TeMário e CSTNews. O modelo adotado foi o PTT5, pré-treinado, em Python, utilizando os módulos Transformers e PyTorch. Além do treinamento monolíngue, aproveitando o fato das bases WikiLingua e XLSum possuírem textos anotados em diversas línguas, também foi realizado um treinamento multilíngue, apenas na base WikiLingua, com ajuste nas bases TeMário e CSTNews, segundo os mesmos parâmetros da implementação monolíngue. Para fins de comparação, os experimentos monolíngues também foram replicados para o modelo mT5, tanto ao se realizar um ajuste-fino direto no modelo mT5 as bases TeMário e CSTNews, como também ao se realizar um primeiro ajuste-fino monolíngue na WikiLingua e XL-Sum. Também foi experimentado o treinamento entre línguas, utilizando textos das bases TeMário e CSTNews traduzidos para o inglês, realizou-se o ajuste-fino do modelo T5 (Base), que já havia sido avaliado nessas bases nos experimentos com sistemas em inglês.

Os sumários candidatos produzidos pelos modelos foram avaliados pelo conjunto de medidas Rouge, além disso, de BERTScore e o MoverScore. Foram calculadas as médias das taxas de compressão e de abstração dos sumários produzidos por cada modelo, com o objetivo de fornecer pistas sobre as características dos comportamentos de cada um deles.

Na avaliação do experimento, preliminarmente, foram feitas as seguintes considerações: (a) as medidas Rouge tendem a beneficiar sumários candidatos que possuam tamanho similar aos sumários de referência, penalizando sumários com tamanhos diferentes, ainda que estes tenham um alto nível de similaridade semântica com os sumários de referência; (b) em razão da ausência de trabalhos de sumarização abstrativa, houve comparação com sumarizadores extrativos; (c) os resultados obtidos pelos sumarizadores abstrativos são inferiores aos extrativos, tendo em vista a complexidade de geração de sumários baseados em sentenças novas, com preservação do nível semântico; (d) os resultados apontaram que o ajuste-fino do modelo PTT5, pré-treinado em português, obteve melhores resultados que todos os outros modelos, tanto na base TeMário como na base CSTNews, com base na qualidade dos sumários gerados.

Por fim, na conclusão do trabalho, o autor afirma que: (a) modelos de aprendizado em profundidade possibilitam resultados satisfatórios de sumarização abstrativa, mesmo em bases de dados com poucas amostras anotadas; (b) a realização de um primeiro ajuste-fino em outro *corpus* de sumarização com maior quantidade de dados (uma vez que utilizou a TeMário e a CSTNews), seguido pelo ajuste-fino na base em si, tende a obter melhorias nos resultados qualitativos; (c) a utilização de um modelo pré-treinado apenas em português conduziu a resultados superiores à utilização de modelos multilíngue ou mesmo à sumarização entre línguas.

Quadro 3 – Comparação de Trabalhos de Avaliação das Ferramentas de Sumarização Automática sob o Ponto de Vista da Qualidade dos Sumários Produzidos

Referência	**Resultado quanto ao critério de qualidade no sentido amplo: qualidade textual, informatividade ou textualidade**
(Rino, Pardo, 2003)	GistSumm: informatividade e textualidade adequada.
(Muller; Granatyr; Lessing, 2015)	GistSumm (algoritmo *Gist*) em comparação com o algoritmo de Luhn: informatividade adequada.
(Espina; Rino, 2002)	TMSumm: informatividade adequada.
(Tosta, 2014)	Sumarização extrativa profunda multilíngue para gerar extratos em português: informatividade adequada.

Referência	Resultado quanto ao critério de qualidade no sentido amplo: qualidade textual, informatividade ou textualidade
(Nascimento Neto; Gomes; Neto, 2007)	SATSumm, em comparação com o GistSumm: textualidade e informatividade adequada.
(Pardo; Rino, 2002)	DMSumm: informatividade adequada.
(Rocha, 2017)	BLMSumm, GistSumm, e o PragmaSUM com o método original e utilizando palavras-chave com os métodos sequência, classificação e TF-ISF: informatividade adequada.
(Nunes; Souza, 2001)	SUMEX com o AutoResumo do Word, em abordagem de sumarização extrativa para a introdução de um artigo científico. Para as estratégias Palavras-chave, Palavras-chave + Localização houve geração de sumário com informatividade adequada.
(Oliveira, 2008).	Sistemas extrativos GistSumm, NeuralSumm, AutoResumo e TextAnalyst, em textos jornalísticos e científicos: GistSumm (informatividade inadequadas textos científicos; NeuralSumm (informatividade inadequada em textos jornalísticos e científicos); AutoResumo e TextAnalyst informatividade adequada para textos jornalísticos e científicos).
(Jorge; Pardo, 2010).	CSTSumm: informatividade adequada.
(Paiola, 2022)	Sumarização com aprendizado de máquina em profundidade, com corpus em língua portuguesa: informatividade adequada, com os resultados inferiores dos sumarizadores abstrativos em comparação com extrativos.

Fonte: elaborado pelo autor (2022)

4.3 Experiência do LegalSumm como sumarizador de texto jurídico

O experimento relativo ao LegalSumm optou pela utilização e avaliação de técnicas de sumarização extrativa e abstrativa sobre material jurídico. Para análise extrativa, foram adotados métodos de Luhn, LexRank, LSA, KLSum, SumBasic e TextRank, com a composição do sumário a partir das frases com maior pontuação. Para análise abstrativa, foi referenciado modelo de rede neural usando o pacote OpenNMT-tf. Para comparar a eficácia com outros sistemas de sumarização (BertSumExt, BertSumAbs e Bart), o autor disponibilizou dois experimentos usando o conjunto de dados do *corpus* RulingBR.

O primeiro experimento sugeriu que o BertSumAbs e Bart produziram resumos competitivos, enquanto o BertSumEx se mostrou deficiente. Mesmo assim, os resultados atestaram que o LegalSumm superou, ou igualou, as pontuações da Rouge-F para todas as métricas. Além disso, o LegalSumm também apresentou eficiência na comparação com outras

abordagens, com modelos extrativos e abstrativos, em trabalho anterior do próprio autor (Feijó; Moreira, 2018).

No campo da sumarização extrativa, os resultados apresentaram desempenho insatisfatório, sendo incapazes de gerar resumos úteis. O desempenho decepcionante estaria diretamente ligado ao método, uma vez que o tamanho das sentenças e a complexidade do discurso na argumentação jurídica dificultam a obtenção de bons resumos apenas pela junção de frases completas do texto-fonte, com deficiências no binômio qualidade textual e informatividade.

No que toca à sumarização abstrativa, houve melhores resultados, uma vez que foram capazes de cobrir corretamente os tópicos principais do texto-fonte (informatividade), não obstante, apresentou deficiência semântica, repetição de expressões e introdução de assuntos fora do contexto com interferência na textualidade. Em geral, as abordagens abstrativas se mostraram mais promissoras se cotejadas com os resumos produzidos pelos seres humanos.

O ambiente, embora promissor, demanda maior amadurecimento, tendo em vista que nenhum dos resultados obtidos poderia substituir sumários produzidos pelos seres humanos. Positivamente, os experimentos poderiam ser utilizados para gerar rascunhos submetidos à supervisão dos seres humanos, como instrumento de auxílio.

A segunda experiência apreciou a capacidade do LegalSumm de melhorar a qualidade dos resumos gerados pelos modelos padrão *Transformer*, por meio de uma pesquisa com os seguintes objetos: (i) quão útil é o módulo de vinculação do LegalSumm; (ii) qual o desempenho apresentado pelo LegalSumm se comparado com as linhas de base para sumarização de texto; (iii) como os juristas avaliaram a qualidade dos resumos gerados automaticamente; (iv) como o número de dados falsos afeta os resultados no LegalSumm.

Em respostas aos questionamentos apontados acima, o autor esclareceu que (i) que o LegalSumm ajudava a melhorar a qualidade dos resumos gerados; (ii) para questão estabelecida, cotejou o LegalSumm com outras bases, incluindo BertSumExt, BertSumAbs (Lapata; Liu, 2019) e Bart (Lewis *et al.*, 2020), com resultados superiores; em (iii) comparou o LegalSumm aos resultados já divulgados no estudo (Feijó; Moreira, 2018) sobre o mesmo conjunto de dados, em adição, requereu a participação de especialistas jurídicos que avaliaram os sumários gerados em termos de cobertura, coerência, fidelidade e substitutibilidade, concluindo que o experimento superou ou igualou todos na métrica Rouge-F. Por fim, em

resposta ao item (iv) foi verificado o impacto do número de treinamentos realizados com exemplos falsos sobre o resultado alcançado.

Em especial, para o objetivo do trabalho, sobressaem os resultados apresentados via avaliação por especialistas jurídicos sobre o binômio qualidade textual e informatividade dos sumários gerados automaticamente. Para a pesquisa empreendida foram selecionados onze voluntários com formação na área jurídica, com mais de dez anos de experiência no trato com decisões judiciais, que deveriam avaliar casos selecionados aleatoriamente, restrito a dez iniciativas, em razão da complexidade do experimento.

Para a hipótese, foram apresentados: (i) o texto completo da decisão, (ii) uma versão resumo original de comparação (criado manualmente), (iii) o resumo gerado por BertSumAbs, e (iii) o resumo produzido por LegalSumm. Especificamente, foi escolhido o BerSumAbs, como padrão de comparação, porque obteve os maiores escores Rouge-F nas linhas utilizadas.

Cada sumário gerado automaticamente era submetido para avaliação de quatro aspectos cruciais, devendo ser atribuída pontuação de um a cinco, sendo certo que, na régua de comparação, a pontuação um (1) equivaleria a uma forte discordância, enquanto cinco (5) representava forte concordância, sobre os seguintes tópicos: (a) o resumo abrange partes importantes do texto; (b) o resumo apresenta um fluxo coerente; (c) o sumário é fiel aos fatos e não introduz fatos estranhos; (d) o sumário automático pode substituir o resumo original, criado manualmente pelo ser humano. De acordo com o conteúdo apresentado, a hipótese (a) trata de aferição de informatividade do sumário, enquanto as hipóteses (b) e (c) estão inseridas no campo da aferição de qualidade textual. Por fim, o item (d) encontra balizamento da própria ideia de qualidade em sentido amplo.

Na comparação com o BertSumAbs, o LegalSumm se mostrou mais eficiente em todos os quatro aspectos levantados, em especial, foi duas vezes melhor no que toca a coerência e fidelidade, sendo três vezes maior quando o tema era substitutibilidade.

No entanto, no que se refere a possibilidade do uso da sumarização automática em substituição ao resumo criado manualmente pelo ser humano, mais da metade das respostas apresentou avaliação negativa quanto à fidelidade dos fatos apresentados, indiciada a falha no binômio qualidade textual e informatividade, afetando o padrão de qualidade em sentido amplo.

Nesse passo, em que pese o processo contínuo de melhoria, o autor concluiu que falta aos sumários automatizados apresentados, no momento, confiabilidade necessária para utilização de forma independente, sem a participação humana. Em perspectiva positiva, os

resumos gerados pelo LegalSumm podem ser usados como rascunho, submetido aos profissionais jurídicos, como forma de aliviar a carga de trabalho.

Em complemento, concluiu que o experimento usando modelos baseados em Bert demostrou a possibilidade de treinamento e ajuste para língua portuguesa, com melhoria da qualidade no domínio jurídico.

Na prospecção de experimento futuro, entendeu ser possível a substituição da arquitetura *Transformer* por um sumarizador baseado em Bart, no módulo de sumarização, com a perspectiva de lidar com documentos mais longos. Outrossim, seria adequado adaptar os estudos para outros documentos legais de caráter mais geral, em comparação com os acórdãos, que dispõem de estrutura interna mais específica.

Indicou, ainda, preocupação na melhoria do treinamento em língua portuguesa, a partir de dados mais refinados, que viabilizassem detecção de erros, com a substituição ou remoção nos sumários. Bem assim, discutiu a possibilidade de tentar um treinamento que favorecesse estilo de escrita distinta, em função da diversidade de abordagem dos juízes, focando no conteúdo, uma vez que o experimento atual buscava identificar um tipo neutro, por meio da mistura de conteúdo e estilo de escrita.

4.4 Conclusões sobre a qualidade dos sumários produzidos

O estado da arte em sumarização automática de texto denotou a existência de um tema em plena evolução, com experimentos variados, com métodos e resultados diversos: abordagem extrativa e abstrativa, mono e multidocumento, de caráter superficial ou profunda, mono ou multilíngue.

Quanto ao idioma de apoio para sumarização automática de texto, na linha citada por Paiola (2022), o desenvolvimento de *corpora* na língua portuguesa favorecerá experiências com sumarização no campo nacional. Atualmente há prevalência da língua inglesa em comparação com o conjunto nacional, sendo certo que o custo envolvido na formação de *corpora* constitui fator limitante.

A maioria dos experimentos em língua portuguesa remetem à abordagem extrativa, com o uso crescente de análise profunda em razão da preocupação com a qualidade dos sumários gerados. A abordagem abstrativa, que surge exatamente da necessidade de suprir as deficiências, com maior preocupação de semântica e coesão, tende a gerar sumários mais

robustos, com conteúdo mais complexo, embora apresente poucas experiências em língua portuguesa e necessite de uma ponderação sobre os custos da atividade computacional.

Os experimentos iniciais utilizavam um documento de entrada (monodocumental), tendo sido constatado evolução da abordagem multidocumento que, a longo prazo, pode ser capaz de auxiliar conhecimentos interdisciplinares, embora a ausência de tecnologia e os custos de atividade computacional imponham barreira ao seu desenvolvimento, no panorama atual.

A abordagem profunda, privilegiando a manipulação e o uso de conhecimento linguístico avançado, tende a gerar sumários mais qualificados do ponto de vista da textualidade (qualidade textual e informatividade), em consonância com o entendimento exposto por Antunes (2018).

Em que pese a dificuldade de analisar a qualidade em sentido amplo, dada a complexidade da tarefa, na linha defendida por Rino e Pardo (2003), os resultados expostos nos diversos experimentos de sumarização automática de textos sugeriram que é possível gerar sumários adequados, sob o ponto de vista da textualidade, qualidade textual e informatividade, admitidos certos padrões de defeitos: repetição de termos, erros gramaticais, redundância, incoerência e outros.

A avaliação perpetrada pelos juízes humanos ainda é capaz de gerar um resultado mais efetivo, sob o ponto de vista da qualidade em sentido amplo, uma vez que textualidade influência no interesse do leitor. De outro lado, os métodos automáticos têm mostrado evolução, com destaque para métrica Rouge, com resultados satisfatórios e uso bem difundido.

Outrossim, a taxa de compressão está diretamente ligada à informatividade dos sumários, capaz de influenciar nos níveis de qualidade, com extratos curtos apresentando resultados mais coerentes que os extratos longos, justamente porque a menor incidência de palavras aponta para redução da chance de erros gramaticais, redundância e incoerência, conforme defendido por Rino e Pardo (2003). Em contrapartida, em termos de informatividade, extratos mais reduzidos parecem gerar resultados piores que os longos, em razão da ausência de espaço suficiente para detalhamento do conteúdo, com deficiência da informação cotejada com o texto-fonte.

É possível a utilização da sumarização automática em texto jurídico, como analisado na experiência do LegalSumm, em que pese a complexidade da linguagem e da estrutura, em caminho que começa a ser trilhado. Ademais, é viável que os operadores do Direito recorram

aos instrumentos tecnológicos para sumarização, seja via ferramenta preordenada ou códigos de programação, como mecanismo de auxílio no combate à sobrecarga de informação jurídica.

5 EXPERIMENTO DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTO JURÍDICO

No ponto atual, já restou esclarecido o problema da sobrecarga de informação jurídica pelo aumento exponencial de dados, cujo enfrentamento passa pela mudança cultural do mundo jurídico, no sentido de somar forças com outras disciplinas, com adaptação da atividade profissional, propiciando solução de fenômenos sociais complexos pelo acolhimento de ferramentas tecnológicas.

Uma vez percorrido o caminho da sumarização automática como ferramenta tecnológica disponível para embate da sobrecarga de informações, é oportuno apresentar a sumarização como experimento, com foco em textos jurídicos, e analisar os resultados no plano definido, em especial, se as iniciativas atendem aos padrões da língua portuguesa, são acessíveis aos operadores do Direito, e se a qualidade dos sumários gerados é adequada.

Notadamente, foi evidenciada a dificuldade de compreensão e processamento da linguagem humana, além da complexidade da avaliação sob o ponto de vista da qualidade em sentido amplo dos sumários, passando pelo diagnóstico do estado atual na geração de sumários de interesse do leitor e pelas peculiaridades dos textos jurídicos, viabilizando questionamento acerca da possibilidade de substituição dos sumários produzidos, pelos operadores do Direito, por mecanismos tecnológicos.

5.1 Metodologia do experimento de sumarização com texto jurídico a partir de acórdãos de julgamento

Para o experimento de sumarização automática de texto jurídico, foram adotados os seguintes padrões: (1) uso plataforma online e gratuita, com programação preconcebida para sumarização automática, com opção para trabalhos em português, escolhida de forma aleatória; (2) em um segundo momento, utilização de mecanismo de Processamento de Linguagem Natural, por meio do Python, linguagem de programação modular, dinâmica e interpretação orientada a objetos (Python Software Foundation, 2020), com o uso de bibliotecas, conjunto de coleções de programações preordenadas para o fim perquirido.

A utilização da plataforma online e gratuita teve por fundamento evidenciar a viabilidade de operadores do Direito se valerem de ferramentas disponíveis, de fácil acesso,

sem maiores conhecimentos em programação. No caso, foi escolhido o *Summarizer* (AISummarizer), ferramenta que possui opção para sumarização de texto em português (Resumidor de Texto, 2023), estabelecendo como principais funcionalidades: definir comprimento do resumo; mostrar palavras, expressões ou termos principais em tópicos; apresentar a melhor linha; o artigo original com preservação do significado; compreensão básica. Com auxílio de inteligência artificial, utiliza abordagem extrativa para detectar, inicialmente, as melhores frases do parágrafo, prosseguindo com o ranqueamento com pontuação das sentenças com base em precisão, estrutura, otimização, qualidade, comprimento e outros fatores técnicos. Posteriormente, há comparação das melhores sentenças com as informações relacionadas ao restante do texto. Por fim, apresenta um resumo de acordo com o conhecimento do texto-fonte.

Para iniciativa, foi realizada a extração ou conversão do arquivo original do acórdão de *portable document format* (PDF) para word, com adoção da taxa de compressão, métrica para medir a dimensão da condensação do conteúdo expresso no resumo, em 85%, seguindo a observação apresentada nos estudos referenciados, com parâmetros de 80-90%, para trilhar um caminho de qualidade textual e informatividade (Rino; Pardo, 2003).

No outro campo, a predileção pelo Python, linguagem de programação modular, encontra justificativa nos conceitos simplificados de código e sintaxe, de conhecimento acessível, com viabilidade de utilização pelos operadores do Direito, possibilitando a geração de sumários mais personalizados, com maior entendimento do processo, necessários para adaptação da cultura jurídica.

A partir de um texto-fonte em *portable document format* (PDF), foi realizada extração e manipulação com PYPDF2; o módulo Re, para operações com expressões regulares; o Natural Language Toolkit (NLTK), conjunto de bibliotecas para Processamento de Linguagem Natural; o PDFmine.six, para mineração de dados; bem assim, o NumPy para possibilitar maior facilidade de processamento de dados em razão das funções numéricas.

Foram aplicadas ferramentas para leitura de *corpus*, pré-processamento, com tratamento da pontuação, formatação do tamanho da tabulação e remoção de palavras desnecessárias, trabalhando com a frequência de palavras do texto para geração de lista de palavras e lista de sentenças, fazendo um ranqueamento das melhores sentenças, com a geração, posteriormente, de um primeiro resumo pela junção das sentenças selecionadas.

Em prosseguimento, sobre o primeiro resumo, houve aplicação da biblioteca Sumy (Sumy, 2015), bastante difundida nas iniciativas de sumarização, adotando abordagem extrativa, com funcionalidade para língua portuguesa, uma vez que inicialmente foi pensada para o idioma tcheco/eslovaco. Em adição, possibilita a escolha de vários algoritmos para tarefas sumarização, tendo sido eleito o de Luhn para que fosse possível ilustrar uma cena clássica de geração de sumário.

O texto jurídico escolhido para o objeto de experimento foi o acórdão, peça jurídica que representa o resultado de um julgamento realizado por um grupo de magistrados (Ministros), um colegiado. A peça é composta por ementa (síntese do acórdão, com os pontos fundamentais), relatório (descrição dos fatos do processo e o direito discutido), motivação (fundamentação da decisão, construída com base em fatos e direito aplicado, com exposições das razões do julgador) e o dispositivo (parte que se traduz em decisão final), tendo sido eleito pelo fato de constituir um desafio gerar sumários coerentes dada a complexidade do conteúdo do texto-fonte.

De modo a possibilitar a comparação de eventuais resultados díspares, partindo de objetos semelhantes, bem assim a complexidade do conteúdo, foram selecionados três acórdãos: Acórdão de Julgamento em REsp 1842613-SP, perante o Superior Tribunal de Justiça – STJ; Acórdão de Julgamento em RE 654833-AC, perante o STF; Acórdão de Julgamento em REsp 1.846.649-MA perante o STJ, observados os critérios: (1) extensão, representativo da *big data*; (2) relevância, considerando para tal desiderato os divulgados nos portais de notícias de sítios especializados em conteúdo jurídico; (3) tocassem campos diversos do Direito, como, por exemplo, direito penal, processual penal e civil; constitucional e ambiental; consumidor e civil.

Por fim, o que toca ao tema avaliação, com foco na realizada pelo ser humano, também conhecida como manual, foram aproveitados os resumos apresentados nos sítios de portais que noticiando matéria jurídica, uma vez que construídos por especialistas humanos (sumário profissional) como referência de comparação.

Em complemento, os resultados foram submetidos ao julgamento de especialista jurídicos, isto é, juízes humanos, em adaptação ao proposto no trabalho do LegalSumm, item 5.3 do trabalho. Nesse passo, foram selecionados onze voluntários com formação na área jurídica, com mais de dez anos de experiência no trato com decisões judiciais, que deveriam responder aos quesitos direcionados para o campo da textualidade, visando perquirir a qualidade dos sumários: coerência, coesão e informatividade (Apêndice D).

