Codex: Possibilidades e Desafios na Análise Automatizada de Jurisprudência

Guilherme Ramos de Morais
Doutorando em Direito na Universidade Federal de Minas Gerais

Inovações, inteligência artificial e tecnologias de informação e comunicação em Sistemas de Justiça

RESUMO

Este relatório técnico analisa o desenvolvimento, as estruturas normativa e algorítmica, as vantagens operacionais e os desafios relacionados à implantação da Plataforma Codex, *data lake* de informações processuais desenvolvido pelo CNJ e que congrega informações processuais de quase todos os tribunais do Brasil. O relatório é subdividido da seguinte forma: introdução; histórico da plataforma Codex; análise de sua base normativa; análise de sua estrutura algorítmica; discussão sobre as vantagens operacionais relacionadas à sua implementação, especificamente a possibilidade de criação de um indexador nacional unificado e de um banco de dados de recuperação semântica adaptado exclusivamente ao direito; desafios relacionados à sua implantação, sobretudo os aspectos regulatórios vinculados ao tratamento das informações desse banco de dados, o que inclui questões sobre a forma de acesso aos dados; segurança informacional (em termos de infraestrutura); e proteção de dados pessoais. Ao final, compilam-se as vantagens e os desafios relacionados à implantação do Codex. Utilizou-se da revisão bibliográfica de literatura especializada e das normas constitucionais e infraconstitucionais sobre o tema.

Palavras-chave: Codex; Processamento de linguagem natural; Recuperação de informações.

ABSTRACT

This technical report analyzes the development, the normative and algorithmic structures, the operational advantages, and the challenges related to the implementation of the Codex Platform, a data lake of procedural information developed by the CNJ that gathers case information from almost all courts in Brazil. The report is divided as follows: introduction; history of the Codex platform; analysis of its normative basis; analysis of its algorithmic structure; discussion of the operational advantages related to its implementation, specifically the possibility of creating a unified national indexer and a semantic retrieval database designed exclusively for law; challenges related to its implementation, especially regulatory aspects concerning the processing of the information in this database, including issues related to data access;

information security (in terms of infrastructure); and personal data protection. Finally, the advantages and challenges related to the implementation of Codex are compiled. The report is based on a bibliographic review of specialized literature and on constitutional and infra-constitutional norms on the subject.

Keywords: Codex; Natural language processing; Information retrieval.

1. Introdução

O conto da Biblioteca de Babel, escrito pelo argentino Jorge Luis Borges em 1941, apresenta uma biblioteca incomum. Trata-se de um prédio composto por um número indefinido, aliás, *infinito*, de galerias hexagonais, das quais se observam infinitos andares, tanto para cima, quanto para baixo. Cada galeria possui vinte prateleiras com cinco estantes; cada estante possui trinta e dois livros uniformes; cada livro possui quatrocentas e dez páginas; cada página possui quarenta linhas; cada linha tem oitenta letras de cor preta. Por ser infinita, tudo que já foi ou será escrito, necessariamente está contido na biblioteca. Para além das análises transcendentais suscitadas pelo conto, a biblioteca possui uma característica única: é um repositório unificado de um conhecimento infinito.

Esse tipo de gestão documental do conhecimento possui vantagens e desvantagens. A sua maior vantagem é a possibilidade de congregar *todo o conhecimento* disponível em um único lugar. Contudo, essa mesma vantagem implica um desafio adicional. O volume total de informações faz com que o resgate daquelas que sejam efetivamente relevantes não seja simples. Adicionalmente, essa centralização também torna o repositório documental um alvo único para investidas de terceiros mal-intencionados.

Semelhante à proposta da Biblioteca de Babel, por volta de 2020, o Conselho Nacional de Justiça (CNJ) iniciou a implementação da Plataforma Codex. Trata-se de um *data lake* de informações processuais de todo o Brasil, ou seja, um repositório unificado para a gestão documental de processos eletrônicos. No momento de elaboração deste relatório técnico, em setembro de 2025, mais de 96% de todas as informações processuais brasileiras já haviam sido carregadas à plataforma [1]. Em razão de sua robustez, o Codex tem aptidão para gerar um sem-número de aplicações positivas para a estruturação e recuperação de informações. Contudo, as poucas informações públicas ao seu respeito geram questionamentos sobre sua finalidade, além de suscitar dúvidas sobre acessibilidade dos dados e segurança informacional.

Feitas essas considerações iniciais, este relatório técnico tem como propósito analisar a história do desenvolvimento, estrutura normativa e algorítmica, vantagens operacionais e desafios relacionados à implantação da Plataforma Codex, além de aspectos regulatórios da aplicação de modelos de processamento de linguagem natural para a análise desse tipo de banco de dados. A organização textual segue essa mesma lógica linear:

introdução; histórico da plataforma Codex; análise de sua base normativa; análise de sua estrutura algorítmica; discussão sobre as vantagens operacionais relacionadas à sua implementação, especificamente a possibilidade de criação de um indexador nacional unificado e de um banco de dados de recuperação semântica adaptado exclusivamente ao direito; desafios relacionados à sua implantação, sobretudo os aspectos regulatórios vinculados ao tratamento das informações desse banco de dados, o que inclui questões sobre a forma de acesso aos dados; segurança informacional (em termos de infraestrutura); e proteção de dados pessoais. Ao final, apresentam-se as considerações finais com a compilação das vantagens e desafios relacionados à implantação do Codex.