De modo a possibilitar a geração de indicadores, identificando tendência, por meio de uma análise quantitativa, foi apresentado um questionário com 13 (treze) perguntas fechadas, sobre a boa qualidade em sentido amplo (coerência, coesão e informatividade), a utilização dos sumários gerados, como material de apoio para trabalho e, derradeiramente, eventual substituição da sumarização humana por mecanismo automatizado, no panorama atual. Foram abordados os seguintes questionamentos:

(1) Analisando o conteúdo do texto-fonte Acórdão de Julgamento em REsp 1842613-SP, perante o STJ, é possível concluir que o sumário A possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?;

(2) Analisando o conteúdo do texto-fonte Acórdão de Julgamento em REsp 1842613-SP, perante o STJ, é possível concluir que o sumário B possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?;

(3) Analisando o conteúdo do texto-fonte Acórdão de Julgamento em RE 654833-AC, perante o STF, é possível concluir que o sumário C possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?;

(4) Analisando o conteúdo do texto-fonte Acórdão de Julgamento em RE 654833-AC, perante o STF, é possível concluir que o sumário D possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?;

(5) Analisando o conteúdo do texto-fonte Acórdão de Julgamento em REsp 1.846.649-MA perante o STJ, é possível concluir que o sumário E possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?;

(6) Analisando o conteúdo do texto-fonte Acórdão de Julgamento em REsp 1.846.649-MA perante o STJ, é possível concluir que o sumário F possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?;

(7) Como operador do Direito, você utilizaria o sumário A como material de apoio para seu trabalho?;

(8) Como operador do Direito, você utilizaria o sumário B como material de apoio para seu trabalho?;

(9) Como operador do Direito, você utilizaria o sumário C como material de apoio para seu trabalho?;

(10) Como operador do Direito, você utilizaria o sumário D como material de apoio para seu trabalho?;

(11) Como operador do Direito, você utilizaria o sumário E como material de apoio para seu trabalho?;

(12) Como operador do Direito, você utilizaria o sumário F como material de apoio para seu trabalho?;

(13) Como operador do Direito, diante dos sumários gerados automaticamente acima expostos, você entende que seria possível substituir, atualmente, pelo artefato tecnológico, o trabalho de sumarização ou resumo realizado pelo ser humano?

Foram utilizados, ainda, como elementos secundários de mineração, buscas de termos específicos sobre os textos do experimento, para detectar ênfase, sentimento ou perfil do julgador, nos votos proferidos.

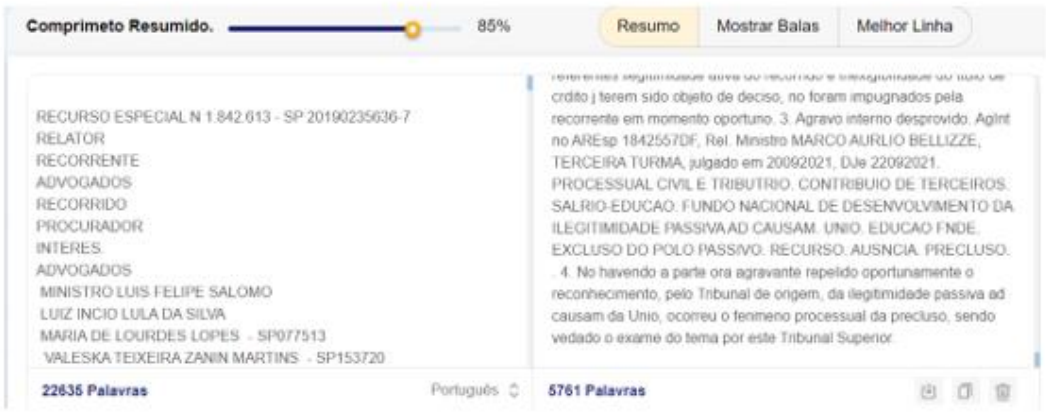
5.2 Sumarização do REsp 1842613-SP perante o Superior Tribunal de Justiça – STJ

O texto-fonte objeto do estudo partiu do Acórdão de Julgamento do REsp 1842613-SP do STJ, julgado pela Quarta Turma, por maioria de votos, em decisão ainda não definitiva, dada recorribilidade no âmbito processual, servindo apenas como referência de tese jurídica, em razão da garantia constitucional da presunção de inocência, diante da ausência do trânsito em julgado.

Cabe pontuar que o recurso especial é uma espécie de ferramenta processual, de natureza extraordinária, de competência do Superior Tribunal de Justiça, com fundamento na Constituição da República Federativa do Brasil de 1988, para revisão dos julgados oriundos dos tribunais inferiores, como forma de controlar a legalidade dos atos praticados diante da legislação federal e da jurisprudência, no que se refere ao erro de aplicação do direito, mantendo a unidade do sistema.

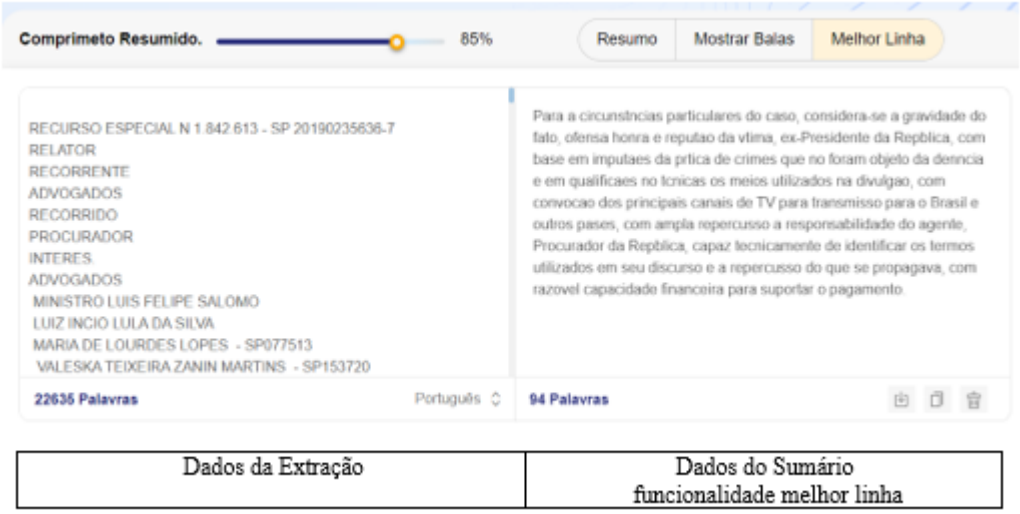
Para iniciativa (figura 1), utilizando o *Summarizer*, foi realizada extração do texto-fonte, inicialmente com 22.635 palavras. O resultado indicou um sumário com 5.761 palavras, de natureza civil e penal (sumário A). Em adição, quando selecionada a funcionalidade “melhor linha”, é possível notar grau de coerência, coesão e informatividade, em resumo com 94 palavras, em que pesem erros gramaticais resultantes do processo de extração do pdf (figura 2):

Figura 1 – Extração do REsp 1842613-SP do STJ e sumário gerado



Fonte: Adaptado pelo autor a partir da interface (AISummarizer, 2022).

Figura 2 – Extração do REsp 1842613-SP do STJ e sumário gerado funcionalidade melhor linha



Fonte: Adaptado pelo autor a partir da interface (AISummarizer, 2022).

O sumário conseguiu estabelecer um padrão adequado na fundamentação empregada: 1) ao identificar o autor do dano, um agente público, Procurador da República, que é capaz tecnicamente de estabelecer tanto os termos utilizados no discurso como a repercussão do caso; 2) a vítima do ato danoso, ex-Presidente da República, que sofreu lesão à sua honra e reputação, pela imputação de crimes e práticas que não foram objeto da denúncia, peça inicial da ação penal pública; 3) e o modo de agir, imputação de crimes não objeto da denúncia, com transmissão para canais de TV para o Brasil e outros países (figura 2):

> Para a circunstncias particulares do caso, considera-se a gravidade do fato, ofensa honra e reputao da vtima, ex-Presidente da Repblica, com base em imputaes da prtica de crimes que no foram objeto da denncia e em qualificaes

> no tcnicas os meios utilizados na divulgao, com convocao dos principais canais de TV para transmisso para o Brasil e outros pases, com ampla repercusso a responsabilidade do agente, Procurador da Repblica, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercusso do que se propagava, com razovel capacidade financeira para suportar o pagamento.

Para um segundo momento, foi realizada extração a partir da manipulação do texto-fonte em formato PDF, via Python, restando 167.607 palavras distribuídas em 59 páginas (Apêndice A). Posteriormente, apresentou-se um primeiro resumo pela junção das sentenças selecionadas, contando com 33.334 palavras. Por fim, tendo sido aplicado o algoritmo de Luhn, foi possível gerar um sumário automático, com apenas 2.058 palavras (sumário B):

> Para a circunstâncias particulares do caso, considera-se a gravidade do fato, ofensa à honra e reputação da vítima, ex-Presidente da República, com base em imputações da prática de crimes que não foram objeto da denúncia e em qualificações não técnicas; os meios utilizados na divulgação, com convocação dos principais canais de TV para transmissão para o Brasil e outros países, com ampla repercussão; a responsabilidade do agente, Procurador da República, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercussão do que se propagava, com razoável capacidade financeira para suportar o pagamento. Para a circunstâncias particulares do caso, considera-se a gravidade do fato, ofensa à honra e reputação da vítima, ex-Presidente da República, com base em imputações da prática de crimes que não foram objeto da denúncia e em qualificações não técnicas; os meios utilizados na divulgação, com convocação dos principais canais de TV para transmissão para o Brasil e outros países, com ampla repercussão; a responsabilidade do agente, Procurador da República, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercussão do que se propagava, com razoável capacidade financeira para suportar o pagamento. _____
> "Além do esquema de corrupção, e denunciado o esquema de lavagem de dinheiro envolvendo o ex-Presidente Lula, o que se constatou foi o repasse de recursos a partir dessa empresa OAS para o ex-Presidente Lula por meio de um upgrade de um apartamento de um imóvel, um triplex no Guarujá, por meio da reforma desse triplex, por meio da decoração desse triplex, e por meio de um contrato de armazenamento de bens pessoais, um contrato milionário firmado para armazenamento, um contrato falso firmado pela OAS como se os bens fossem dela e não do ex-Presidente

É possível verificar tese jurídica consistente em saber se houve excesso por parte do membro do Ministério Público por ocasião de entrevista em que divulgava os termos da denúncia contra o denunciado. No caso, caberia ser avaliado se a divulgação fora dos limites do fato seria capaz de gerar danos morais ao denunciado, sob o prisma do excesso no exercício do direito de informar.

O representativo julgamento foi noticiado em veículos de imprensa, em especial, o *site* do próprio Superior Tribunal de Justiça, servindo de padrão de comparação, via avaliação

humana, possibilitando aferir o padrão de qualidade em sentido amplo. Conforme publicado (Dallagnol pagará R$ 75 mil a Lula por entrevista do Powerpoint, 2022), houve notícia de que “STJ condena ex-procurador Dallagnol a indenizar Lula em R$ 75 mil por entrevista do PowerPoint”. Em continuidade, esclareceu que o ministro relator do recurso especial, Luis Felipe Salomão, explicou que, quando o agente público pratica ato com potencial para se tornar um ilícito civil, sua condição de agente de Estado perde relevância, ainda que a conduta tenha se dado com o uso da condição pública, respondendo pelo fato (ação indenizatória) não o ente público, mas o próprio servidor. Além disso, Dallagnol incorreu em abuso de direito ao caracterizar Lula, durante as falas na entrevista coletiva, como "comandante máximo do esquema de corrupção" e "maestro da organização criminosa", bem como ao anunciar fatos que não faziam parte do objeto da denúncia, concluindo ter havido dano moral contra o ex-presidente.

O entendimento do julgado foi firmado no sentido de que não há espaço para dúvidas de que todos os agentes envolvidos nas delimitadas etapas da persecução penal devem cuidar para que o procedimento não se desvie de fundamentos éticos, assim como trabalhar pela preponderância dos princípios do devido processo legal, contraditório e ampla defesa. Assim, como a peça acusatória deve ser o espelho das investigações, a divulgação fora dos termos revela-se inadequada, evidenciando o abuso de direito na conduta do membro do Ministério Público, gerando dano moral à vítima, passível de sancionamento civil.

Pelas razões expostas, é possível concluir que os sumários A e B apresentaram qualidade em sentido amplo e poderiam ser utilizados como material de apoio para os operadores do Direito.

Por fim, é possível buscar termos específicos que refletem o objeto do julgamento, como, por exemplo, 136 ocorrências da palavra direito, 28 para abuso, 28 para violação, 18 para atuação, 10 para limite, 7 para funcional e 5 para dever, referentes aos temas “abuso de direito”, “limite de atuação” e “violação do dever funcional”, que constituíram objeto do julgamento (Apêndice A).

5.3 Sumarização do RE 654833-AC perante o STF

O texto-fonte objeto do estudo partiu do Acórdão de Julgamento do RE 654833-AC perante o STF, julgado pelo Plenário, que tratava de dano causado por madeireiros na exploração de terras indígenas, no Acre, nos anos 1980, no qual se buscava afastar a tese da

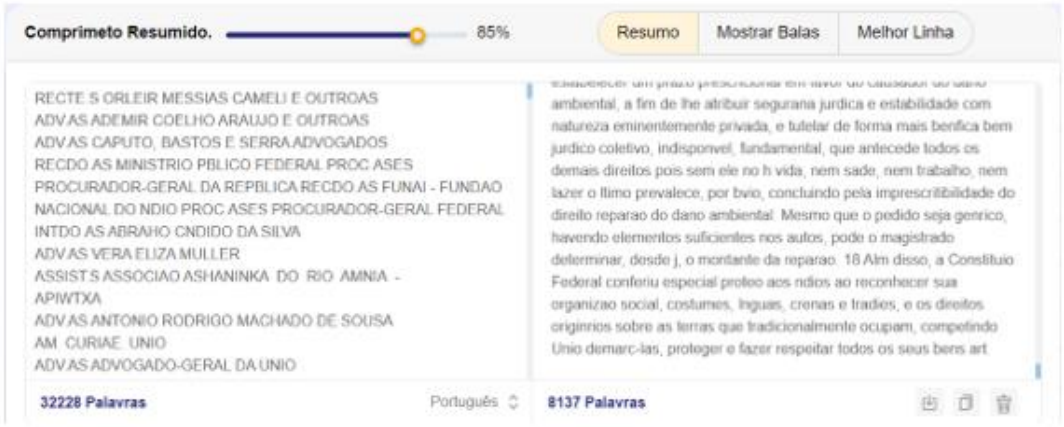
imprescritibilidade, reconhecida em repercussão geral, relativa ao pedido de reparação de dano ambiental (tema 999).

O Recurso Extraordinário (Re) é uma espécie de ferramenta processual, de natureza extraordinária, de competência exclusiva do Supremo Tribunal Federal, intérprete máximo Constituição da República Federativa do Brasil de 1988, para revisão dos julgados oriundos dos tribunais inferiores, quando a decisão: contrariar dispositivo da Constituição; declarar a inconstitucionalidade de tratado ou lei federal; julgar válida lei ou ato de governo local contestado em face da Constituição, julgar válida lei local contestada em face de lei federal. Trata-se, portanto, de verificação de hipóteses de lesão aos fundamentos constitucionais, garantindo-se a higidez do sistema.

Para iniciativa (figura 3) utilizando o *Summarizer* foi realizada extração do texto-fonte, preliminarmente com 32.228 palavras. O resultado indicou um sumário com 8.137 palavras, atentando ainda para um acórdão de 126 páginas, envolvendo debate profundo de matéria ambiental, civil, processual e constitucional, em que pesem erros gramaticais resultantes do processo de extração, devendo destacar os seguintes trechos (sumário C):

> Em nosso ordenamento jurdico, a regra a prescrio da pretenso reparatria. A imprescritibilidade, por sua vez, exceo. Depende, portanto, de fatores externos, que o ordenamento jurdico reputa inderrogveis pelo tempo. Embora a Constituio e as leis ordinrias no disponham acerca do prazo prescricional para a reparao de danos civis ambientais, sendo regra a estipulao de prazo para pretenso ressarcitria, a tutela constitucional a determinados valores impe o reconhecimento de pretenses imprescritveis. O meio ambiente deve ser considerado patrimnio comum de toda humanidade, para a garantia de sua integral proteo, especialmente em relao s geraes futuras. Todas as condutas do Poder Pblico estatal devem ser direcionadas no sentido de integral proteo legislativa interna e de adeso aos pactos e tratados internacionais protetivos desse direito humano fundamental de 3 gerao, para evitar prejuzo da coletividade em face de uma afetao de certo bem recurso natural a uma finalidade individual. A reparao do dano ao meio ambiente direito fundamental indisponvel, sendo imperativo o reconhecimento da imprescritibilidade no que toca recomposio dos danos ambientais.
> [...]
> Havendo prova dos danos e de terem os rus sido os responsveis pelas condutas lesivas, devem ser eles condenados a pagarem as indenizaes correspondentes. irrelevante o fato de o territrio indgena ainda no estar demarcado ao tempo dos fatos, pois as normas constitucionais e legais conferem aos ndios a exclusiva explorao econmica das riquezas naturais existentes nas terras por eles tradicionalmente ocupadas, mesmo que ainda no tenham sido submetidas a demarcao. Ningum pode extrair madeira de imvel pertencente a terceiros indgenas ou no sem a autorizao do seu proprietrio ou legtimo possuidor seja ele conhecido ou no. O montante da indenizao normalmente no se submete a limites mnimo e mximo, tendo como parmetros bsicos a extenso e o valor do dano. Apelaes no providas.

Figura 3 – Extração do RE 654833-AC perante o STF e sumário gerado



Fonte: Adaptado pelo autor a partir da interface (AISummarizer, 2022)

O sumário conseguiu estabelecer um padrão adequado na fundamentação empregada: 1) explicitar que no sistema jurídico, a regra é a prescrição, sendo a imprescritibilidade, exceção; 2) que o meio ambiente é um bem coletivo, transindividual, comum a toda a humanidade; 3) que é valor constitucionalmente tutelado, razão pela qual a reparação ao meio ambiente é um direito fundamental indisponível, sedo imperativo o reconhecimento de sua imprescritibilidade; 4) que estava no contexto de dano ambiental causado na exploração de terras indígenas:

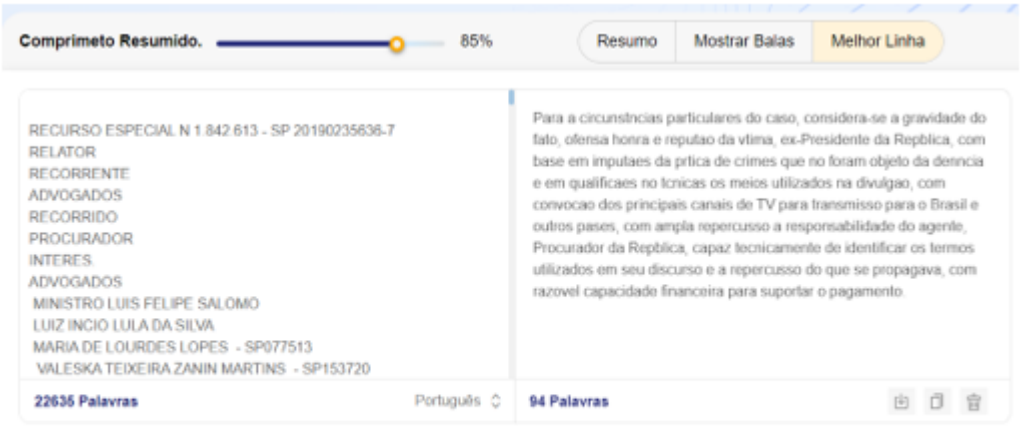
> Em nosso ordenamento jurdico, a regra a prescrio da pretenso reparatria. A imprescritibilidade, por sua vez, exceo. Depende, portanto, de fatores externos, que o ordenamento jurdico reputa inderrogveis pelo tempo. Embora a Constituio e as leis ordinrias no disponham acerca do prazo prescricional para a reparao de danos civis ambientais, sendo regra a estipulao de prazo para pretenso ressarcitria, a tutela constitucional a determinados valores impe o reconhecimento de pretenses imprescritveis. O meio ambiente deve ser considerado patrimnio comum de toda humanidade, para a garantia de sua integral proteo, especialmente em relao s geraes futuras. Todas as condutas do Poder Pblico estatal devem ser direcionadas no sentido de integral proteo legislativa interna e de adeso aos pactos e tratados internacionais protetivos desse direito humano fundamental de 3 gerao, para evitar prejuzo da coletividade em face de uma afetao de certo bem recurso natural a uma finalidade individual. A reparao do dano ao meio ambiente direito fundamental indisponvel, sendo imperativo o reconhecimento da imprescritibilidade no que toca recomposio dos danos ambientais.
> [...]
> Havendo prova dos danos e de terem os rus sido os responsveis pelas condutas lesivas, devem ser eles condenados a pagarem as indenizaes correspondentes. irrelevante o fato de o territrio indgena ainda no estar demarcado ao tempo dos fatos, pois as normas constitucionais e legais conferem aos ndios a exclusiva explorao econmica das riquezas naturais existentes nas terras por eles tradicionalmente ocupadas, mesmo que ainda no tenham sido submetidas a demarcao. Ningum pode extrair madeira de imvel pertencente a terceiros

> indgenas ou no sem a autorizao do seu proprietrio ou legtimo possuidor seja ele conhecido ou no. O montante da indenizao normalmente no se submete a limites mnimo e mximo, tendo como parmetros bsicos a extenso e o valor do dano. Apelaes no providas.

Não obstante, quando selecionada a funcionalidade “melhor linha”, em que pese notar certo grau de coerência, coesão e informatividade, em resumo com 214 palavras, as conclusões expostas, embora tratem de matéria ambiental, não traduzem o objeto como está debatido no acórdão (figura 4):

> 1 Para assegurar a efetividade desse direito, incumbe ao Poder Pblico – preservar e restaurar os processos ecolgicos essenciais e prover o manejo ecolgico das espcies e ecossistemas Regulamento – preservar a diversidade e a integridade do patrimnio gentico do Pas e fiscalizar as entidades dedicadas pesquisa e manipulao de material genticoRegulamento Regulamento Regulamento Regulamento – definir, em todas as unidades da Federao, espaos territoriais e seus componentes a serem especialmente protegidos, sendo a alterao e a supresso permitidas somente atravs de lei, vedada qualquer utilizao que comprometa a integridade dos atributos que justifiquem sua proteo 4 RE 654833 AC Regulamento – exigir, na forma da lei, para instalao de obra ou atividade potencialmente causadora de significativa degradao do meio ambiente, estudo prvio de impacto ambiental, a que se dar publicidade Regulamento – controlar a produo, a comercializao e o emprego de tcnicas, mtodos e substncias que comportem risco para a vida, a qualidade de vida e o meio ambiente Regulamento – promover a educao ambiental em todos os nveis de ensino e a conscientizao pblica para a preservao do meio ambiente – proteger a fauna e a flora, vedadas, na forma da lei, as prticas que coloquem em risco sua funo ecolgica, provoquem a extino de espcies ou submetam os animais a crueldade.

Figura 4 – Extração do RE 654833-AC perante o STF e sumário gerado



Fonte: Adaptado pelo autor a partir da interface (AISummarizer, 2022)

Em continuidade, houve manipulação do texto-fonte em formato PDF, via Python. Realizada a extração, surgiram 507.242 palavras distribuídas em 126 páginas que, no primeiro

resumo de junção das melhores sentenças apresentava 45.701 palavras. Em prosseguimento, aplicado o algoritmo de Luhn, houve geração de um sumário automático com apenas 5.542 palavras, com qualidade textual, embora apresente falhas no que tange aos requisitos da coesão e coerência, pois possui boa informatividade, no que se refere às conclusões do julgado (sumário D). Assim, destacam-se os seguintes momentos (APÊNDICE B):

> De fato, o artigo 2º, inciso I, da Lei nº 6.831/81, ao definir a Política Nacional do Meio Ambiente, dispõe como um de seus princípios a "ação governamental na manutenção do equilíbrio ecológico, considerando o meio ambiente como um patrimônio público a ser necessariamente assegurado e protegido, tendo em vista o uso coletivo".
> [...]
> Ademais, o próprio Fundo de Defesa de Direitos Difusos, previsto pelo artigo 13 da Lei nº 7.347/85 e regulamentado pela Lei nº 9.008/95, encontra-se na estrutura do Ministério da Justiça e tem natureza pública, e é o destino das condenações ao pagamento do dano ambiental, como ocorre no presente caso.
> [...]
> Finalmente, no que toca ao dano ambiental relativo às terras indígenas, ressalte-se que a propriedade das áreas é da União, gerenciado pela Fundação Nacional do Índio e pelas próprias comunidades, o que já atrai o regime de direito público para todas as questões referentes à proteção da área, incluindo questões ressarcitórias dos danos socioambientais às comunidades indígenas. Assim, é imprescritível o dano ambiental, nos termos do artigo 37, §5º da Constituição da República.
> [...]
> O *caput* do artigo 225 da Constituição Federal assim traduz o direito ao meio ambiente: "Todos têm direito ao meio ambiente ecologicamente
> [...]
> São reconhecidos aos índios sua organização social, costumes, línguas, crenças e tradições, e os direitos originários sobre as terras que tradicionalmente ocupam, competindo à União demarcá-las, proteger e fazer respeitar todos os seus bens.

Na síntese, é possível identificar tese jurídica centrada na imprescritibilidade dos danos causados ao meio ambiente, patrimônio público que deve ser assegurado e protegido, tendo em vista a fruição coletiva. Ademais, o regime de direito público atrai questões referentes à proteção da área, bem com as ressarcitórias dos danos socioambientais às comunidades indígenas, reconhecidos aos índios sua organização social, costumes, línguas, crenças e tradições, e os direitos originários sobre as terras que tradicionalmente ocupam, competindo à União demarcá-las, proteger e fazer respeitar todos os seus bens.

A matéria foi objeto de divulgação digital, via *site* do Supremo Tribunal Federal, com o título Especial meio ambiente: ressarcimento por dano ambiental não está sujeito à prescrição (STF, 2023), em que se esclareceu que o recurso, interposto pelos madeireiros, questionava acórdão do Superior Tribunal de Justiça (STJ), que reconheceu a imprescritibilidade do dano

ambiental, tendo sido o tema afetado, sob o prisma da repercussão geral, na relatoria do Ministro Alexandre de Moraes.

No julgamento da repercussão propriamente dita, perante o plenário virtual, o relator, Ministro Alexandre de Moraes, delimitou a controvérsia, ponderando sobre o embate do princípio da segurança jurídica, em benéfico apenas do autor do dano ambiental, diante da inércia do poder público em adotar posturas fiscalizatórias (tese da prescritibilidade da reparação) e do princípio constitucional de proteção ao meio ambiente, que beneficia toda a coletividade (tese da imprescritibilidade).

Em seu voto, identificando o meio ambiente como patrimônio comum da humanidade, as condutas do poder público deveriam ser direcionadas para a integral proteção legislativa interna e internacional, evitando prejuízo da coletividade decorrente do uso de recurso natural para finalidades individuais.

Assim, embora haja previsão de prazo prescricional para a reparação de danos ambientais, como a Constituição Federal expressamente proteger o meio ambiente, como direito fundamental, torna o direito à indenização, nesses casos, imprescritível.

Pelas razões expostas, é possível concluir que os sumários C e D apresentaram qualidade em sentido amplo e poderiam ser utilizados como material de apoio para os operadores do Direito.

Finalmente, a busca por termos específicos referentes ao objeto do julgamento, retornou: 165 ocorrências para o tema ‘direito ambiental’, 38 para ‘direito fundamental’, 20 para ‘coletivo’, 189 para ‘imprescritibilidade’, 14 para ‘terra indígena’. (APÊNDICE B).

5.4 Sumarização do REsp 1.846.649-MA perante o STJ

O terceiro ato consistiu na análise do texto-fonte objeto do Acórdão de Julgamento do REsp 1.846.649-MA perante o STJ, Tema Repetitivo 1061, em que foi debatida a inversão do ônus da prova, impondo à instituição financeira/ré a obrigação de comprovar a autenticidade da assinatura constante do contrato juntado ao processo, quando o consumidor/autor tivesse negado a veracidade da assinatura.

O Recurso Especial foi analisado sob o rito da sistemática de recurso repetitivos, descrita no Código de Processo Civil (Lei n. 13.105/2015), em que a decisão final serve de

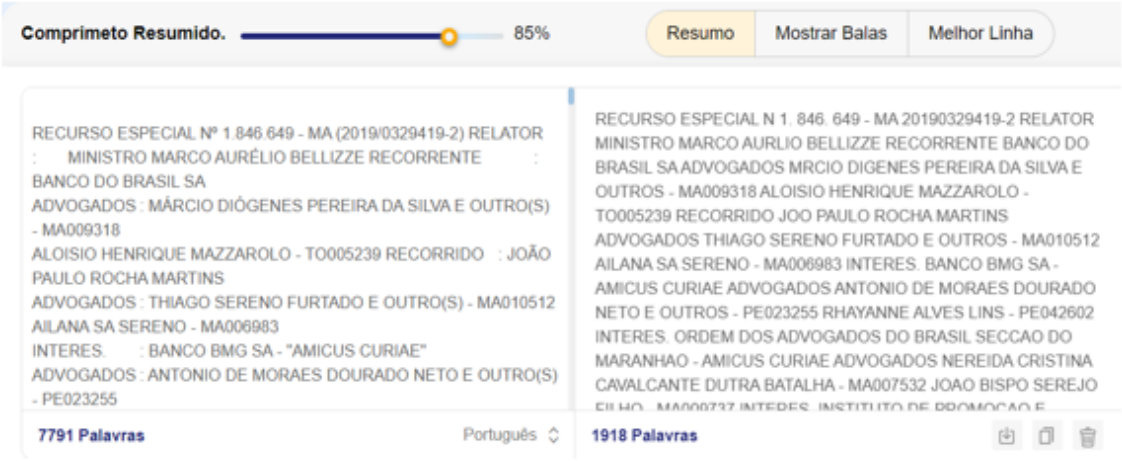
paradigma a ser aplicado aos milhares de processos que tratam do tema, garantindo assim eficiência e a razoável duração, com a entrega da prestação jurisdicional em tempo hábil.

Para iniciativa (figura 5), foi realizada extração do texto-fonte de 7.791 palavras, utilizando o *Summarizer*, que indicou um sumário com 1.918 palavras, envolvendo direito do consumidor e processo civil. O resumo foi capaz de manter, em seu início, o resultado do julgado, com a tese afinal fixada, bem assim, apresentou, em seu corpo, um relatório com as 4 teses delimitadas no texto-fonte (sumário E), que serviram de parâmetro no julgado que foi objeto de recurso especial, revelando, portanto, grau de textualidade, conforme colacionado, a seguir:

> Para os fins do art. 1. 036 do CPC2015, a tese firmada a seguinte Na hiptese em que o consumidorautor impugnar a autenticidade da assinatura constante em contrato bancrio juntado ao processo pela instituio financeira, caber a esta o nus de provar a sua autenticidade CPC, arts. 6, 368 e 429,
> [...]
> Para os fins repetitivos, foi aprovada a seguinte tese Na hiptese em que o consumidorautor impugnar a autenticidade da assinatura constante em contrato bancrio juntado ao processo pela instituio financeira, caber a esta o nus de provar a autenticidade CPC, arts. 6, 369 e 429, II
> [...]
> A primeira tese restou assim fixada Independentemente da inverso do nus da prova – que deve ser decretada apenas nas hipteses autorizadas pelo art. 6 VIII do CDC, segundo avaliao do magistrado no caso concreto –, cabe instituio financeirar, enquanto fato impeditivo e modificativo do direito do consumidorautor CPC, art. 373, II, o nus de provar que houve a contratao do emprstimo consignado, mediante a juntada do contrato ou de outro documento capaz de revelar a manifestao de vontade do consumidor no sentido de firmar o negcio jurdico, permanecendo com o consumidorautor, quando alegar que no recebeu o valor do emprstimo, o dever de colaborar com a Justia CPC, art. 6 e fazer a juntada do seu extrato bancrio, embora este no deva ser considerado, pelo juiz, como documento essencial para a propositura
> [...]
> A segunda tese restou assim fixada A pessoa analfabeta plenamente capaz para os atos da vida civil CC, art. 2 e pode exarar sua manifestao de vontade por quaisquer meios admitidos em direito, no sendo necessria a utilizao de procurao pblica ou de escritura pblica para a contratao de emprstimo consignado, de sorte que eventual vcio existente na contratao do emprstimo deve ser discutido luz das hipteses legais que autorizam a anulao por defeito do negcio jurdico CC, arts. 138, 145, 151, 156, 157 e 158
> [...]
> A terceira tese restou assim fixada cabvel a repetio de indbito em dobro nos casos de emprstimos consignados quando a instituio financeira no conseguir comprovar a validade do contrato celebrado com a parte autora, restando configurada a m-f da instituio, resguardadas as hipteses de enganos justificáveis
> [...]
> A quarta tese restou assim fixada 4. No estando vedada pelo ordenamento jurdico, lcita a contratao de quaisquer modalidades de mtuo financeiro, de modo que, havendo vcio na contratao, sua anulao deve ser discutida luz das

> hipteses legais que versam sobre os defeitos do negcio jurdico CC, arts. 138, 145, 151, 156, 157 e 158 e dos deveres legais de probidade, boa-f CC, art. 422 e de informao adequada e clara sobre os diferentes produtos, especificando corretamente as caractersticas do contrato art. 4 IV e art. 6, III, do CDC, observando-se, todavia, a possibilidade de convalidao do negcio anulvel, segundo os princpios da conservao dos negcios jurdicos CC, art. 170

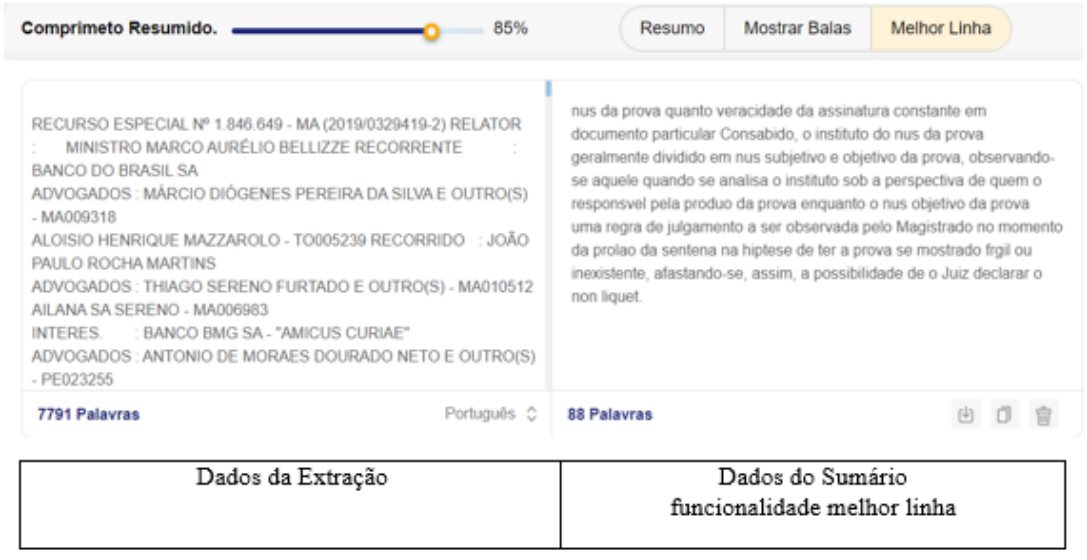
Figura 5 – Extração do REsp 1.846.649-MA perante o STJ e sumário gerado



Fonte: Adaptado pelo autor a partir da interface (AISummarizer, 2022)

Não obstante, utilizando a funcionalidade “melhor linha”, o texto, com 88 palavras, não incorporou o resultado alcançado pelo julgado, tratando apenas de regra acerca do ônus probatório (figura 6):

Figura 6 – Extração do REsp 1.846.649-MA perante o STJ e sumário gerado funcionalidade melhor linha



Fonte: Adaptado pelo autor a partir da interface (AISummarizer, 2022).

> nus da prova quanto veracidade da assinatura constante em documento particular Consabido, o instituto do nus da prova geralmente dividido em nus

> subjetivo e objetivo da prova, observando-se aquele quando se analisa o instituto sob a perspectiva de quem o responsvel pela produo da prova enquanto o nus objetivo da prova uma regra de julgamento a ser observada pelo Magistrado no momento da prolao da sentena na hiptese de ter a prova se mostrado frgil ou inexistente, afastando-se, assim, a possibilidade de o Juiz declarar o non liquet.

No que toca à manipulação do texto-fonte em formato PDF, via Python, foi realizada extração do acórdão com 23 páginas e 58.209 palavras, tratando de orientações no campo do direito do consumidor, cujo microssistema jurídico impõe observância de regras e princípios especiais, além de elementos de direito processual, sobremaneira, no campo da relevância do ônus probatório para solução do conflito. Na fase de pré-elaboração, houve junção das melhores sentenças, apresentadas 26.597 palavras. O resultado a seguir colacionado indicou um sumário final com 1.577 palavras, após aplicação do algoritmo de Luhn (sumário F):

> 2°) e pode exarar sua manifestação de vontade por quaisquer meios admitidos em direito, não sendo necessária a utilização de procuração pública ou de escritura pública para a contratação de empréstimo consignado, de sorte que eventual vício existente na contratação do empréstimo deve ser discutido à luz das hipóteses legais que autorizam a anulação por defeito do negócio jurídico (CC, arts. Embargos declaratórios conhecidos, sendo os 1°ˢ, 2°ˢ, 5°ˢ e 7°ˢ desprovidos; os 4°s embargos parcialmente providos para excluir do acórdão os precedentes deste sodalício de n°ˢ 5499/2016 (Embargos de Declaração) e 18905/2015 (Apelação Cível); e os 3°ˢ, 4°ˢ, 6°ˢ e 8°ˢ parcialmente providos para aclarar a 3ª tese que passará a ter a seguinte redação: "Nos casos de empréstimos consignados, quando restar configurada a inexistência ou invalidade do contrato celebrado entre a financeira e a parte autora, bem como, demonstrada a má-fé da instituição bancária, será cabível a repetição de indébito em dobro, resguardadas as hipóteses de enganos justificáveis". 2°) e pode exarar sua manifestação de vontade por quaisquer meios admitidos em direito, não sendo necessária a utilização de procuração pública ou de escritura pública para a contratação de empréstimo consignado, de sorte que eventual vício existente na contratação do empréstimo deve ser discutido à luz das hipóteses legais que autorizam a anulação por defeito do negócio jurídico (CC, arts.

Conforme se depreende, no resultado não é possível identificar tese jurídica centrada na inversão do ônus probatório, após a negativa de autenticidade de assinatura pelo consumidor/autor, com imposição da obrigação de comprovação da validade pela instituição financeira/ré, muito embora a explicação faça sentido do ponto de vista probatório. Assim, o resultado exibiu qualidade adequada, uma vez que as conclusões do julgado não são as mesmas contidas no sumário (Apêndice C), presente, no caso, vício de coerência e informatividade.

O tema objeto do experimento foi veiculado no *site* Migalhas, com o título “Cabe ao banco comprovar autenticidade de assinatura em contrato” (STJ, 2021), servindo de resumo de

comparação. De acordo com o exposto na notícia, o Ministro Marco Belizze, relator do julgamento, esclareceu que, em geral, compete ao arguidor provar a falsidade de assinatura, segundo as regras do ônus da prova. Não obstante, no microssistema do consumidor, em razão da hipossuficiência, uma vez negada a assinatura do contrato pelo consumidor/autor, compete ao banco/réu comprovar a autenticidade do autógrafo no documento, por meio de perícia grafotécnica ou mediante os meios de prova legais ou moralmente legítimos, firmando-se uma tese paradigma.

Pelas razões expostas, é possível concluir que o sumário E apresenta qualidade em sentido amplo e poderia ser utilizado como material de apoio para os operadores do Direito. Entretanto, no que tange ao sumário F, identificado vício de coerência e informatividade, de maneira geral, padece de qualidade em sentido amplo e não poderia ser utilizado como material de apoio seguro para os operadores do Direito.

Na oportunidade, seguindo a linha de pesquisa, foram efetivadas buscas por termos específicos referentes ao objeto do julgamento: 30 ocorrências para o tema ‘consumidor’, 5 para ‘ônus da prova’, 10 para ‘inversão’, 43 para ‘autenticidade’, 43 para ‘assinatura’ (Apêndice C).

5.5 Avaliação da qualidade dos sumários por especialistas jurídicos

Conforme especificado no item 6.2, os sumários gerados a partir dos experimentos apresentados foram submetidos aos especialistas jurídicos para avaliação da qualidade em sentido amplo (Apêndice D), no campo da textualidade, relativamente à coerência (entender o conteúdo), coesão (entender o relacionamento entre os elementos linguísticos e informatividade (entender se a mensagem apresentada continha o conteúdo necessário).

Os resultados sobre os questionamentos supra-apresentados, acerca da qualidade em sentido amplo dos sumários A, B, C, D, E e F, foram assim respondidos (Apêndice D):

I. Relativamente ao sumário A, experimento com *Summarizer*, itens (1) e (7), para 81,8% dos participantes, o texto resultado possui boa qualidade em sentido amplo, sendo que 81,8% também o utilizaria como material de apoio para o trabalho;
II. Relativamente ao sumário B, experimento via Python com aplicação do algoritmo de Luhn, itens (2) e (8), para 72,7% dos participantes, o texto resultado possui boa qualidade em sentido amplo, sendo que 72,7% também o utilizaria como material de apoio para o trabalho;

III. Relativamente ao sumário C, experimento com *Summarizer*, itens (3) e (9), para 72,7% dos participantes, o texto resultado possui boa qualidade em sentido amplo, sendo que 72,7% também o utilizaria como material de apoio para o trabalho;

IV. Relativamente ao sumário D, experimento via Python com aplicação do algoritmo de Luhn, itens (4) e (10), para 81,8% dos participantes, o texto resultado possui boa qualidade em sentido amplo, sendo que 72,7% também o utilizaria como material de apoio para o trabalho;

V. Relativamente ao sumário E, experimento com *Summarizer*, itens (5) e (11), para 81,8% dos participantes, o sumário E possui boa qualidade em sentido amplo; sendo que 72,7% também o utilizaria como material de apoio para o trabalho;

VI. Relativamente ao sumário F, experimento via Python com aplicação do algoritmo de Luhn, itens (6) e (12), para apenas 45,5% dos participantes, o texto resultado possui boa qualidade em sentido amplo, sendo que 45,5% também o utilizaria como material de apoio para o trabalho;

Finalmente, quanto ao item derradeiro do questionário, para 72,7% dos participantes seria possível substituir, atualmente, o trabalho de sumarização ou resumo realizado pelo ser humano, pelo desempenhado via artefato tecnológico.

Quadro 4 – Comparação de Resultados Avaliação da qualidade dos sumários por especialistas jurídicos

Sumário Gerado	Ferramenta de sumarização	Item do questionário	Resultado quanto ao critério de qualidade no sentido amplo	Utilização como material de apoio para o trabalho
A	*Summarizer*	(1) e (7)	para 81,8% dos participantes, o texto resultado possui boa qualidade em sentido amplo	81,8% utilizariam como material de apoio para o trabalho
B	Python com aplicação do algoritmo de Luhn	(2) e (8)	para 72,7% dos participantes, o texto resultado possui boa qualidade em sentido amplo	72,7% utilizariam como material de apoio para o trabalho

Sumário Gerado	Ferramenta de sumarização	Item do questionário	Resultado quanto ao critério de qualidade no sentido amplo	Utilização como material de apoio para o trabalho
C	*Summarizer*	(3) e (9)	para 72,7% dos participantes, o texto resultado possui boa qualidade em sentido amplo	72,7% utilizariam como material de apoio para o trabalho
D	Python com aplicação do algoritmo de Luhn	(4) e (10)	para 81,8% dos participantes, o texto resultado possui boa qualidade em sentido amplo	72,7% utilizariam como material de apoio para o trabalho
E	*Summarizer*	(5) e (11)	para 81,8% dos participantes, o texto resultado possui boa qualidade em sentido amplo	72,7% utilizariam como material de apoio para o trabalho
F	Python com aplicação do algoritmo de Luhn	(6) e (12)	para 45,5% dos participantes, o texto resultado possui boa qualidade em sentido amplo	45,5% utilizariam como material de apoio para o trabalho

Fonte: elaborado pelo autor (2022).

5.6 Conclusões sobre o experimento com sumarização de texto jurídico

Por meio das atividades realizadas, seja via ferramenta preordenada ou códigos de programação, foram gerados sumários automáticos de textos jurídicos (acórdãos) envolvendo campos diversos do Direito: direito penal, processual penal e civil; constitucional e ambiental; consumidor e civil. De modo a possibilitar a comparação de eventuais resultados díspares, partindo de objetos semelhantes, foram selecionados apenas três acórdãos. Os sumários resultados A, C e E encontraram origem na ferramenta preordenada *Summarizer*, enquanto os resultados B, D e F foram construídos via Python com aplicação do algoritmo de Luhn.

Com foco na avaliação, atentando para coerência, coesão e informatividade, houve comparação dos sumários gerados com os resumos apresentados nos portais de notícias de sítios especializados em matéria jurídica, construídos por especialistas humanos (sumário

profissional), ficando sugerido que é possível trilhar textos que possam preencher os requisitos da textualidade, servindo como material de apoio para os operadores do Direito.

A primeira conclusão foi de que os sumários A, B, C, D e E apresentaram qualidade em sentido amplo e poderiam ser utilizados como material de apoio para os operadores do Direito. Exceção ficou por conta do resultado do sumário F, identificado vício de coerência e informatividade, padecendo de qualidade em sentido amplo e negativa para utilização segura como material de apoio para os operadores do Direito.

Em um segundo momento, os sumários gerados foram submetidos ao julgamento por especialista jurídicos, sobre a qualidade em sentido amplo e utilização como material de apoio para o trabalho, como forma de confirmar as conclusões da primeira avaliação.

De forma geral, a avaliação dos especialistas jurídicos confirmou a primeira conclusão pois decidiu, por maioria, que os sumários A, B, C, D e E apresentaram qualidade em sentido amplo e poderiam ser utilizados como material de apoio para os operadores do Direito. No mesmo passo, a maioria entendeu que o sumário F não apresentou qualidade em sentido amplo, negando sua utilização como material de apoio seguro.

Em ponto derradeiro da avaliação, 72,7% dos participantes entenderam que seria possível substituir, atualmente, o trabalho de sumarização ou resumo realizado pelo ser humano, pelo desempenhado via artefato tecnológico.

Por último, por meio do Python, o experimento propiciou ainda a apresentação de elementos secundários, como a mineração de termos específicos nos textos, para detectar ênfase, sentimento ou perfil do julgador: para o (1) Acórdão de Julgamento do REsp 1842613-SP do STJ foram localizadas 136 ocorrências da palavra direito, 28 para abuso, 28 para violação, 18 para atuação, 10 para limite, 7 para funcional e 5 para dever, referentes aos temas “abuso de direito”, “limite de atuação” e “violação do dever funcional”, que constituíram objeto do julgamento; para o (2) Acórdão de Julgamento do RE 654833-AC perante o STF foram localizadas 165 ocorrências para o tema “direito ambiental”, 38 para “direito fundamental”, 20 para “coletivo”, 189 para “imprescritibilidade”, 14 para “terra indígena”; enquanto para o (3) Acórdão de Julgamento do REsp 1.846.649-MA perante o STJ retornou 30 ocorrências para o tema “consumidor”, 5 para “ônus da prova”, 10 para “inversão”, 43 para “autenticidade”, 43 para “assinatura”.

A busca por termos específicos exemplificou a importância que a mineração de texto pode assumir na análise de julgamentos proferidos pelos tribunais, uma vez que é possível

determinar ênfase, sentimento ou perfil do julgador em relação a determinados temas, trilhando um caminho mais pavimentado dos operadores do Direito em direção ao órgão julgador, na busca de decisões favoráveis

Assim, foi possível analisar ferramentas computacionais em língua portuguesa acessíveis aos operadores do Direito, em caminho a ser trilhado no combate à sobrecarga de informação jurídica.

Em que pese a complexidade envolvida na produção de sumários a partir de acórdãos, foi possível gerar conteúdo de qualidade para apoio aos operadores do Direito, não obstante eventuais erros gramaticais, repetição de conteúdo, vícios de coesão, coerência e informatividade.

CONCLUSÕES E NOVAS PESQUISAS

A grande integração representada pela globalização, impulsionada pela amplificação tecnológica, foi capaz de elevar o nível de informação disponível. O desenvolvimento da Computação, com instrumentos tecnológicos, programas e processos automatizados, cada vez mais digital, foi gerando uma enorme massa de dados, de origem pública ou privada, chamada de *big data*, com grande oportunidade para a pesquisa.

Todo esse movimento provocou sobrecarga da informação, com alteração do panorama cultural jurídico, por meio da inserção dos novos conceitos e preocupações, com impacto no desempenho das atividades dos operadores do Direito, com reflexos positivos, como melhoramento dos processos produtivos, e negativos, como contaminação da relação entre trabalho e tempo disponível, que inclui não apenas o descanso, mas as oportunidades para aperfeiçoamento da formação profissional, provocando um desequilíbrio a ser superado. A solidificação do sistema jurídico, a estrutura e extensão dos seus documentos, estilo da produção textual necessária para validade do ato jurídico, a complexidade da língua portuguesa e outros, refletem desafios adicionais a serem enfrentados.

Assim, o Direito deve participar do momento histórico de ligação com os recursos oriundos da Computação, possibilitando o alcance de novos patamares do conhecimento, no campo das Humanidades Digitais, na resolução de um fenômeno social complexo.

A análise do grande manancial de dados jurídicos, mormente, pelo registro de informações na forma escrita, diante da necessidade de apresentação de razões escritas como forma de justificar decisões judiciais, administrativas ou legislativas no ambiente do Estado de Direito, lança luz sobre ferramentas tecnológicas voltadas para a mineração dos textos, em especial a sumarização automática.

O processo busca apoio em ferramentas que possam entender e manipular um *corpus*, documento acerca de determinado tema, cujo desenvolvimento tem favorecido a sumarização automática, reforçando a necessidade de material especializado em língua portuguesa, como parte integrante da evolução no âmbito nacional. De relevo citar que há diversas iniciativas nesse plano, de caráter interdisciplinar, referenciando, por exemplo, o *corpus* RulinBR, formado a partir de decisões judiciais em língua portuguesa.

O ato de sumarizar faz parte da essência do ser humano, que armazena apenas os acontecimentos mais importantes, em trabalho que envolve interpretação, seleção de informações relevantes e junção dos conteúdos, de forma simplificada. O sistema de sumarização automática está estruturado nas etapas de análise, transformação e síntese. A sumarização extrativa resulta da seleção das sentenças mais importantes, relevantes ou representativas do conteúdo do texto-fonte, apresentando-se como uma atividade menos complexa, de justaposição ou junção de partes selecionadas. A sumarização abstrativa, por sua vez, acaba por produzir alterações, com novas sentenças, que tem por objetivo melhorar a coerência do resultado do texto-fonte reduzido, de conteúdo mais complexo. A abordagem profunda está conectada com o conhecimento linguístico, enquanto a abordagem superficial é fundamentada em dados estatísticos, com pouca preocupação com a complexidade da compreensão do conteúdo.

O estado da arte revelou que os primeiros estudos, que apareceram no final da década de 1950, estavam ligados ao critério da relevância de palavras, sentenças ou partes do texto-fonte, isto é, métodos extrativos. Modernamente, foi introduzida a utilização do aprendizado de máquina, métodos baseados em grafos e redes neurais, em movimento de evolução e franca expansão, sobremaneira, em métodos abstrativos.

A avaliação dos sistemas de sumarização é um tema complexo, que pode ser realizado de forma manual, via ser humano, ou por meio de mecanismos automáticos ou semiautomáticos, importando na verificação de quais são as métricas que melhor definem o sumário gerado.

Em especial, para o objetivo do desenvolvimento neste estudo, interessou investigar a preocupação com a qualidade em sentido amplo dos sumários gerados, no campo da textualidade, contemplada no binômio qualidade textual e informatividade (Pardo, 2008), em razão do grau do interesse que textos provocam nos leitores.

Importante lembrar que não há uniformidade no tratamento do tema, sendo certo que alguns trabalham com o conceito de textualidade, outros se referindo aos componentes qualidade textual (coerência e coesão) e informatividade. Por fim, a avaliação de qualidade, em boa proporção, está inserida no campo da subjetividade do julgador.

As ferramentas e trabalhos de sumarização automática de texto relacionados, especialmente utilizando *corpus* em língua portuguesa, indiciaram, em primeiro plano, a relevância do tema no contexto atual de sobrecarga da informação. Há diversidade de

abordagens: experiências extrativas e abstrativas, mono e multidocumento, de caráter mono, multilíngue ou entre línguas.

Comparando os estudos apresentados, sob o ponto de vista da qualidade em sentido amplo, é possível concluir que os experimentos, tanto envolvendo abordagem extrativa quanto abstrativa, foram capazes de gerar sumários com razoável nível de adequação, viabilizando o entendimento do texto-fonte a partir do resumo. Não obstante, abordagens abstrativas e profundas tendem a gerar sumários mais sólidos, com conteúdo mais complexos, embora importem maior custo computacional. Bem assim, a abordagem multidocumento tende a ser, a longo prazo, um instrumento potente de auxílio, notadamente, nos conhecimentos interdisciplinares.

A análise do LegalSumm, como sumarizador de texto jurídico, revelou a existência de iniciativa voltada para auxílio dos operadores do Direito na elaboração de resumos de decisões judiciais (acórdãos) em língua portuguesa. O estudo, que inclui a capacidade de trabalhar com documentos longos, divididos em seções predefinidas, por exemplo, os acórdãos, tem o mérito de aplicar o treinamento em língua portuguesa, por meio da construção de um *corpus* jurídico especializado, o RulingBR, gerando resumos com maior grau de semelhança com o comportamento humano, adotando um viés neutro de discurso.

Os resultados apresentados sugeriram que, no campo da sumarização extrativa, a qualidade em sentido amplo foi insatisfatória, pelo tamanho das sentenças e a complexidade do discurso na argumentação jurídica, que dificultaram a obtenção de bons resumos pela junção de frases completas do texto-fonte. No que toca à sumarização abstrativa, foi capaz de cobrir corretamente os tópicos principais do texto-fonte (preservando a informatividade). Entretanto, apresentou deficiência semântica, redundância e introdução de assuntos fora do contexto, comprometendo a aferição da textualidade.

Em uma segunda experiência abordada no referido estudo, restou apreciada a capacidade do LegalSumm de melhorar a qualidade dos resumos gerados pelos modelos padrão Transformer. Na comparação com o BertSumAbs, o LegalSumm mostrou-se significativamente mais eficiente, sendo duas vezes melhor em coerência e fidelidade, e três vezes superior quando o tema era substitutibilidade. No entanto, no que se refere à possibilidade do uso da sumarização automática em substituição ao resumo criado manualmente pelo ser humano, mais da metade das respostas pelos juízes humanos apresentou avaliação negativa quanto à fidelidade dos fatos apresentados, ficando latente a falha na qualidade (sentido amplo), concluindo que os resumos

gerados pelo LegalSumm poderiam ser usados como rascunho, submetido aos profissionais jurídicos, como forma de aliviar a carga de trabalho.

O experimento de sumarização automática utilizando texto jurídico, via ferramenta preconcebida ou programação por meio do Python, sugere que já há mecanismos tecnológicos, em língua portuguesa, disponíveis aos operadores do Direito para auxílio no combate à sobrecarga de informação, capazes de gerar sumários de qualidade em sentido amplo e resumos de apoio ao trabalho, em que pese a complexidade da sumarização de acórdãos.

De forma geral, a avaliação dos especialistas jurídicos confirmou a primeira conclusão, por maioria, de que os sumários possuem qualidade em sentido amplo e poderiam ser utilizados como material de apoio para os operadores do Direito, ainda que constatado o resultado negativo apresentado no sumário F.

Em adição, embora a avaliação humana possa envolver aspectos subjetivos de julgamento, 72,7% dos participantes entenderam que seria possível substituir, atualmente, o trabalho de sumarização ou resumo realizado por humanos pelo desempenhado via artefatos tecnológicos.

A iniciativa de mineração de texto ainda surpreendeu pela possibilidade de selecionar determinados termos, de forma a oportunizar a construção de teses jurídicas segundo ênfase, sentimento ou perfil do julgador, em caminho de melhor comunicação dos operadores do Direito com o órgão julgador, que importa melhor qualidade do trabalho desenvolvido e maior chance de êxito na demanda.

Numa perspectiva de trabalho futuro, seria interessante explorar a sumarização multidocumento voltada para experiências interdisciplinares no campo das Humanidades Digitais. Outrossim, deve ser vislumbrada oportunidade de produção de sumários jurídicos envolvendo as diversas áreas de estudo do Direito, bem assim a seleção e construção de teses jurídicas com ênfase nos termos dos sumários gerados. A abordagem, comparativamente mais complexa, com diversidade de tópicos e extensão dos textos, representa uma estrada interessante no caminho para produção de peças e decisões judiciais automatizadas.

Há horizonte para investigação de técnicas com maior profundidade linguística, com exploração das características semânticas e estruturas textuais, redução de redundâncias e incoerências, buscando formas de gerar sumários mais qualificados, com vistas à substituição do sumário humano.

REFERÊNCIAS

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. *In*: THE 1993 ACM SIGMOD INTERNATIONAL CONFERENCE, 1993. **Proceedings of the 1993 ACM SIGMOD international conference on Management of data – SIGMOD '93** [...]. Washington, D.C., United States: ACM Press, 1993. p. 207-216. Disponível em: http://portal.acm.org/citation.cfm?doid=170035.170072. Acesso em: 26 dez. 2022.

ALEIXO, P.; PARDO, T. A. S. **CSTNews:** um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-document Structure Theory). São Carlos: USP; UFSCar; UNESP, 2008. (Série Relatórios Técnicos do Núcleo Interinstitucional de Linguística Computacional – NILC).

AMARAL, Fernando. **Aprenda mineração de dados:** teoria e prática. Rio de Janeiro: Alta Books, 2016.

AMO, S. **Técnicas de mineração de dados**. [S. l.], 24 dez. 2022.

ANTONITISCH, André *et al.* Summ-it++:anenrichedversionoftheSumm-itcorpus. PUCRS University. Porto Alegre, 2016. Disponível em: https://repositorio.pucrs.br/dspace/handle/10923/14015. Acesso em: 20 nov. 2023.

ANTUNES, J. B. Uma abordagem para sumarização automática semi-extrativa. 12 nov. 2018. [doctoralThesis]. Disponível em: https://repositorio.ufpe.br/handle/123456789/33305. Acesso em: 11 jan. 2023.

BARBOSA, W. L. **Análise de estudos sobre aplicações e desafios da implementação de *big data* na administração pública**. [*S. l.*], 2017. Disponível em: http://dspace.uniube.br:8080/jspui/handle/123456789/439. Acesso em: 24 dez. 2022.

BAUMAN, Zygmunt. **Modernidade líquida**. Rio de Janeiro: Zahar, 2001.

BAXENDALE, P. B. Machine-Made Index for Technical Literature—An Experiment. **IBM Journal of Research and Development**, v. 2, n. 4, p. 354-361, out. 1958, Disponível em: https://ieeexplore.ieee.org/document/5392648/authors#authors. Acesso em: 20 nov. 2023.

BENIN, Keli Rodrigues do Amaral. **Processamento de linguagem natural e a Ciência da Informação:** inter-relações e contribuições. 2023. Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual de Londrina, Londrina, 2023. Disponível em: https://repositorio.utfpr.edu.br/jspui/bitstream/1/32098/3/processamentolinguagemnaturalcienciainformacao.pdf. Acesso em: 20 nov. 2023.

BERRY, M. J. A.; LINOFF, G. **Data mining techniques:** for marketing, sales, and customer relationship management. 2. ed. Indianapolis, Ind: Wiley Pub, 2004.

BOENTE, A. N. P.; GOLDSCHMIDT, R. R.; ESTRELA, V. V. **Uma metodologia de suporte ao processo de descoberta de conhecimento em bases de dados**. [*S. l.*], 2008. Disponível em: http://boente.eti.br/publica/seget2008kdd.pdf. Acesso em: 20 nov. 2023.

CABRAL, L. de S. **Uma plataforma para sumarização automática de textos independente de idioma**. 27 fev. 2015. [doctoralThesis]. Disponível em: https://repositorio.ufpe.br/handle/123456789/14968. Acesso em: 5 jan. 2023.

CAMARGO, Renata Tironi de. **Investigação de estratégias de sumarização humana multidocumento**. 2013. 135 f. Dissertação (Mestrado em Ciências Humanas) – Universidade Federal de São Carlos, São Carlos, 2013. Disponível em: https://bdtd.ibict.br/vufind/Record/SCAR_6637bc5db024ee254f3183eedccc6a2a. Acesso em: 16 nov. 2023.

CAMARGO, Y., V.; DI-FELIPPO, A. Enriquecendo o *corpus* CM2News: construção e anotação de coleções bilíngues de notícias. **Proceedings of the 6th Workshop on Portuguese Description** (JDP), as a collocated event of the 12th Brazilian Symposium in Information and Human Language Technology (STIL), Salvador/BA, 15-18, p. 239-243, out. 2019. Disponível em: http://www.nilc.icmc.usp.br/nilc/download/ariani/CamargoDiFelippo_JDP_2019. Acesso em: 20 nov. 2023.

CÂNDIDO, T. G. de; WEBBER, C. G. Avaliação da Coesão Textual: Desafios para Automatizar a Correção de Redações. **Renote**, [*S. l.*], v. 16, n. 1, 21 jul. 2018. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/86013. Acesso em: 9 mar. 2023.

CARVALHO, H. L. M. de *et al.* Legibilidade das notas explicativas: uma análise dos maiores municípios mineiros. **RAGC**, [*S. l.*], v. 11, n. 45, 28 mar. 2023. Disponível em: https://www.revistas.fucamp.edu.br/index.php/ragc/article/view/2901. Acesso em: 28 abr. 2023.

CHOMSKY, N. **Syntactic structures**. 2. ed. Berlin ; New York: Mouton de Gruyter, 2002. Disponível em: https://tallinzen.net/media/readings/chomsky_syntactic_structures.pdf. Acesso em: 18 abr. 2023.

COLLOVINI, S. *et al.* Summit: um corpus anotado com informações discursivas visando à sumarização automática. **5º Workshop em Tecnologia da Informação e da Linguagem Humana** (TIL'2007). Rio de Janeiro: Proceedings of the SBC, 2007. Disponível em: http://www.nilc.icmc.usp.br/til/til2007/arq0165.pdf. Acesso em: 20 nov. 2023.

CONDORI, R. E. L. **Sumarização automática de opiniões baseada em aspectos**. [*S. l.*], 2015. Disponível em: https://sites.icmc.usp.br/taspardo/Dissertation2015-LopezCondori.pdf. Acesso em: 20 nov. 2023.

CONTI, F. D. **Mineração de dados no *moodle*:** análise de prazos de entrega de atividades. [*S. l.*], 2011. Disponível em: http://repositorio.ufsm.br/handle/1/5389.

CÓRDULA, E. B. de L.; NASCIMENTO, G. C. C. do. A produção do conhecimento na construção do saber sociocultural e científico. **Educação Pública**, 2018. Disponível em: https://educacaopublica.cecierj.edu.br/artigos/18/12/a-produo-do-conhecimento-na-construo-do-saber-sociocultural-e-cientfico. Acesso em: 21 dez. 2022.

COSTA VAL, M. G. **Texto e textualidade.** Redação e textualidade. São Paulo: Martins Fontes, 1991. Disponível em: https://docplayer.com.br/21410495-Texto-e-textualidade-costa-val-m-g-redacao-e-textualidade-s-paulo-martins-fontes-1991-1-o-que-e-texto.html. Acesso em: 8 mar. 2023.

DALLAGNOL pagará R$ 75 mil a Lula por entrevista do Powerpoint. 22 mar. 2022. Disponível em: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/22032022-STJ-condena-ex-procurador-Dallagnol-a-indenizar-Lula-em-R--75-mil-por-entrevista-do-PowerPoint.aspx. Acesso em: 22 mar. 2023.

DALLARI, Dalmo de Abreu. **O poder dos juízes**. 3. ed., rev. São Paulo: Saraiva, 2007.

DE LUCA, R. C. **Aplicação de conhecimento léxico-conceitual na sumarização automática multidocumento**. [*S. l.*], 2019. Disponível em: https://repositorio.ufscar.br/bitstream/handle/ufscar/11163/Rejeane%20C.%20de%20Luca%20-%20Disserta%C3%A7%C3%A3o%20Final.pdf?sequence=1&isAllowed=y. Acesso em: 22 mar. 2023.

DE MAURO, A.; GRECO, M.; GRIMALDI, M. **What is Big Data?** A Consensual Definition and a Review of Key Research Topics. [*S. l.*: *s. n.*], 2014.

DEVLIN, J. *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In*: NAACL-HLT 2019, jun. 2019. **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:** Human Language Technologies, v. 1 (Long and Short Papers) [...]. Minneapolis, Minnesota: Association for Computational Linguistics, jun. 2019. p. 4171–4186. Disponível em: https://aclanthology.org/N19-1423. Acesso em: 18 abr. 2023.

EAGLETON, T. **A idéia de cultura**. Tradução Sandra Castello Branco. São Paulo: UNESP, 2005.

EDMUNDSON, H. P. New Methods in Automatic Extracting. **Journal of ACM** 16(2): 264-285, 1969. Disponível em: https://dl.acm.org/doi/10.1145/321510.321519#sec-cit. Acesso em: 20 nov. 2023.

EIFLER SARAIVA, C. A.; LIMA ARGIMON, I. I. de. Ciência da computação e ciência cognitiva: um paralelo de semelhanças. **Ciências & Cognição**, [*S. l.*], v. 12, p. 150-155, nov. 2007. Disponível em: http://pepsic.bvsalud.org/scielo.php?script=sci_abstract&pid=S1806-58212007000300014&lng=pt&nrm=iso&tlng=pt. Acesso em: 16 fev. 2023.

ENGELAGE, T. P. **O discurso jurídico e as estratégias argumentativas nas práticas**. 2016. Disponível em: http://www.eumed.net/rev/cccss/2016/03/conflito.html. Acesso em: 16 fev. 2023.

ESPINA, Alice Picon; RINO, Lucia Helena Machado. Utilização de métodos extrativos na sumarização automática de textos. **Nilc-TR-02-06**, mar, 2002. Disponível em: http://www.nilc.icmc.usp.br/nilc/download/NILCTR0206-EspinaRino.pdf. Acesso em: 20 nov. 2023.

ESTIVAL, Dominique; SPARCK JONES, Karen; GALLIERS, Julia R. Evaluating Natural Language Processing Systems: An Analysis and Review. **Lecture Notes in Artificial Intelligence** 1083. Machine Translation. 12. 375-379, 1997. Disponível em: https://www.researchgate.net/publication/220418956_Karen_Sparck_Jones_Julia_R_Galliers_Evaluating_Natural_Language_Processing_Systems_An_Analysis_and_Review_Lecture_Notes_in_Artificial_Intelligence_1083. Acesso em: 21 nov. 2023.

FALAVIGNA, A. *et al.* Apreensibilidade e qualidade da informação: bases de uma avaliação textual automática na área da saúde. *In*: ANAIS DO XVI WORKSHOP DE INFORMÁTICA MÉDICA, 13 fev. 2020. **Anais do Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)** [...]. [*S. l.*]: SBC, 13 fev. 2020. p. 2617–2620. Disponível em: https://sol.sbc.org.br/index.php/sbcas/article/view/9911. Acesso em: 9 mar. 2023.

FÁVERO, L. L. **A informatividade como elemento de textualidade.** [*S. l.*]: Letras de hoje, v. 20, n2, 1985. Disponível em: https://revistaseletronicas.pucrs.br/index.php/fale/article/download/17487/11220.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory.; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, *[S. l.]*, v. 17, n. 3, p. 37, 1996. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 16 nov. 2023.

FEIJÓ, D. de V. **Summarizing legal rulings**. Accepted: 2021-10-09T05:04:49Z, 2021. Disponível em: https://lume.ufrgs.br/handle/10183/230669. Acesso em: 18 abr. 2023.

FEIJÓ, Diego; MOREIRA, Viviane. RulingBR: A Summarization Dataset for Legal Texts: 13th International Conference, PROPOR 2018, Canela, Brazil, set. 24-26, 2018, Proceedings. 10.1007/978-3-319-99722-3_26. Disponível em: https://www.researchgate.net/publication/327224046_RulingBR_A_Summarization_Dataset_for_Legal_Texts_13th_International_Conference_PROPOR_2018_Canela_Brazil_September_24-26_2018_Proceedings/citation/download. Acesso em: 20 nov. 2023.

FERNANDES, J. G. dos S. *et al.* Malha de saberes: Memória, narrativa e história oral na produção e na transmissão do conhecimento. **Projeto História : Revista do Programa de Estudos Pós-Graduados de História**, [*S. l.*], v. 72, p. 284-308, 15 dez. 2021. Disponível em: https://revistas.pucsp.br/index.php/revph/article/view/54797. Acesso em: 10 ago. 2023.

FERNANDES, S. C.; MEIRINHOS, M. Superabundância de informação: um dilema na sociedade digital. **VII Conferência Ibérica de Inovação na Educação com TIC: ieTIC2021**, p. 45-55, 2021. Disponível em: https://bibliotecadigital.ipb.pt/handle/10198/24629. Acesso em: 19 abr. 2023.

FONSECA, E. R. **Reconhecimento de implicação textual em português**. 2018. Doutorado em Ciências de Computação e Matemática Computacional – Universidade de São Paulo, São Carlos, 2018. Disponível em: http://www.teses.usp.br/teses/disponiveis/55/55134/tde-18102018-144020/. Acesso em: 19 abr. 2023.

FRANÇA, R. S. de; AMARAL, H. J. C. do. Mineração de Dados na Identificação de Grupos de Estudantes com Dificuldades de Aprendizagemno Ensino de Programação. **RENOTE**, [*S. l.*], v. 11, n. 1, 5 ago. 2013. Disponível em: https://seer.ufrgs.br/renote/article/view/41634. Acesso em: 26 dez. 2022.

FREITAS JUNIOR, J. C. da S. *et al.* ***Big data*** **e gestão do conhecimento:** definições e direcionamentos de pesquisa, 2016. Disponível em: https://lume.ufrgs.br/handle/10183/163533. Acesso em: 10 ago. 2023.

GALVÃO, N. D.; MARIN, H. de F. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, [*S. l.*], v. 22, n. 5, p. 686-690, out. 2009. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-21002009000500014&lng=pt&tlng=pt. Acesso em: 26 dez. 2022.

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. **Artificial Intelligence Review**, [*S. l.*], v. 47, 1º jan. 2017.

GOFF, Jacquer L. **História e memória**. [*S. l.*], 1990.

GOLDIM, J. R. Consentimento e informação: a importância da qualidade do texto utilizado. **Clinical and Biomedical Research**, [*S. l.*], v. 26, n. 3, 2006. Disponível em: https://www.seer.ufrgs.br/index.php/hcpa/article/view/99986. Acesso em: 9 mar. 2023.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining**: um guia prático. Rio de Janeiro: Elsevier, 2005.

GOMES, T. A. **Avaliação de técnicas de similaridade textual na uniformização de jurisprudência**. [*S. l.*], 2020. Disponível em: https://repositorio.unb.br/bitstream/10482/40798/1/2020_ThiagoAlencarGomes.pdf.

HARTOG, François. Tempo e patrimônio. **Varia História**, 1º dez. 2006. Disponível em: https://www.researchgate.net/publication/250992200_Tempo_e_patrimonio/citation/download. Acesso em: 20 nov. 2023.

HASAN, Tahmid *et al.* XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages, 4693-4703. 10.18653/v1/2021.findings-acl.413, 2021. Disponível em: https://www.researchgate.net/publication/353485951_XL-Sum_Large-

Scale_Multilingual_Abstractive_Summarization_for_44_Languages/citation/download. Acesso em: 20 nov. 2023.

HOCKEY, S. The History of Humanities Computing. *In*: SCHREIBMAN, S.; SIEMENS, R.; UNSWORTH, J. (org.). **A Companion to Digital Humanities**. Malden, MA, USA: Blackwell Publishing Ltd, 2004. p. 1-19. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/9780470999875.ch1. Acesso em: 23 dez. 2022.

HOVY, E.; LIN, C. Automated Text Summarization in SUMMARIST. 1997. **Intelligent Scalable Text Summarization** [...]. [*S. l.*: *s. n.*], 1997. Disponível em: https://aclanthology.org/W97-0704. Acesso em: 19 abr. 2023.

HUTCHINS J. Summarization: Some Problems and Methods. *In:* JONES, Karen Spark (ed.). **Meaning:** the frontier of informatics. London: Aslib,1987. p. 151-173. Disponívem em: http://www.mariapinto.es/ciberabstracts/Articulos/55.pdf. Acesso em: 20 nov. 2023.

INÁCIO, M. L. **Sumarização de opinião com base em abstract meaning representation**. 2021. Mestrado (Ciências de Computação e Matemática Computacional) – Universidade de São Paulo, São Carlos, 2021. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-13092021-141741/. Acesso em: 5 jan. 2023.

INATOMI, C. C. A abordagem da mobilização do direito entre a crítica necessária e a crítica possível. **Lua Nova: Revista de Cultura e Política**, [*S. l.*], n. 108, p. 101-119, dez. 2019. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-64452019000300101&tlng=pt. Acesso em: 23 dez. 2022.

JORGE, Maria Lucía Castro; PARDO, Thiago. Experiments with CST-Based Multidocument Summarization. **Proceedings of TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing**, p. 74-82, Uppsala, Sweden. Association for Computational Linguistic, 2010. Disponível em: https://aclanthology.org/W10-2312/. Acesso em: 20 nov. 2023.

JURAFSKY, Daniel; MARTIN, James. **Speech and Language Processing:** An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2008. Disponível em: https://www.researchgate.net/publication/200111340_Speech_and_Language_Processing_An_Introduction_to_Natural_Language_Processing_Computational_Linguistics_and_Speech_Recognition/citation/download. Acesso em: 16 nov. 2023.

KUPIEC, Julian; PEDERSEN, Jan; CHEN, Francine. A trainable document summarizer. **Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (**SIGIR '95). Association for Computing Machinery, New York, NY, USA, 68-73, 1995. Disponível em: https://dl.acm.org/doi/10.1145/215206.215333. Acesso em: 16 nov. 2023.

LADHAK, Faisal *et al.* WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization. **Findings of the Association for Computational Linguistics:**

EMNLP, 2020. p. 4034-4048, Online. Association for Computational Linguistics. Disponível em: https://aclanthology.org/2020.findings-emnlp.360.pdf. Acesso em: 16 nov. 2023.

LAPATA, M.; LIU, Y. (2019). Text Summarization with Pretrained Encoders. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. Disponível em: https://aclanthology.org/D19-1387.pdf. Acesso em: 16 nov. 2023.

LAROCCA NETO, Joel *et al.* **Contribuição ao estudo de técnicas para sumarização automática de textos**. Dissertação (Mestrado) – Pontifícia Universidade Católica do Paraná, Brasil, 2002.

LAROCCA NETO, Joel *et al.* **Generating Text Summaries through the Relative Importance of Topics**, 2000. Disponível em: https://www.researchgate.net/publication/226125333_Generating_Text_Summaries_through_the_Relative_Importance_of_Topics/link/00b4953211ae3dbb19000000/download. Acesso em: 16 nov. 2023.

LEITÃO, Márcio Martins. Processamento anafórico. *In:* MAIA, Marcus. **Psicolinguística, psicolinguística:** uma introdução. São Paulo: Contexto, 2015.

LEITE, D. S. **Um estudo comparativo de modelos baseados em estatísticas textuais, grafos e aprendizado de máquina para sumarização automática de textos em português**. 21 dez. 2010. Disponível em: https://repositorio.ufscar.br/handle/ufscar/459. Acesso em: 2 jan. 2023.

LEITE, Daniel; RINO, Lucia. **Selecting a Feature Set to Summarize Texts in Brazilian Portuguese**, 2006. Disponível em: https://www.researchgate.net/publication/220943111_Selecting_a_Feature_Set_to_Summarize_Texts_in_Brazilian_Portuguese. Acesso em: 16 nov. 2023.

LLORET, Elena; SANZ, Manuel. Text summarisation in progress: **A literature review. Artif. Intell. Rev.**, 37, 2012. Disponível em: https://www.researchgate.net/publication/220637739_Text_summarisation_in_progress_A_literature_review/link/63109e0261e4553b9557e906/download. Acesso em: 16 nov. 2023.

LÓPEZ CONDORI, R. E.; SALGUEIRO PARDO, T. A. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. **Expert Systems with Applications**, [*S. l.*], v. 78, p. 124–134, 15 jul. 2017. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417417300829. Acesso em: 5 jan. 2023.

LUHN, H. P. A Business Intelligence System. **IBM Journal of Research and Development**, v.. 2, n. 4, p. 314-319, out. 1958. Disponível em: https://ieeexplore.ieee.org/document/5392644. Acesso em: 16 nov. 2023.

MACHADO, Aydano P. *et al.* Mineração de texto em redes sociais aplicada à educação a distância. **Colabor@ - Revista Digital da CVA-RICESU**, v. 6, n. 23, jul. 2010. Disponível em: https://silo.tips/download/mineraao-de-texto-em-redes-sociais-aplicada-a-educaao-a-distancia. Acesso em: 16 nov. 2023.

MANI, I.; MAYBURY, M.T. **Advances in automatic text summarization**. The MIT Press, Cambridge, MA., 1999. Disponível em: https://www.academia.edu/12815768/Advances_in_automatic_text_summarization. Acesso em: 16 nov. 2023.

MANN, W. C.; THOMPSON, S. A. Rhetorical Structure Theory: Toward a functional theory of text organization. **Text – Interdisciplinary Journal for the Study of Discourse**, [*S. l.*], v. 8, n. 3, 1988. Disponível em: https://www.sfu.ca/rst/05bibliographies/bibs/Mann_Thompson_1988.pdf. Acesso em: 7 jan. 2023.

MARANHÃO, J. S. de A.; FLORÊNCIO, J. A.; ALMADA, M. Inteligência artificial aplicada ao direito e o direito da inteligência artificial. **Suprema – Revista de Estudos Constitucionais**, [*S. l.*], v. 1, n. 1, p. 154-180, 30 jun. 2021. Disponível em: https://suprema.stf.jus.br/index.php/suprema/article/view/20. Acesso em: 16 fev. 2023.

MARCONI, D. Visual law e legal design provocam revolução no Poder Judiciário. 3 jan. 2022. **Consultor Jurídico**. Disponível em: https://www.conjur.com.br/2022-jan-03/darci-visual-law-legal-design-provocam-revolucao-judiciario. Acesso em: 24 fev. 2023.

MARCU, D. From discourse structures to text summaries. 1997. **Intelligent Scalable Text Summarization** [...]. [*S. l.*: *s. n.*], 1997. Disponível em: https://aclanthology.org/W97-0713. Acesso em: 7 jan. 2023.

MARTINS, C.B. *et al.* Introdução à sumarização automática. **Relatório Técnico RT-DC 002/2001**, Departamento de Computação, Universidade Federal de São Carlos, 2001. Disponível em: https://sites.icmc.usp.br/taspardo/rtdc00201-cmartinsetal.pdf. Acesso em: 16 nov. 2023.

MARTINS, Camilla; RINO, Lucia. **Pruning UNL texts for Summarizing Purposes**, 2001. p. 539-544. Disponível em: https://www.researchgate.net/publication/220706830_Pruning_UNL_texts_for_Summarizing_Purposes. Acesso em: 16 nov. 2023.

MARTINS, T. B. F. *et al*. **Readability formulas applied to textbooks in brazilian portuguese**. São Carlos: ICMSC-USP. 1996. Disponível em: https://repositorio.usp.br/item/000906089. Acesso em: 16 nov. 2023.

MATTHEWS, P. H. **Linguistics:** a very short introduction. Oxford: Oxford University Press, 2003.

MAZIERO, E.G. *et al*. TeMário 2006: Estendendo o Córpus TeMário. **NILC-TR-07-06.** NILC, ago. 2007 Disponível em: http://www.nilc.icmc.usp.br/nilc/download/NILCTR0706-MazieroEtAl(2).pdf. Acesso em: 16 nov. 2023.

MAZIERO, Erick G.; PARDO, Thiago A. S.; ALUÍSIO, Sandra M. Ferramenta Automática de Simplificação. Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos Departamento de Ciências da Computação (SCC), Núcleo Interinstitucional de Lingüística Computacional, 2009. Disponível em: https://sites.icmc.usp.br/taspardo/WICT2009-MazieroEtAl.pdf. Acesso em: 20 nov. 2023.

McKEOWN, K.R. **Text Generation:** Using Discourse Strategies and Focus Constraints to Generate Natural Language Text. Cambridge University Press, Cambridge, 1985.

MICROSOFT Office Word – AutoResumo, 2003. Disponível em: https://support.microsoft.com/pt-br/word. Acesso em: 15 abr. 2022.

MIHALCEA, R. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. jul. 2004. **Proceedings of the ACL Interactive Poster and Demonstration Sessions** [...]. Barcelona, Spain: Association for Computational Linguistics, jul. 2004. p. 170-173. Disponível em: https://aclanthology.org/P04-3020. Acesso em: 7 jan. 2023.

MIRANDA, C. J. R. **Processamento em Streaming:** Avaliação de Frameworks em contexto Big Data. [*S. l.*], 2018. Disponível em: https://hdl.handle.net/1822/59130.

MOENS, Marc; TEUFEL, Simone. **Argumentative Classification of Extracted Sentences**, 2001. Disponível em: https://www.researchgate.net/publication/2571959_Argumentative_Classification_of_Extracted_Sentences. Acesso em: 16 nov. 2023.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de textos**. Instituto de Informática Universidade Federal de Goiás, 2007. Disponível em: https://docplayer.com.br/4489889-Mineracao-de-textos-e-a-m-morais-a-p-l-ambrosio-instituto-de-informatica-universidade-federal-de-goias-www-inf-ufg-br.html. Acesso em: 15 mar. 2023.

MULLER, E.; GRANATYR, J.; LESSING, O. R. Comparativo entre o algoritmo de Luhn e o algoritmo GistSumm para sumarização de documentos. **Revista de Informática Teórica e Aplicada**, [*S. l.*], v. 22, n. 1, p. 75, 4 maio 2015. Disponível em: http://seer.ufrgs.br/index.php/rita/article/view/RITA-VOL22-NR1-75. Acesso em: 27 abr. 2023.

NACIMENTO, Marcelo; GUELPELI, Marcus. (2011). **BLMSumm** – Métodos de Busca Local e Metaheurísticas na Sumarização de Textos, 2011. Disponível em: https://www.researchgate.net/publication/236686149_BLMSumm_-_Metodos_de_Busca_Local_e_Metaheuristicas_na_Sumarizacao_de_Textos. Acesso em: 21 nov. 2023.

NASCIMENTO NETO, Alisson; GOMES, Andrea; NETO, Manoel. SatSumm – **Uma ferramenta para sumarização automática de textos jornalísticos**, 2007. Disponível em:

https://www.researchgate.net/publication/266031342_SATSUMM_-UMA_FERRAMENTA_PARA_SUMARIZACAO_AUTOMATICA_DE_TEXTOS_JORNALISTICOS. Acesso em: 16 nov. 2023.

NASCIMENTO, Darlan Xavier. **Explorando a avaliação de sumários automáticos multidocumento multilíngues**. Dissertação (Mestrado em Linguística) – Universidade Federal de São Carlos, São Carlos, 2020. Disponível em: https://bdtd.ibict.br/vufind/Record/SCAR_faaff2e5145492287b5fcc0842e8794c. Acesso em: 16 nov. 2023.

NASCIMENTO, Rafaella do *et al.* Mineração de dados na identificação de empresas irregulares quanto ao pagamento de impostos. **Revista de Engenharia e Pesquisa Aplicada**, 2018. Disponível em: https://www.researchgate.net/publication/329782903_Mineracao_de_Dados_na_Identificacao_de_Empresas_Irregulares_Quanto_ao_Pagamento_de_Impostos. Acesso em: 16 nov. 2023.

NENKOVA, Ani; PASSONNEAU, Rebecca. **Evaluating Content Selection in Summarization:** The Pyramid Method, 2004. p. 145-152. Disponível em: https://www.researchgate.net/publication/220817661_Evaluating_Content_Selection_in_Summarization_The_Pyramid_Method. Acesso em: 16 nov. 2023.

NUNES, Maria *et al.* **GistSumm:** A Summarization Tool Based on a New, 2003. Disponível em: https://www.researchgate.net/publication/2869190_GistSumm_A_Summarization_Tool_Based_on_a_New. Acesso em: 16 nov. 2023.

OLIVEIRA, Bruno Vilela. **Uma análise de estratégias de sumarização automática**. 9f. Dissertação (Mestrado em Engenharia Civil) – Universidade Federal do Rio de Janeiro, 2008. Disponível em: https://www.livrosgratis.com.br/ler-livro-online-73771/uma-analise-de-estrategias-de-sumarizacao-automatica. Acesso em: 21 jun. 2023.

OLIVEIRA, N. *et al*. **Processamento de linguagem natural para identificação de notícias falsas em redes sociais:** ferramentas, tendências e desafios. [S.l.: s.n.], 2020. Disponível em: http://sbseg.sbc.org.br/2020/capitulos/capitulo%202.pdf. Acesso em: 16 nov. 2023.

OLIVEIRA, R. P. de; ARAUJO, G. C. de. Qualidade do ensino: uma nova dimensão da luta pelo direito à educação. **Revista Brasileira de Educação**, Rio de Janeiro, n. 28, p. 5-23, abr. 2005. Disponível em: Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-24782005000100002&lng=en&nrm=iso. Acesso em: 21 fev. 2021.

OLIVEIRA, R. P. de; ARAUJO, G. C. de. Qualidade do ensino: uma nova dimensão da luta pelo direito à educação. **Revista Brasileira de Educação**, [*S. l.*], n. 28, p. 5–23, abr. 2005. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-24782005000100002&lng=pt&tlng=pt. Acesso em: 9 mar. 2023.

OST, F. François Ost. **O tempo do direito**, 2005. Tradução Élcio Fernandes; revisão técnica Carlos Aurélio Mota de Souza. Disponível em: https://www.lexml.gov.br/urn/urn:lex:br:rede.virtual.bibliotecas:livro:2005;000729587. Acesso em: 16 fev. 2023.

PAIOLA, P. H. **Sumarização abstrativa de textos em português utilizando aprendizado de máquina**. [*S. l.*], 2022.

PARDO, T. A. S. **Sumarização automática:** principais conceitos e sistemas para o português brasileiro. [S. l.], 2008. Disponível em: https://sites.icmc.usp.br/taspardo/NILCTR0804-Pardo.pdf. Acesso em: 27 dez. 2022.

PARDO, T.A.S. GistSumm – GIST SUMMarizer: extensões e novas funcionalidades. **Série de Relatórios do NILC. NILC-TR-05-05**. São Carlos-SP, fevereiro, 8p. 2005. Disponível em: https://sites.icmc.usp.br/taspardo/NILCTR0505-Pardo.pdf. Acesso em: 21 nov. 2023.

PARDO, Thiago Alexandre Salgueiro; RINO, Lucia Helena Machado; NUNES, Maria das Graças Volpe. NeuralSumm: uma abordagem conexionista para a sumarização automática de textos. **Anais**... Campinas: SBC, 2003. Disponível em: https://repositorio.usp.br/item/001339266. Acesso em: 16 nov. 2023.

PARDO, Thiago; RINO, Lucia. **DMSumm: Review and Assessment**. 263-274, 2002. Disponível em: https://www.researchgate.net/publication/221522998_DMSumm_Review_and_Assessment. Acesso em: 21 nov. 2023.

PASSOS CRUZ, Carla Cristina; VINÍCIUS, Janick; LEITE DOS SANTOS, Jéssica. A utilização de mineração textual e redes semânticas para identificação de tendências em segurança e defesa. 2019. Disponível em: https://www.researchgate.net/publication/353193013_A_UTILIZACAO_DE_MINERACAO_TEXTUAL_E_REDES_SEMANTICAS_PARA_IDENTIFICACAO_DE_TENDENCIAS_EM_SEGURANCA_E_DEFESA/citation/download. Acesso em: 19 nov. 2023.

PEREIRA, Silvio do Lago. **Processamento de linguagem natural**. [S. I.], v. 31. São Paulo: Universidade de São Paulo, 2011. Disponível em: https://www.ime.usp.br/~slago/IA-pln.pdf. Acesso em: 21 nov. 2023.

POLLOCK, J.J.; ZAMORA, A. Automatic Abstracting Research at Chemical Abstracts Service. **Journal of Chemical Information and Compute Sciences** 15(4): 226-232, 1995. Disponível em: https://pubs.acs.org/doi/10.1021/ci60004a008. Acesso em: 21 nov. 2023.

PORCARO, R. M.; LIFSCHITZ, S.; MCC, P.-R. **Mineração de dados** – Funcionalidades, técnicas e abordagens. [*S. l.*], 2002. Disponível em: https://docplayer.com.br/694226-Mineracao-de-dados-funcionalidades-tecnicas-e-abordagens.html. Acesso em: 27 dez. 2022.

PYTHON SOFTWARE FOUNDATION, 2020. Disponível em: https://docs.python.org/3/ https://www.python.org/. Acesso em: 21 nov. 2023.

QUEIROZ, M. P.; FIALHO, F.; REMOR, C. A. Hierarquia DIKW e Capital Humano. **SUCEG – Seminário de Universidade Corporativa e Escolas de Governo**, [*S. l.*], v. 1, n. 1, p. 149-166, 4 dez. 2017. Disponível em: https://anais.suceg.ufsc.br/index.php/suceg/article/view/20. Acesso em: 28 abr. 2023.

REIS, L. F. S. **O direito surgiu antes da escrita *law came before of writing***. [*S. l.*], 2019.

RESUMIDOR DE TEXTO. [*s. d.*]. Disponível em: https://www.summarizer.org/br/resumidor-de-texto. Acesso em: 25 mar. 2023.

RIBEIRO, C. J. S. *Big data:* os novos desafios para o profissional da informação. **Big Data**, [*S. l.*], 2014.

RINO, L.H.M. **Modelagem de discurso para o tratamento da concisão e preservação da idéia central na geração de textos**. Tese (Doutorado) – IFSC-USP. São Carlos-SP, 1996. Disponível em: https://www.teses.usp.br/teses/disponiveis/76/76132/tde-07012009-095918/pt-br.php. Acesso em: 21 nov. 2023.

RINO, Lucia H Machado. **Sobre geração e sumarização de textos**. [*S. l.*], 2005.

RINO, Lucia Helena Machado *et al.* A Comparison of Automatic Summarizers of Texts in Brazilian Portuguese. *In*: BAZZAN, A. L. C.; LABIDI, S. (org.). **Advances in Artificial Intelligence – SBIA 2004**. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. v. 3171, p. 235-244. Disponível em: http://link.springer.com/10.1007/978-3-540-28645-5_24. Acesso em: 2 jan. 2023.

RINO, Lucia Helena Machado; PARDO, T. A. S. A **Sumarização automática de textos:** Principais características e metodologias. [*S. l.*], 2003. Disponível em: https://sites.icmc.usp.br/taspardo/jaia2003-rinopardo.pdf. Acesso em: 2 jan. 2023.

ROCHA DA SILVA, R.; SALGUEIRO PARDO, T. A. Building Contrastive Summaries of Subjective Text Via Opinion Ranking. **Revista de Informática Teórica e Aplicada**, [*S. l.*], v. 29, n. 2, p. 11–34, 14 maio 2022. Disponível em: https://seer.ufrgs.br/index.php/rita/article/view/118372. Acesso em: 5 jan. 2023.

ROCHA, Valdir Júnior Cordeiro. **PragmaSUM:** novos métodos na utilização de palavras-chave na sumarização automática. 2017. 88 p. Dissertação (Mestrado Profissional) – Programa de Pós-Graduação em Educação, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, 2017. Disponível em: http://acervo.ufvjm.edu.br/jspui/handle/1/1658. Acesso em: 21 nov. 2023.

RODRIGUES, M. Memória, patrimônio, bibliotecas nacionais e a construção da identidade coletiva. **Em Questão**, [*S. l.*], , p. 243–262, 17 set. 2015. Disponível em: https://www.seer.ufrgs.br/index.php/EmQuestao/article/view/54754. Acesso em: 10 ago. 2023.

SABER, Marina *Medina. Efeitos da sobrecarga da informação no cotidiano de jornalistas em Campo Grande - MS*. 2006. 228 f. Dissertação (Mestrado em Ciência da Informação) – Universidade de Brasília, Brasília, 2006. Disponível em: http://repositorio2.unb.br/jspui/handle/10482/5520. Acesso em: 10 ago. 2023.

SAGGION, H.; POIBEAU, T. Automatic Text Summarization: Past, Present and Future. *In*: POIBEAU, T. *et al.* (org.). **Multi-source, Multilingual Information Extraction and Summarization**. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 3–21. Disponível em: http://link.springer.com/10.1007/978-3-642-28569-1_1. Acesso em: 27 abr. 2023.

SALGANIK, M. J. ***Bit by bit:*** social research in the digital age. Princeton: Princeton University Press, 2018.

SANTOS, diogo. **Formalismo e processo –** Uma brevíssima visão. 1 jul. 2013. Editora JC. Disponível em: https://www.editorajc.com.br/formalismo-e-processo-uma-brevissima-visao/. Acesso em: 23 dez. 2022.

SANTOS, Â. F. da S. dos. **Sumarização automática de texto**. masterThesis – 2012. Disponível em: https://ubibliorum.ubi.pt/handle/10400.6/3738. Acesso em: 4 jan. 2023.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. [*S. l.*], v. 1, n. 1, 1996.

SCHWAB, Klaus. **A quarta revolução industrial.** São Paulo: Edipro, 2016. Disponível em: https://ria.ufrn.br/jspui/handle/123456789/1826. Acesso em: 21 nov. 2023.

SILLA JR., Carlos N.; KAESTNER, Celso A. A. **kNNSumm:** um sumarizador automático de documentos utilizando aprendizado baseado em instâncias. Pontifícia Universidade Católica do Paraná (PUC-PR), 2007. Disponível em: Disponível em: https://core.ac.uk/download/15779322.pdf. Acesso em: 16 nov. 2023.

SILVA JUNIOR, A. R. da. Cultura: a palavra e as idéias. **Sociedade e Estado**, [*S. l.*], v. 23, n. 1, p. 171–178, abr. 2008. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-69922008000100008&lng=pt&tlng=pt. Acesso em: 23 dez. 2022.

SILVA, R. A. F. e *et al.* Inteligência artificial e carreiras jurídicas no Brasil: uma revisão e proposta de agenda de pesquisa. **Humanidades & Inovação**, [*S. l.*], v. 8, n. 48, p. 187–203, 6 out. 2021. Disponível em: https://revista.unitins.br/index.php/humanidadeseinovacao/article/view/5568. Acesso em: 23 dez. 2022.

SILVA, L. C. da *et al.* Utilização de técnicas de Mineração de Dados para detectar possíveis relacionamentos entre empresas participantes de licitações nas Forças Armadas. **Acanto em Revista**, [*S. l.*], v. 7, n. 7, p. 85–85, 5 out. 2020. Disponível em: https://portaldeperiodicos.marinha.mil.br/index.php/acantoemrevista/article/view/1146. Acesso em: 26 nov. 2023.

SIMONASSI, R. **Uma abordagem de sumarização automática de textos aplicada a debates online**. 3 dez. 2016. Disponível em: https://bdtd.ucb.br:8443/jspui/handle/123456789/1458. Acesso em: 5 jan. 2023.

SIMSON, O. R. de M. Memória, cultura e poder na sociedade do esquecimento. **Augusto Guzzo Revista Acadêmica**, [*S. l.*], n. 6, p. 14-18, 7 maio 2003. Disponível em: http://fics.edu.br/index.php/augusto_guzzo/article/view/57. Acesso em: 10 ago. 2023.

SMOLKA, A. L. B. A memória em questão: uma perspectiva histórico-cultural. **Educação & Sociedade**, [*S. l.*], v. 21, p. 166–193, jul. 2000. Disponível em: https://www.scielo.br/j/es/a/KVJmjgPbDQt56Jz3XXK9BRF/. Acesso em: 10 ago. 2023.

SOUSA, R. N. de; PRATA, D. N. Resumo automático de textos jurídicos usando grafos com vocabulário controlado e algoritmo k-means com words embedding. **Revista Esmat**, [*S. l.*], v. 11, n. 18, p. 65–80, 14 out. 2019. Disponível em: http://esmat.tjto.jus.br/publicacoes/index.php/revista_esmat/article/view/304. Acesso em: 27 dez. 2022.

SOUZA, C.F.R.; NUNES, M.G.V. Avaliação de algoritmos de sumarização extrativa de textos em português. **Relatórios Técnicos do ICMC-USP (NILC-TR-01-9)**, n. 153, out. 2001. Disponível em: https://icmc.usp.br/institucional/estrutura-administrativa/biblioteca/publicacoes/relatorios-tecnicos. Acesso em: 21 nov. 2023.

SOUZA, Raul Carvalho de. **Uma comparação entre métodos e classificadores em documentos jurídicos de atividades processuais repetitivas na PGDF**. 2021. 110p. Dissertação (Mestrado em Ciência da Computação) - Universidade de Brasília, Brasília, 2021. Disponível em: https://repositorio.unb.br/handle/10482/42507. Acesso em: 21 nov. 2023.

SPARCK JONES, Karen. Discourse Modelling for Automatic Summarisation. **Tech. Report** No. 290. University of Cambridge. UK, February. 1993a. Disponível em: https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-290.pdf. Acesso em: 21 nov. 2023.

SPIRLING, A.; RODRIGUEZ, P. L. Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. **Journal of Politics**, 84(1), 101-115, 2022. Disponível em: https://arthurspirling.org/documents/embed.pdf. Acesso em: 21 nov. 2023.

STF. Especial meio ambiente: ressarcimento por dano ambiental não está sujeito à prescrição. A decisão levou em conta a supremacia do interesse público em relação à conservação de um meio ambiente ecologicamente equilibrado e sadio. **Portal STF**, 16 jun. 2023. Disponível em: https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=509005&ori=1. Acesso em: 20 nov. 2023.

STJ: Cabe ao banco provar autenticidade de assinatura em contrato. **Migalhas**, 24 nov. 2021. Disponível em: https://www.migalhas.com.br/quentes/355495/stj-cabe-ao-banco-provar-autenticidade-de-assinatura-em-contrato. Acesso em: 10 ago. 2023.

SUMY. 2015. Disponível em: https://github.com/miso-belica/sumy. Acesso em: 2 mar. 2023.

TAVARES, Rodrigo de Souza. **ChatGPT para advogados:** guia prático para utilizar a inteligência artificial generativa no direito. 2023. E-book Kindle.

TEXTANALYST (Megaputer). Disponível em: https://www.megaputer.com/what-is-text-analytics/ Acesso em: 10 abr. 2023.

TOSTA, Fabricio Elder da Silva. **Aplicação de conhecimento léxico-conceitual na sumarização multidocumento multilíngue**. 2014. 119 f. Dissertação (Mestrado em Ciências Humanas) – Universidade Federal de São Carlos, São Carlos, 2014. Disponível em: https://repositorio.ufscar.br/handle/ufscar/5796?show=full. Acesso em: 10 abr. 2023.

UZÊDA, V.R.; PARDO, T.A.S.; NUNES, M.G.V. Avaliação de métodos de sumarização automática de textos baseados na *Rhetorical Structure Theory*. **Revista de Iniciação Científica – RIC**. Centro de Tecnologia Educacional para Engenharia – CETEPE, USP/São Carlos. 2007. Disponível em: https://sites.icmc.usp.br/taspardo/NILCTR0707-UzedaEtAl.pdf. Acesso em: 10 abr. 2023.

VALENTIM, M. L. P. **Inteligência competitiva em organizações:** dado, informação e conhecimento. Disponível em: https://www.brapci.inf.br/index.php/article/download/7468. Acesso em: 20 nov. 2023.

VILELA, E. M.; MENDES, I. J. M. Interdisciplinaridade e saúde: estudo bibliográfico. **Revista Latino-Americana de Enfermagem**, [*S. l.*], v. 11, n. 4, p. 525-531, ago. 2003. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-11692003000400016&lng=pt&tlng=pt. Acesso em: 23 dez. 2022.

Outros sites consultados

http://bdtd.ibict.br/vufind/

https://scholar.google.com/

https://scielo.org/

https://www.teses.usp.br/

https://www-periodicos-capes-gov-br.ezl.periodicos.capes.gov.br/index.php?

APÊNDICE A – Experimento em Python de Sumarização com texto jurídico a partir do acórdão de julgamento em REsp 1842613-SP perante o STJ

```
# instalando bibliotecas necessárias
! pip install PyPDF2
# PyPDF2 – trabalhando com arquivos PDF: realizar extração
# re – Operações com expressões regulares
# nltk  – Kit de ferramentas de Processamento de Linguagem Natural
import PyPDF2
import re
import nltk
from PyPDF2 import PdfReader
# montando caminho para acessar o arquivo no onedrive
from google.colab import drive
drive.mount('/content/drive')
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
# carregando a biblioteca de extração de texto em documentos PDF para mineração de dados
!pip install PDFminer.six
# extraindo o texto em pdf e contando número de palavras
from pdfminer.high_level import extract_text
pdf_file = open('/content/drive/MyDrive/REsp n 1842613 STJ.pdf', 'rb')
texto = extract_text(pdf_file)
print (texto)
# extraindo o número de páginas do texto
reader = PdfReader(pdf_file)
len(reader.pages)
print (" O número de páginas é: ", len(reader.pages))
print (texto, " \n" , " O número de palavras é: ", len(texto))
Página 59
Superior Tribunal de Justiça
O número de palavras é:  167607
```

```
# carregando NumPy biblioteca para a linguagem Python com funções numéricas possibilitando maior processamento de dados
!pip install numpy
import numpy
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (1.22.4)
#carregando biblioteca de código aberto chamada Natural Language Toolkit (NLTK) para Processamento de Linguagem Natural
# ferramentas de tratamento de pontuação, remoção de palavras desnecessárias, leitor de corpus
import string
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('stopwords')
string.punctuation
stopwords = nltk.corpus.stopwords.words('portuguese')
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
# preprocessamento do texto
def preprocessamento(texto1):
  texto_formatado = texto1.lower()
  tokens = []
  for token in nltk.word_tokenize(texto_formatado):
    tokens.append(token)
  tokens = [palavra for palavra in tokens if palavra not in stopwords and palavra not in string.punctuation]
  texto_formatado = ' '.join([str(elemento) for elemento in tokens])
  return texto_formatado
texto_formatado = preprocessamento(texto)
texto_formatado
```

‘recurso especial nº 1.842.613 sp 2019/0235636-7 relator recorrente advogados recorrido procurador interes advogados ministro luis felipe salomão luiz inácio lula silva maria lourdes lopes sp077513 valeska teixeira zanin martins sp153720 cristiano zanin martins sp172730 bruno sales biscuola sp302602 deltan martinazzo dallagnol advocacia-geral união agu associacao nacional procuradores republica flavio luiz yarshell sp088098 elizandra mendes camargo ana sp210065 luiza orsolon galardo outro s sp376474 ementa recurso especial ação indenização danos morais entrevista coletiva informar oferecimento denúncia criminal ex-presidente república denunciados divulgação comandada procurador república entrevista destacada narrativa ofensiva técnica utilização powerpoint declaração crimes constavam peça acusatória alegação cerceamento defesa ilegitimidade passiva agente público causador dano matéria ordem impugnada oportunamente pública decidida preclusão assistência conformidade assistido limites ace...’

```
# frequência de palavras
from nltk.tokenize import word_tokenize
frequencia_palavras = nltk.FreqDist(nltk.word_tokenize(texto_formatado))
frequencia_palavras
frequencia_maxima = max(frequencia_palavras.values())
frequencia_maxima
for palavra in frequencia_palavras.keys():
  frequencia_palavras[palavra] = (frequencia_palavras[palavra] / frequencia_maxima)
frequencia_palavras
FreqDist({'``': 1.0, 'direito': 0.6931818181818182, 'público': 0.5909090909090909, 'tribunal': 0.5511363636363636, 'ação': 0.5227272727272727, '--': 0.4715909090909091, 'acórdão': 0.4602272727272727, 'dje': 0.4602272727272727, 'justiça': 0.4318181818181818, 'denúncia': 0.3977272727272727, ...})
# geração de lista
lista = [frequencia_palavras]
lista.sort(reverse=True)
print(lista)
ordenada = sorted(lista, reverse=True)
print(lista)
print(ordenada)
for lista in range(0,5):
  print(lista, ordenada)
```

[FreqDist({'``': 1.0, 'direito': 0.6931818181818182, 'público': 0.5909090909090909, 'tribunal': 0.5511363636363636, 'ação': 0.5227272727272727, '--': 0.4715909090909091, 'acórdão': 0.4602272727272727, 'dje': 0.4602272727272727, 'justiça': 0.4318181818181818, 'denúncia': 0.3977272727272727, ...})]
[FreqDist({'``': 1.0, 'direito': 0.6931818181818182, 'público': 0.5909090909090909, 'tribunal': 0.5511363636363636, 'ação': 0.5227272727272727, '--': 0.4715909090909091, 'acórdão': 0.4602272727272727, 'dje': 0.4602272727272727, 'justiça': 0.4318181818181818, 'denúncia': 0.3977272727272727, ...})]
[FreqDist({'``': 1.0, 'direito': 0.6931818181818182, 'público': 0.5909090909090909, 'tribunal': 0.5511363636363636, 'ação': 0.5227272727272727, '--': 0.4715909090909091, 'acórdão': 0.4602272727272727, 'dje': 0.4602272727272727, 'justiça': 0.4318181818181818, 'denúncia': 0.3977272727272727, ...})]
0 [FreqDist({'``': 1.0, 'direito': 0.6931818181818182, 'público': 0.5909090909090909, 'tribunal': 0.5511363636363636, 'ação': 0.5227272727272727, '--': 0.4715909090909091, 'acórdão': 0.4602272727272727, 'dje': 0.4602272727272727, 'justiça': 0.4318181818181818, 'denúncia': 0.3977272727272727, ...})]
1 [FreqDist({'``': 1.0, 'direito': 0.6931818181818182, 'público': 0.5909090909090909, 'tribunal': 0.5511363636363636, 'ação': 0.5227272727272727, '--': 0.4715909090909091, 'acórdão': 0.4602272727272727, 'dje': 0.4602272727272727, 'justiça': 0.4318181818181818, 'denúncia': 0.3977272727272727, ...})]
2 [FreqDist({'``': 1.0, 'direito': 0.6931818181818182, 'público': 0.5909090909090909, 'tribunal': 0.5511363636363636, 'ação': 0.5227272727272727, '--': 0.4715909090909091, 'acórdão': 0.4602272727272727, 'dje': 0.4602272727272727, 'justiça': 0.4318181818181818, 'denúncia': 0.3977272727272727, ...})]
3 [FreqDist({'``': 1.0, 'direito': 0.6931818181818182, 'público': 0.5909090909090909, 'tribunal': 0.5511363636363636, 'ação': 0.5227272727272727, '--': 0.4715909090909091, 'acórdão': 0.4602272727272727, 'dje': 0.4602272727272727, 'justiça': 0.4318181818181818, 'denúncia': 0.3977272727272727, ...})]
4 [FreqDist({'``': 1.0, 'direito': 0.6931818181818182, 'público': 0.5909090909090909, 'tribunal': 0.5511363636363636, 'ação': 0.5227272727272727, '--': 0.4715909090909091, 'acórdão': 0.4602272727272727, 'dje': 0.4602272727272727, 'justiça': 0.4318181818181818, 'denúncia': 0.3977272727272727, ...})]
# lista de sentenças
lista_sentencas = nltk.sent_tokenize(texto_formatado)

```
print ("\n", lista_sentencas)
```

["recurso especial nº 1.842.613 sp 2019/0235636-7 relator recorrente advogados recorrido procurador interes advogados ministro luis felipe salomão luiz inácio lula silva maria lourdes lopes sp077513 valeska teixeira zanin martins sp153720...

```
# atribuindo notas para sentenças
nota_sentencas = {}
for sentenca in lista_sentencas:
  #print(sentenca)
  for palavra in nltk.word_tokenize(sentenca.lower()):
    #print(palavra)
    if palavra in frequencia_palavras.keys():
      if sentenca not in nota_sentencas.keys():
        nota_sentencas[sentenca] = frequencia_palavras[palavra]
      else:
        nota_sentencas[sentenca] += frequencia_palavras[palavra]

nota_sentencas
```

{"recurso especial nº 1.842.613 sp 2019/0235636-7 relator recorrente advogados recorrido procurador interes advogados ministro luis felipe salomão luiz inácio lula silva maria lourdes lopes sp077513 valeska teixeira zanin martins sp153720 cristiano zanin martins sp172730 bruno sales biscuola sp302602 deltan martinazzo dallagnol advocacia-geral união agu associacao nacional procuradores republica flavio luiz yarshell sp088098 elizandra mendes camargo ana sp210065 luiza orsolon galardo outro s sp376474 ementa recurso especial ação indenização danos morais entrevista coletiva informar oferecimento denúncia criminal ex-presidente república denunciados divulgação comandada procurador república entrevista destacada narrativa ofensiva técnica utilização powerpoint declaração crimes constavam peça acusatória alegação cerceamento defesa ilegitimidade passiva agente público causador dano matéria ordem impugnada oportunamente pública decidida preclusão assistência conformidade assistido limites acessoriedade teoria asserção ilegitimidade alegada contestação determinação instrução probatória decisão meritória stf tema n. 940...

```
# gerando as melhores sentenças
import heapq
melhores_sentencas = heapq.nlargest(1, nota_sentencas, key=nota_sentencas.get)
melhores_sentencas
```

["recurso especial nº 1.842.613 sp 2019/0235636-7 relator recorrente advogados recorrido procurador interes advogados ministro luis felipe salomão luiz inácio lula silva maria lourdes lopes sp077513 valeska teixeira zanin martins sp153720 cristiano zanin martins sp172730 bruno sales biscuola sp302602 deltan martinazzo dallagnol advocacia-geral união agu associacao nacional procuradores republica flavio luiz yarshell sp088098 elizandra mendes camargo ana sp210065 luiza orsolon galardo outro s sp376474 ementa recurso especial ação indenização danos morais entrevista coletiva informar oferecimento denúncia criminal ex-presidente república denunciados divulgação comandada procurador república entrevista destacada narrativa ofensiva técnica utilização powerpoint declaração crimes constavam peça acusatória alegação cerceamento defesa ilegitimidade passiva agente público ...

```
# gerando o sumário
resumo = ' '.join(melhores_sentencas)
resumo
```

recurso especial nº 1.842.613 sp 2019/0235636-7 relator recorrente advogados recorrido procurador interes advogados ministro luis felipe salomão luiz inácio lula silva maria lourdes lopes sp077513 valeska teixeira zanin martins sp153720 cristiano zanin martins sp172730 bruno sales biscuola sp302602 deltan martinazzo dallagnol advocacia-geral união agu associacao nacional procuradores republica flavio luiz yarshell sp088098 elizandra mendes camargo ana sp210065 luiza orsolon galardo outro s sp376474 ementa recurso especial ação indenização danos morais entrevista coletiva informar oferecimento denúncia criminal ex-presidente república denunciados divulgação comandada procurador república entrevista destacada narrativa ofensiva técnica utilização powerpoint declaração crimes constavam peça acusatória alegação cerceamento defesa ilegitimidade passiva agente público causador dano matéria ordem impugnada oportunamente pública decidida preclusão assistência conformidade assistido limites ace ...

```
# contando o número de palavas do primeiro resumo
print(resumo, " \n" , " O número de palavras é: ", len(resumo))
```

recurso especial nº 1.842.613 sp 2019/0235636-7 relator recorrente advogados recorrido procurador interes advogados ministro luis felipe salomão luiz inácio lula silva maria lourdes lopes sp077513 valeska teixeira zanin martins sp153720 cristiano zanin martins sp172730 bruno sales biscuola sp302602 deltan martinazzo dallagnol

O número de palavras é: 33334

```
# carregando a biblioteca sumy para Sumarização de Textos em Português
!pip install sumy
```

```
# aplicando a biblioteca do algoritmo de Luhn para sumarização de texto
from sumy.parsers.plaintext import PlaintextParser
from sumy.nlp.tokenizers import Tokenizer
from sumy.summarizers.luhn import LuhnSummarizer
parser = PlaintextParser.from_string(texto, Tokenizer("portuguese"))
sumarizador = LuhnSummarizer()
resumo1 = sumarizador(parser.document, 3)
# gerando resumo
resumo1
```

(<Sentence: Para a circunstâncias particulares do caso, considera-se a gravidade do fato, ofensa à honra e reputação da vítima, ex-Presidente da República, com base em imputações da prática de crimes que não foram objeto da denúncia e em qualificações não técnicas; os meios utilizados na divulgação, com convocação dos principais canais de TV para transmissão para o Brasil e outros países, com ampla repercussão; a responsabilidade do agente, Procurador da República, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercussão do que se propagava, com razoável capacidade financeira para suportar o pagamento.>, <Sentence: Para a circunstâncias particulares do caso, considera-se a gravidade do fato, ofensa à honra e reputação da vítima, ex-Presidente da República, com base em imputações da prática de crimes que não foram objeto da denúncia e em qualificações não técnicas; os meios utilizados na divulgação, com convocação dos principais canais de TV para transmissão para o Brasil e outros países, com ampla repercussão; a responsabilidade do agente, Procurador da República, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercussão do que se propagava, com razoável capacidade financeira para suportar o pagamento.>, <Sentence: _____ "Além do esquema de corrupção, e denunciado o esquema de lavagem de dinheiro envolvendo o ex-Presidente Lula, o que se constatou foi o repasse de recursos a partir dessa empresa OAS para o ex-Presidente Lula por meio de um upgrade de um apartamento de um imóvel, um triplex no Guarujá, por meio da reforma desse triplex, por meio da decoração desse triplex, e por meio de um contrato de armazenamento de bens pessoais, um contrato milionário firmado para armazenamento, um contrato falso firmado pela OAS como se os bens fossem dela e não do ex-Presidente.>)

```
resumo1 = ' '.join([str(elemento) for elemento in sumarizador(parser.document, 3)])
resumo1
```

Para a circunstâncias particulares do caso, considera-se a gravidade do fato, ofensa à honra e reputação da vítima, ex-Presidente da República, com base em imputações da prática de crimes que não foram objeto da denúncia e em qualificações não técnicas; os meios utilizados na divulgação, com convocação dos principais canais de TV para transmissão para o Brasil e outros países, com ampla repercussão;...

```
# contando o número de palavras do texto
print(texto, "  \n  " , " O número de palavras é de: ", len(texto))
```

```
 Superior Tribunal de Justiça
   O número de palavras é de:  167607
```

```
print(resumo1, " \n" , " O número de palavras é: ", len(resumo1))
```

Para a circunstâncias particulares do caso, considera-se a gravidade do fato, ofensa à honra e reputação da vítima, ex-Presidente da República, com base em imputações da prática de crimes que não foram objeto da denúncia e em qualificações não técnicas; os meios utilizados na divulgação, com convocação dos principais canais de TV para transmissão para o Brasil e outros países, com ampla repercussão; a responsabilidade do agente, Procurador da República, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercussão do que se propagava, com razoável capacidade financeira para suportar o pagamento. Para a circunstâncias particulares do caso, considera-se a gravidade do fato, ofensa à honra e reputação da vítima, ex-Presidente da República, com base em imputações da prática de crimes que não foram objeto da denúncia e em qualificações não técnicas; os meios utilizados na divulgação, com convocação dos principais canais de TV para transmissão para o Brasil e outros países, com ampla repercussão; a responsabilidade do agente, Procurador da República, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercussão do que se propagava, com razoável capacidade financeira para suportar o pagamento. _____ "Além do esquema de corrupção, e denunciado o esquema de lavagem de dinheiro envolvendo o ex-Presidente Lula, o que se constatou foi o repasse de recursos a partir dessa empresa OAS para o ex-Presidente Lula por meio de um upgrade de um apartamento de um imóvel, um triplex no Guarujá, por meio da reforma desse triplex, por meio da decoração desse triplex, e por meio de um contrato de armazenamento de bens pessoais, um contrato milionário firmado para armazenamento, um contrato falso firmado pela OAS como se os bens fossem dela e não do ex-Presidente.

```
 O número de palavras é:  2058
```

```
# pesquisar determinada palavra no corpo do texto e contar o número de ocorrências
palavra = str(input("digite a palavra a ser pesquisada:  "))
if palavra in texto:
    print('\n palavra encontrada: ', palavra)
elif palavra not in texto:
    print("\n palavra não encontrada!")
print("n\ o número de palavras é: ", texto.count(palavra))
digite a palavra a ser pesquisada:  direito
 palavra encontrada:  direito
n\ o número de palavras é:  136
```

APÊNDICE B – Experimento em Python de Sumarização com texto jurídico a partir do acórdão de julgamento do RE 654833-AC perante o STF

```
 # SUMÁRIO GERADO A PARTIR DO ACÓRDÃO NO JULGAMENTO DO REsp nº RE 654833 DO SUPREMO TRIBUNAL FEDERAL – STF
# instalando bibliotecas necessárias
! pip install PyPDF2
 # PyPDF2 –  trabalhar com arquivos PDF e realizar de extração
 # re – Operações com expressões regulares
 # nltk  – Kit de ferramentas de Processamento de Linguagem Natural
import PyPDF2
import re
import nltk
from PyPDF2 import PdfReader
# montando caminho para acessar o arquivo no onedrive
from google.colab import drive
drive.mount('/content/drive')
# carregando a biblioteca de extração de texto em documentos PDF para mineração de dados
!pip install PDFminer.six
# extraindo o texto em pdf e contando número de palavras
from pdfminer.high_level import extract_text
pdf_file = open('/content/drive/MyDrive/RE 654833 imprescritibilidade danos ao meio ambiente.pdf', 'rb')
texto = extract_text(pdf_file)
print (texto)
# extraindo o número de páginas do texto
reader = PdfReader(pdf_file)
len(reader.pages)
print("O número de páginas é: ", len(reader.pages))
print(texto, " \n" , " O número de palavras é: ", len(texto))
Inteiro Teor do Acórdão - Página 126 de 126
[...]

  O número de palavras é:  507242
# carregando NumPy biblioteca para a linguagem Python com funções numéricas possibilitando maior processamento de dados
!pip install numpy
import numpy
#carregando biblioteca biblioteca de código aberto chamada Natural Language Toolkit (NLTK) para Processamento de Linguagem Natural
# ferramentas de tratamento de pontuação, remoção de palavras desnecessárias, leitor de corpus
import string
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('stopwords')
```

```
string.punctuation
stopwords = nltk.corpus.stopwords.words('portuguese')
# preprocessamento do texto
def preprocessamento(texto1):
  texto_formatado = texto1.lower()
  tokens = []
  for token in nltk.word_tokenize(texto_formatado):
    tokens.append(token)
  tokens = [palavra for palavra in tokens if palavra not in stopwords and palavra not in
string.punctuation]
  texto_formatado = ' '.join([str(elemento) for elemento in tokens])
  return texto_formatado
texto_formatado = preprocessamento(texto)
texto_formatado
# frequencia de palavras
from nltk.tokenize import word_tokenize
frequencia_palavras = nltk.FreqDist(nltk.word_tokenize(texto_formatado))
frequencia_palavras
frequencia_maxima = max(frequencia_palavras.values())
frequencia_maxima
for palavra in frequencia_palavras.keys():
  frequencia_palavras[palavra] = (frequencia_palavras[palavra] / frequencia_maxima)
frequencia_palavras
# frequencia de palavras
from nltk.tokenize import word_tokenize
frequencia_palavras = nltk.FreqDist(nltk.word_tokenize(texto_formatado))
frequencia_palavras
frequencia_maxima = max(frequencia_palavras.values())
frequencia_maxima
for palavra in frequencia_palavras.keys():
  frequencia_palavras[palavra] = (frequencia_palavras[palavra] / frequencia_maxima)
frequencia_palavras
# lista de sentenças
lista_sentencas = nltk.sent_tokenize(texto_formatado)
print ("\n", lista_sentencas)
# atribuindo notas para sentenças
nota_sentencas = {}
for sentenca in lista_sentencas:
  #print(sentenca)
  for palavra in nltk.word_tokenize(sentenca.lower()):
    #print(palavra)
    if palavra in frequencia_palavras.keys():
      if sentenca not in nota_sentencas.keys():
        nota_sentencas[sentenca] = frequencia_palavras[palavra]
      else:
        nota_sentencas[sentenca] += frequencia_palavras[palavra]
nota_sentencas
# gerando as melhores sentenças
import heapq
```

```
melhores_sentencas = heapq.nlargest(1, nota_sentencas, key=nota_sentencas.get)
melhores_sentencas
# gerando o sumário
resumo = ' '.join(melhores_sentencas)
resumo
# contando o número de palavas do primeiro resumo
print(resumo, " \n" , " O número de palavras é: ", len(resumo))
# carregando a biblioteca sumy para Sumarização de Textos em Português
!pip install sumy
 # aplicando a biblioteca do algoritmo de Luhn para sumarização de texto
from sumy.parsers.plaintext import PlaintextParser
from sumy.nlp.tokenizers import Tokenizer
from sumy.summarizers.luhn import LuhnSummarizer
parser = PlaintextParser.from_string(texto, Tokenizer("portuguese"))
sumarizador = LuhnSummarizer()
resumo1 = sumarizador(parser.document, 3)
 # gerando resumo
resumo1
resumo1 = ' '.join([str(elemento) for elemento in sumarizador(parser.document, 3)])
resumo1
# contando o número de palavras do texto
print(texto, "  \n  " , " O número de palavras é de: ", len(texto))
print(resumo1, " \n" , " O número de palavras é: ", len(resumo1))
A resposta parece-me positiva.De fato, o artigo 2°, inciso I, da Lei n° 
6.831/81, ao definir a Política Nacional do Meio Ambiente, dispõe como 
um de seus princípios a "ação governamental na manutenção do equilíbrio 
ecológico, considerando o meio ambiente [...]
  O número de palavras é:  5542
# pesquisar determinada palavra no corpo do texto e contar o número de ocorrências
palavra = str(input("digite a palavra a ser pesquisada:  "))
if palavra in texto:
   print('\n palavra encontrada: ', palavra)
elif palavra not in texto:
   print("\n palavra não encontrada!")
print("n\ o número de palavras é: ", texto.count(palavra))
```

Fonte: Autoria própria

APÊNDICE C – Experimento em Python de Sumarização com texto jurídico a partir do acórdão de julgamento do REsp 1.846.649-MA perante o STJ

```
# SUMÁRIO GERADO A PARTIR DO ACÓRDÃO NO JULGAMENTO DO REsp
1.846.649
# instalando bibliotecas necessárias
! pip install PyPDF2
 # PyPDF2 –  trabalhar com arquivos PDF e realizar de extração
 # re – Operações com expressões regulares
 # nltk  – Kit de ferramentas de Processamento de Linguagem Natural
import PyPDF2
import re
import nltk
from PyPDF2 import PdfReader
# montando caminho para acessar o arquivo no onedrive
from google.colab import drive
drive.mount('/content/drive')
# carregando a biblioteca de extração de texto em documentos PDF para mineração de dados
!pip install PDFminer.six
# extraindo o texto em pdf e contando número de palavras
from pdfminer.high_level import extract_text
pdf_file = open('/content/drive/MyDrive/RESP 1.846.649 impugnação autenticidade
assinatura.pdf', 'rb')
texto = extract_text(pdf_file)
print (texto)
# extraindo o número de páginas do texto
reader = PdfReader(pdf_file)
len(reader.pages)
print("O número de páginas é: ", len(reader.pages))
print(texto, " \n" , " O número de palavras é: ", len(texto))
‘O número de páginas é:  23
RECURSO ESPECIAL Nº 1.846.649 – MA  (2019/0329419-2)
RELATOR
RECORRENTE
ADVOGADOS’

[...]

Superior Tribunal de Justiça

 O número de palavras é:  58209

# carregando NumPy biblioteca para a linguagem Python com funções numéricas
possibilitando maior processamento de dados
!pip install numpy
import numpy
#carregando biblioteca biblioteca de código aberto chamada Natural Language Toolkit
(NLTK) para Processamento de Linguagem Natural
```

```
# ferramentas de tratamento de pontuação, remoção de palavras desnecessárias, leitor de
corpus
import string
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('stopwords')
string.punctuation
stopwords = nltk.corpus.stopwords.words('portuguese')
# preprocessamento do texto
def preprocessamento(texto1):
  texto_formatado = texto1.lower()
  tokens = []
  for token in nltk.word_tokenize(texto_formatado):
    tokens.append(token)
  tokens = [palavra for palavra in tokens if palavra not in stopwords and palavra not in
string.punctuation]
  texto_formatado = ' '.join([str(elemento) for elemento in tokens])
  return texto_formatado
texto_formatado = preprocessamento(texto)
texto_formatado
# frequencia de palavras
from nltk.tokenize import word_tokenize
frequencia_palavras = nltk.FreqDist(nltk.word_tokenize(texto_formatado))
frequencia_palavras
frequencia_maxima = max(frequencia_palavras.values())
frequencia_maxima
for palavra in frequencia_palavras.keys():
  frequencia_palavras[palavra] = (frequencia_palavras[palavra] / frequencia_maxima)
frequencia_palavras
# geração de lista
lista = [frequencia_palavras]
lista.sort(reverse=True)
print(lista)
ordenada = sorted(lista, reverse=True)
print(lista)
print(ordenada)
for lista in range(0,5):
  print(lista, ordenada)
# lista de sentenças
lista_sentencas = nltk.sent_tokenize(texto_formatado)
print ("\n", lista_sentencas)
# atribuindo notas para sentenças
nota_sentencas = {}
for sentenca in lista_sentencas:
  #print(sentenca)
  for palavra in nltk.word_tokenize(sentenca.lower()):
    #print(palavra)
    if palavra in frequencia_palavras.keys():
      if sentenca not in nota_sentencas.keys():
```

```
        nota_sentencas[sentenca] = frequencia_palavras[palavra]
      else:
        nota_sentencas[sentenca] += frequencia_palavras[palavra]
nota_sentencas
# gerando as melhores sentenças
import heapq
melhores_sentencas = heapq.nlargest(1, nota_sentencas, key=nota_sentencas.get)
melhores_sentencas
# gerando o sumário
resumo = ' '.join(melhores_sentencas)
resumo
# contando o número de palavas do primeiro resumo
print(resumo, " \n" , " O número de palavras é: ", len(resumo))
# carregando a biblioteca sumy para Sumarização de Textos em Português
!pip install sumy
 # aplicando a biblioteca do algoritmo de Luhn para sumarização de texto
from sumy.parsers.plaintext import PlaintextParser
from sumy.nlp.tokenizers import Tokenizer
from sumy.summarizers.luhn import LuhnSummarizer
parser = PlaintextParser.from_string(texto, Tokenizer("portuguese"))
sumarizador = LuhnSummarizer()
resumo1 = sumarizador(parser.document, 3)
 # gerando resumo
resumo1
resumo1 = ' '.join([str(elemento) for elemento in sumarizador(parser.document, 3)])
resumo1
# contando o número de palavras do texto
print(texto, "  \n  " , " O número de palavras é de: ", len(texto))
 print(resumo1, " \n" , " O número de palavras é: ", len(resumo1))
‘O número de palavras é:  1577’

# pesquisar determinada palavra no corpo do texto e contar o número de ocorrências
palavra = str(input("digite a palavra a ser pesquisada:  "))
if palavra in texto:
    print('\n palavra encontrada: ', palavra)
elif palavra not in texto:
    print("\n palavra não encontrada!")
print("n\ o número de palavras é: ", texto.count(palavra))
```

Fonte: Autoria própria

APÊNDICE D – Avaliação manual de qualidade de sumário automático de texto jurídico

Textos-fonte: (1) Acórdão de Julgamento em Recurso Especial nº 1842613-SP do Superior Tribunal de Justiça –STJ – Sumário A e B; (2) Acórdão de Julgamento em RE 654833-AC perante o Supremo Tribunal Federal – STF – Sumário C e D; (3) Acórdão de Julgamento em Recurso Especial nº 1.846.649-MA perante o STJ – Sumário E e F

Sumários automáticos: **SUMÁRIO A, SUMÁRIO B, SUMÁRIO C, SUMÁRIO D, SUMÁRIO E, SUMÁRIO F**

bmperdigao@gmail.com Alternar conta
Não compartilhado

SUMÁRIO A

SUMÁRIO A

Figura 2 – Extração do Recurso Especial nº 1842613 - SP do Superior Tribunal de Justiça – STJ e sumário gerado funcionalidade melhor linha

Dados da Extração	Dados do Sumário funcionalidade melhor linha

Fonte: Adaptado pelo autor a partir da interface (AISummarizer, 2022)

Sumário gerado com a funcionalidade 'melhor linha':

"Para a circunstncias particulares do caso, considera-se a gravidade do fato, ofensa honra e reputao da vtima, ex-Presidente da Repblica, com base em imputaes da prtica de crimes que no foram objeto da denncia e em qualificaes no tcnicas os meios utilizados na divulgao, com convocao dos principais canais de TV para transmisso para o Brasil e outros pases, com ampla repercusso a responsabilidade do agente, Procurador da Repblica, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercusso do que se propagava, com razovel capacidade financeira para suportar o pagamento."

SUMÁRIO B

SUMÁRIO B

Experimento em Python de Sumarização com texto jurídico a partir do acórdão de julgamento em recurso especial nº 1842613 - SP perante o Superior Tribunal de Justiça - STJ

"Para a circunstâncias particulares do caso, considera-se a gravidade do fato, ofensa à honra e reputação da vítima, ex-Presidente da República, com base em imputações da prática de crimes que não foram objeto da denúncia e em qualificações não técnicas; os meios utilizados na divulgação, com convocação dos principais canais de TV para transmissão para o Brasil e outros países, com ampla repercussão; a responsabilidade do agente, Procurador da República, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercussão do que se propagava, com razoável capacidade financeira para suportar o pagamento. Para a circunstâncias particulares do caso, considera-se a gravidade do fato, ofensa à honra e reputação da vítima, ex-Presidente da República, com base em imputações da prática de crimes que não foram objeto da denúncia e em qualificações não técnicas; os meios utilizados na divulgação, com convocação dos principais canais de TV para transmissão para o Brasil e outros países, com ampla repercussão; a responsabilidade do agente, Procurador da República, capaz tecnicamente de identificar os termos utilizados em seu discurso e a repercussão do que se propagava, com razoável capacidade financeira para suportar o pagamento. _____ "Além do esquema de corrupção, e denunciado o esquema de lavagem de dinheiro envolvendo o ex-Presidente Lula, o que se constatou foi o repasse de recursos a partir dessa empresa OAS para o ex-Presidente Lula por meio de um upgrade de um apartamento de um imóvel, um triplex no Guarujá, por meio da reforma desse triplex, por meio da decoração desse triplex, e por meio de um contrato de armazenamento de bens pessoais, um contrato milionário firmado para armazenamento, um contrato falso firmado pela OAS como se os bens fossem dela e não do ex-Presidente."

O número de palavras é: 2058

SUMÁRIO C

SUMÁRIO C

Figura 3 – Extração do Recurso Extraordinário RE nº 654833 - AC perante o Supremo Tribunal Federal - STF e sumário gerado

Dados da Extração	Dados do Sumário

Fonte: Adaptado pelo autor a partir da interface (AISummarizer, 2022)

Trechos do sumário gerado:

> Em nosso ordenamento jurídico, a regra a prescrição da pretensão reparatória. A imprescritibilidade, por sua vez, exceção. Depende, portanto, de fatores externos, que o ordenamento jurídico reputa inderrogáveis pelo tempo. Embora a Constituição e as leis ordinárias não disponham acerca do prazo prescricional para a reparação de danos cíveis ambientais, sendo regra a estipulação de prazo para pretensão ressarcitória, a tutela constitucional a determinados valores impõe o reconhecimento de pretensões imprescritíveis. O meio ambiente deve ser considerado patrimônio comum de toda humanidade, para a garantia de sua integral proteção, especialmente em relação a gerações futuras. Todas as condutas do Poder Público estatal devem ser direcionadas no sentido de integral proteção legislativa interna e de adesão aos pactos e tratados internacionais protetivos desse direito humano fundamental de 3 geração, para evitar prejuízo da coletividade em face de uma afetação de certo bem recurso natural a uma finalidade individual. A reparação do dano ao meio ambiente direito fundamental indisponível, sendo imperativo o reconhecimento da imprescritibilidade no que toca à recomposição dos danos ambientais.
>
> (...)
>
> Havendo prova dos danos e de terem os réus sido os responsáveis pelas condutas lesivas, devem ser eles condenados a pagarem as indenizações correspondentes. irrelevante o fato de o território indígena ainda não estar demarcado ao tempo dos fatos, pois as normas constitucionais e legais conferem aos índios a exclusiva exploração econômica das riquezas naturais existentes nas terras por eles tradicionalmente ocupadas, mesmo que ainda não tenham sido submetidas à demarcação. Ninguém pode extrair madeira de imóvel pertencente a terceiros indígenas ou não sem a autorização do seu proprietário ou legítimo possuidor seja ele conhecido ou não. O montante da indenização normalmente não se submete a limites mínimo e máximo, tendo como parâmetros básicos a extensão e o valor do dano. Apelações não providas.

SUMÁRIO D

SUMÁRIO D

Experimento em Python de Sumarização com texto jurídico a partir do acórdão de julgamento em Recurso Extraordinário RE nº 654833 - AC perante o Supremo Tribunal Federal - STF

Trechos do sumário produzido:

"De fato, o artigo 2º, inciso I, da Lei nº 6.831/81, ao definir a Política Nacional do Meio Ambiente, dispõe como um de seus princípios a "ação governamental na manutenção do equilíbrio ecológico, considerando o meio ambiente como um patrimônio público a ser necessariamente assegurado e protegido, tendo em vista o uso coletivo".

(...)

Ademais, o próprio Fundo de Defesa de Direitos Difusos, previsto pelo artigo 13 da Lei nº 7.347/85 e regulamentado pela Lei nº 9.008/95, encontra-se na estrutura do Ministério da Justiça e tem natureza pública, e é o destino das condenações ao pagamento do dano ambiental, como ocorre no presente caso.

(...)

Finalmente, no que toca ao dano ambiental relativo às terras indígenas, ressalte-se que a propriedade das áreas é da União, gerenciado pela Fundação Nacional do Índio e pelas próprias comunidades, o que já atrai o regime de direito público para todas as questões referentes à proteção da área, incluindo questões ressarcitórias dos danos socioambientais às comunidades indígenas. Assim, é imprescritível o dano ambiental, nos termos do artigo 37, §5º da Constituição da República.

(...)

O caput do artigo 225 da Constituição Federal assim traduz o direito ao meio ambiente: "Todos têm direito ao meio ambiente ecologicamente

(...)

São reconhecidos aos índios sua organização social, costumes, línguas, crenças e tradições, e os direitos originários sobre as terras que tradicionalmente ocupam, competindo à União demarcá-las, proteger e fazer respeitar todos os seus bens."

O número de palavras é: 5542

SUMÁRIO E

SUMÁRIO E

Figura 5 – Extração do Recurso Especial nº 1.846.649 - MA perante o Superior Tribunal de Justiça – STJ e sumário gerado.

Fonte: Adaptado pelo autor a partir da interface (AISummarizer, 2022)

Trechos do sumário gerado:

> Para os fins do art. 1. 036 do CPC2015, a tese firmada a seguinte Na hiptese em que o consumidorautor impugnar a autenticidade da assinatura constante em contrato bancrio juntado ao processo pela instituio financeira, caber a esta o nus de provar a sua autenticidade CPC, arts. 6, 368 e 429;
>
> (...)
>
> Para os fins repetitivos, foi aprovada a seguinte tese Na hiptese em que o consumidorautor impugnar a autenticidade da assinatura constante em contrato bancrio juntado ao processo pela instituio financeira, caber a esta o nus de provar a autenticidade CPC, arts. 6, 369 e 429, II
>
> (...)
>
> A primeira tese restou assim fixada Independentemente da inverso do nus da prova - que deve ser decretada apenas nas hipteses autorizadas pelo art. 6 VIII do CDC, segundo avaliao do magistrado no caso concreto -, cabe instituio financeirar, enquanto fato impeditivo e modificativo do direito do consumidorautor CPC, art. 373, II, o nus de provar que houve a contratao do emprstimo consignado, mediante a juntada do contrato ou de outro documento capaz de revelar a manifestao de vontade do consumidor no sentido de firmar o negcio jurdico, permanecendo com o consumidorautor, quando alegar que no recebeu o valor do emprstimo, o dever de colaborar com a Justia CPC, art. 6 e fazer a juntada do seu extrato bancrio, embora este no deva ser considerado, pelo juiz, como documento essencial para a propositura
>
> (...)
>
> A segunda tese restou assim fixada A pessoa analfabeta plenamente capaz para os atos da vida civil CC, art. 2 e pode exarar sua manifestao de vontade por quaisquer meios admitidos em direito, no sendo necessria a utilizao de procurao pblica ou de escritura pblica para a contratao de emprstimo consignado, de sorte que eventual vcio existente na contratao do emprstimo deve ser discutido luz das hipteses legais que autorizam a anulao por defeito do negcio jurdico CC, arts. 138, 145, 151, 156, 157 e 158
>
> (...)
>
> A terceira tese restou assim fixada cabvel a repetio de indbito em dobro nos casos de emprstimos consignados quando a instituio financeira no conseguir comprovar a validade do contrato celebrado com a parte autora, restando configurada a m-f da instituio, resguardadas as hipteses de enganos justificáveis
>
> (...)
>
> A quarta tese restou assim fixada 4. No estando vedada pelo ordenamento jurdico, lcita a contratao de quaisquer modalidades de mtuo financeiro, de modo que, havendo vcio na contratao, sua anulao deve ser discutida luz das hipteses legais que versam sobre os defeitos do negcio jurdico CC, arts. 138, 145, 151, 156, 157 e 158 e dos deveres legais de probidade, boa-f CC, art. 422 e de informao adequada e clara sobre os diferentes produtos, especificando corretamente as caractersticas do contrato art. 4 IV e art. 6, III, do CDC, observando-se, todavia, a possibilidade de convalidao do negcio anulvel, segundo os princpios da conservao dos negcios jurdicos CC, art. 170

SUMÁRIO F

SUMÁRIO F

Experimento em Python de Sumarização com texto jurídico a partir do acórdão de julgamento em Recurso Especial RESP nº 1.846.649 - MA perante o Superior Tribunal de Justiça - STJ

Trechos do sumário produzido:

> "2°) e pode exarar sua manifestação de vontade por quaisquer meios admitidos em direito, não sendo necessária a utilização de procuração pública ou de escritura pública para a contratação de empréstimo consignado, de sorte que eventual vício existente na contratação do empréstimo deve ser discutido à luz das hipóteses legais que autorizam a anulação por defeito do negócio jurídico (CC, arts. Embargos declaratórios conhecidos, sendo os 1°s, 2°s, 5°s e 7°s desprovidos; os 4°s embargos parcialmente providos para excluir do acórdão os precedentes deste sodalício de n°s 5499/2016 (Embargos de Declaração) e 18905/2015 (Apelação Cível); e os 3°s, 4°s, 6°s e 8°s parcialmente providos para aclarar a 3a tese que passará a ter a seguinte redação: "Nos casos de empréstimos consignados, quando restar configurada a inexistência ou invalidade do contrato celebrado entre a financeira e a parte autora, bem como, demonstrada a má-fé da instituição bancária, será cabível a repetição de indébito em dobro, resguardadas as hipóteses de enganos justificáveis". 2°) e pode exarar sua manifestação de vontade por quaisquer meios admitidos em direito, não sendo necessária a utilização de procuração pública ou de escritura pública para a contratação de empréstimo consignado, de sorte que eventual vício existente na contratação do empréstimo deve ser discutido à luz das hipóteses legais que autorizam a anulação por defeito do negócio jurídico (CC, arts."

O número de palavras é: 1577

AVALIAÇÃO MANUAL DE QUALIDADE DE SUMÁRIO AUTOMÁTICO DE TEXTO JURÍDICO

Os questionamentos a seguir importam na avaliação dos sumários gerados em experimento de sumarização automática de texto jurídico, que consistiu em resumir os textos-fonte (acórdão) com aplicação de duas técnicas: uma ferramenta preconcebida, denominada Summarizer (AISummarizer), e uma ferramenta constituída a partir da programação modular do Python, restando gerados os seguintes sumários: (1) Acórdão de Julgamento em Recurso Especial nº 1842613 – SP do Superior Tribunal de Justiça – STJ - Sumário A e B; (2) Acórdão de Julgamento em Recurso Extraordinário RE nº 654833 – AC perante o Supremo Tribunal Federal – STF - Sumário C e D; (3) Acórdão de Julgamento em Recurso Especial nº 1.846.649 - MA perante o Superior Tribunal de Justiça - STJ - Sumário E e F.

Em especial, para o objetivo desenvolvido no estudo, interessou investigar a preocupação com a qualidade em sentido amplo dos sumários gerados (PARDO, 2008), notadamente espelhada na coerência (entender o conteúdo), na coesão (entender o relacionamento entre os elementos linguísticos) e na informatividade (entender se a mensagem apresentada continha o conteúdo necessário), no campo da textualidade. Assim, é preciso investigar o real grau do interesse que os textos resultados da experiência podem provocam nos operadores do Direito, sobremaneira, como ferramenta de apoio.

Exatamente pela dificuldade de compreensão e geração de sumários de qualidade, a avaliação humana, também conhecida como manual, ainda costuma ser a mais apropriada, razão pela qual você, como operador do Direito, com mais de dez anos de experiência, foi selecionado.

Diante do panorama apresentado, responda aos seguintes questionamentos:

1. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Recurso Especial nº 1842613-SP, perante o Superior Tribunal De Justiça – STJ, é possível concluir que o sumário A possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?
SIM
NÃO

2. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Recurso Especial nº 1842613-SP, perante o Superior Tribunal De Justiça – STJ, é possível concluir que o sumário B possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?
SIM
NÃO

3. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Recurso Extraordinário RE nº 654833-AC, perante o Supremo Tribunal Federal – STF, é possível concluir que o sumário C possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?
SIM
NÃO

4. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Recurso Extraordinário RE nº 654833-AC, perante o Supremo Tribunal Federal – STF, é possível concluir que o sumário D possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?
SIM
NÃO

5. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Acórdão de Julgamento em Recurso Especial nº 1.846.649-MA perante o Superior Tribunal de Justiça – STJ, é possível concluir que o sumário E possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?
SIM
NÃO

6. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Acórdão de Julgamento em Recurso Especial nº 1.846.649-MA perante o Superior Tribunal de Justiça – STJ, é possível concluir que o sumário F possui boa qualidade em sentido amplo (coerência, coesão e informatividade)?
SIM
NÃO

7. Como operador do Direito, você utilizaria o sumário A como material de apoio para seu trabalho?
SIM
NÃO

8. Como operador do Direito, você utilizaria o sumário B como material de apoio para seu trabalho?
SIM
NÃO

9. Como operador do Direito, você utilizaria o sumário C como material de apoio para seu trabalho?
SIM
NÃO

10. Como operador do Direito, você utilizaria o sumário D como material de apoio para seu trabalho?
SIM
NÃO

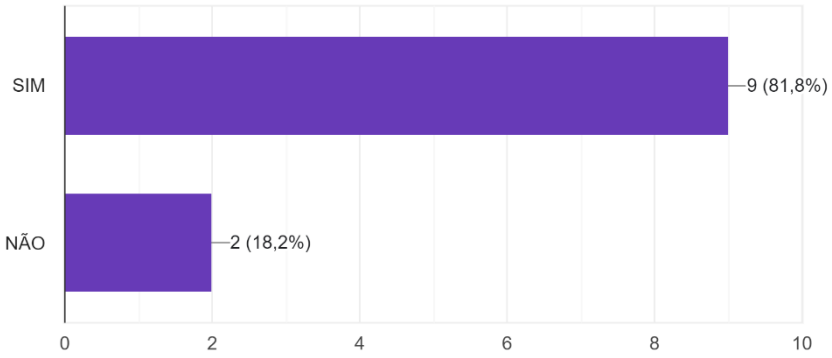
11. Como operador do Direito, você utilizaria o sumário E como material de apoio para seu trabalho?
SIM
NÃO

12. Como operador do Direito, você utilizaria o sumário F como material de apoio para seu trabalho?
SIM
NÃO

13. Como operador do Direito, diante dos sumários gerados automaticamente acima expostos, você entende que seria possível substituir, atualmente, pelo artefato tecnológico, o trabalho de sumarização ou resumo realizado pelo ser humano?
SIM
NÃO

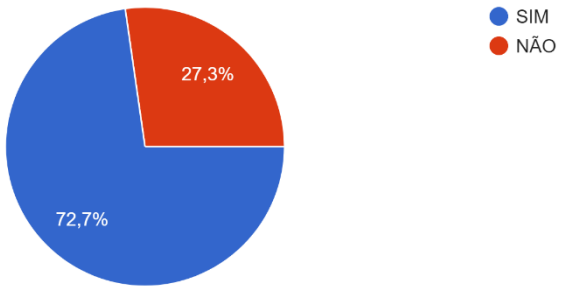
1. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Recurso Especial nº 1842613 – SP, perante o Superior Tribunal De Justi...tido amplo (coerência, coesão e informatividade)?

11 respostas



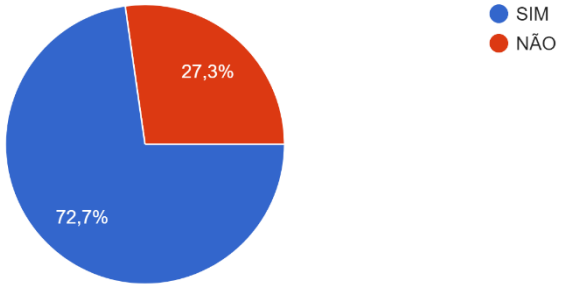
2. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Recurso Especial nº 1842613 – SP, perante o Superior Tribunal De Justi...do amplo (coerência, coesão e informatividade)?

11 respostas

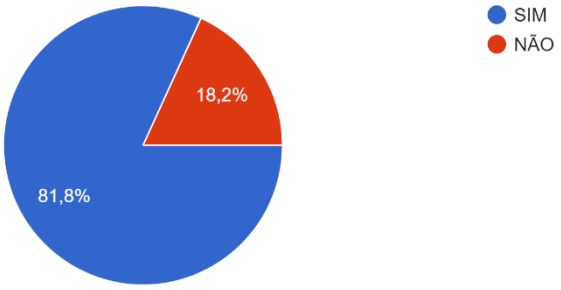


3. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Recurso Extraordinário RE nº 654833 – AC, perante o Supremo Tribunal Federa...ido amplo (coerência, coesão e informatividade)?

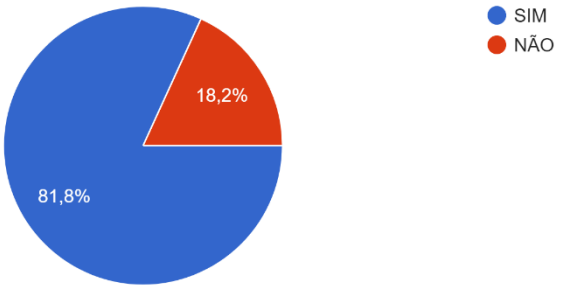
11 respostas



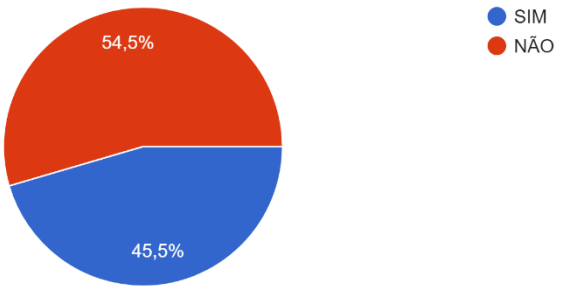
4. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Recurso Extraordinário RE nº 654833 – AC, perante o Supremo Tribunal Federa...ido amplo (coerência, coesão e informatividade)?
11 respostas



5. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Acórdão de Julgamento em Recurso Especial nº 1.846.649 - MA perante o Supe...ido amplo (coerência, coesão e informatividade)?
11 respostas

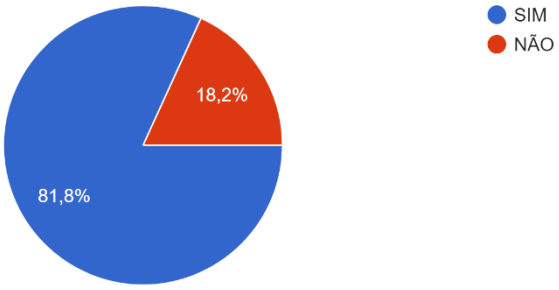


6. Analisando o conteúdo do texto-fonte Acórdão de Julgamento em Acórdão de Julgamento em Recurso Especial nº 1.846.649 - MA perante o Supe...ido amplo (coerência, coesão e informatividade)?
11 respostas



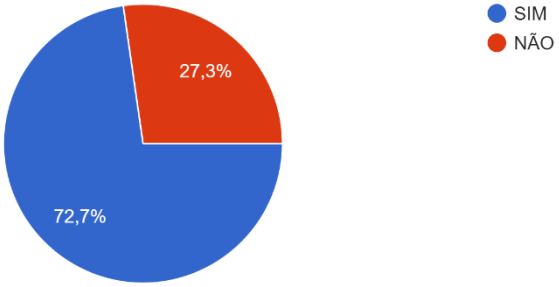
7. Como operador do Direito, você utilizaria o sumário A como material de apoio para seu trabalho?

11 respostas



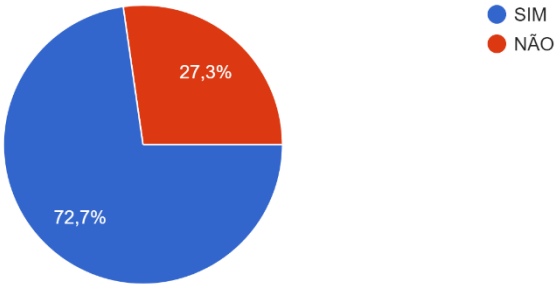
8. Como operador do Direito, você utilizaria o sumário B como material de apoio para seu trabalho?

11 respostas



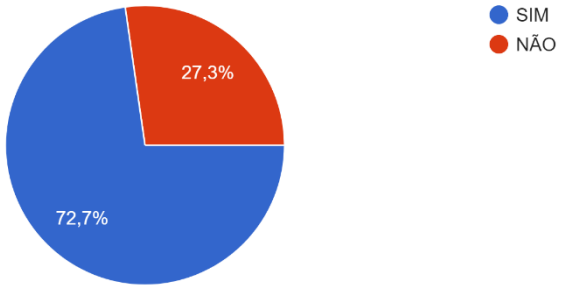
9. Como operador do Direito, você utilizaria o sumário C como material de apoio para seu trabalho?

11 respostas



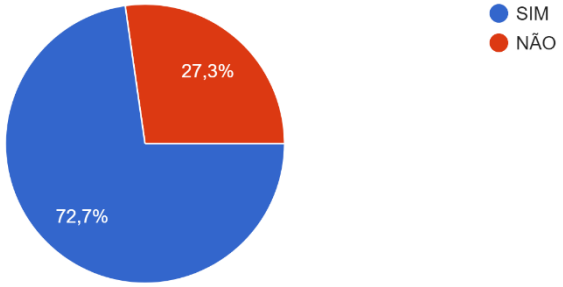
10. Como operador do Direito, você utilizaria o sumário D como material de apoio para seu trabalho?

11 respostas



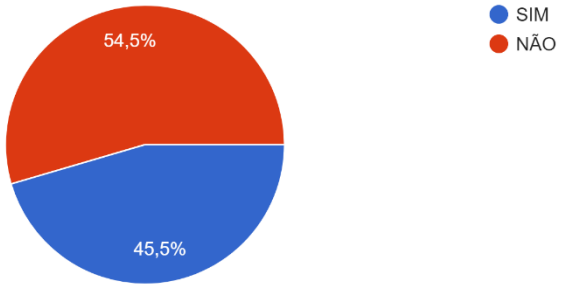
11. Como operador do Direito, você utilizaria o sumário E como material de apoio para seu trabalho?

11 respostas



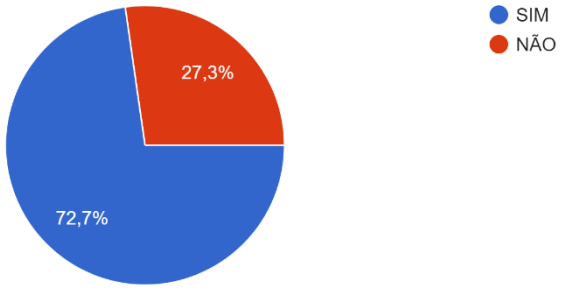
12. Como operador do Direito, você utilizaria o sumário F como material de apoio para seu trabalho?

11 respostas



13. Como operador do Direito, diante dos sumários gerados automaticamente acima expostos, você entende que seria possível substituir, atual... sumarização ou resumo realizado pelo ser humano?

11 respostas



Fonte: Autoria própria