A metodologia adotada é de revisão da literatura sobre o tema e dos atos normativos relacionados à plataforma. Para fins de esclarecimento ético e transparente no uso da inteligência artificial, ressalta-se que a revisão gramatical e sugestões de eventuais melhorias na redação original foram feitas por ferramentas de inteligência artificial generativa, como o GPT 5.0 e o Gemini AI Pro 2.5. Nenhuma inteligência artificial generativa foi utilizada como substituta do autor ou para a produção originária de ideias.

2. Planejamento, desenvolvimento e implementação do Codex

O Codex, segundo o próprio CNJ, é resultado de uma parceria entre o CNJ e o Tribunal de Justiça de Rondônia (TJRO), celebrada por volta de 2020 [2]. A parceria entre o CNJ e o TJRO para o desenvolvimento de soluções tecnológicas de gestão de dados judiciais, contudo, remonta ao ano de 2018, momento em que houve a assinatura de termo de cooperação técnica que permitiu ao CNJ utilizar em âmbito nacional o Sistema Sinapses do TJRO [3]. Apesar desses fatos, o CNJ, oficialmente, no portfólio do Programa Justiça 4.0, afirma que o Codex foi criado em 2022 [4]. Essa informação contradiz o próprio Anexo II da Portaria n. 118 do CNJ, que, desde 13 de abril de 2021, dispõe especificamente sobre o portfólio de soluções de tecnologia da informação e comunicação e serviços digitais do CNJ. Nessa portaria, o Codex foi listado entre as tecnologias do CNJ.

No que diz respeito ao seu histórico, as informações disponíveis sobre a sua origem aqui se encerram: há pouco, quase nenhuma, informação pública confiável sobre sua efetiva origem, pessoas, órgãos públicos envolvidos, e recursos financeiros relacionados ao seu desenvolvimento.

Do ponto de vista normativo, sem intenção de exaurir a análise das normas constitucionais e infraconstitucionais sobre o assunto, convém destacar o contexto atual em que se inseriu o Codex. No ano de 2020, o CNJ instituiu a Base Nacional de Dados do Poder Judiciário (DATAJUD) pela Resolução CNJ n. 331/2020, cujos dados eram carregados pelos tribunais ainda de forma não automatizada (Art. 4º). O principal objeto de análise do

DATAJUD eram os metadados processuais (numeração processual unificada, conforme a Resolução CNJ n. 65/2008; código da Tabela Processual Unificada (TPU), conforme a Resolução CNJ n. 46/2007; dados das partes, e afins, art. 6º). Logo em seguida, na Resolução CNJ n. 335/2020, o CNJ estabeleceu a Plataforma Digital do Poder Judiciário Brasileiro (PDPJ-Br) como forma de integração dos sistemas de todos os tribunais do país, priorizando-se, contudo, o desenvolvimento do PJe (art. 15).

Por não haver uniformização das plataformas processuais, um sem-número de sistemas de tramitação processuais surgiram: PJe, e-Proc, Projudi, Themis, e-SAJ, além de plataformas autorais dos tribunais. Convém destacar que, de fato, houve tentativa de uniformizar o PJe como sistema único de tramitação processual. A Resolução CNJ n. 185/2013 instituía o PJe como sistema nacional de tramitação processual (art. 1º), e os arts. 34 e seguintes determinavam a implantação do PJe em prazos específicos, alcançando-se a implementação total em 2018. Ao final do prazo, os Tribunais resistiram à implementação definitiva do PJe sob justificativas diversas, que variavam das deficiências técnicas do PJe à perda de investimentos realizados nas plataformas próprias já desenvolvidas (Migalhas, 2019). Essa resistência não foi superada ao longo do tempo, o que levou à adoção do modelo de interoperabilidade da PDPJ-Br.

3. Características técnicas do Codex

É nesse contexto de integração, automatização e aumento exponencial do volume de informações produzidas, é que se insere o Codex, cuja Resolução CNJ n. 446, de 14 de março de 2022, constitui seu principal subsídio legal. Nessa resolução, o CNJ instituiu o Codex como ferramenta oficial de extração de dados (estruturados e não estruturados) dos processos judiciais eletrônicos em tramitação no Poder Judiciário Nacional, com a exceção aos dados do Supremo Tribunal Federal e do próprio Conselho Nacional de Justiça [5]. Por sua vez, a Portaria CNJ n. 183, de 2 de junho de 2022, dispôs que, também com exceção do Supremo Tribunal Federal, todos os tribunais deveriam integrar seus sistemas de gestão de processos judiciais ao Codex até o dia 30 de junho de 2022, excepcionados, ainda, os sistemas administrativos, os de acompanhamento processual de processos físicos e os de acompanhamento de processos eletrônicos que fossem integralmente desativados até o dia 30 de junho de 2022. O art. 7º da Resolução CNJ n. 446/2022 finalmente instituiu o envio automatizado das informações das bases dos Tribunais [6].

Quanto à estrutura algorítmica, o código-fonte da plataforma não é aberto. Contudo, é possível deduzir parcialmente sua arquitetura, seja pelo fato de ser um *data lake,* seja pelas informações divulgadas de forma esparsa pelos tribunais.

Conforme a Resolução CNJ n. 446/2022, cada tribunal implementa a ferramenta de extração de dados estruturados e não estruturados em sua infraestrutura (art. 3º, § 1º). Como o próprio art. 1º sugere, os dados não estruturados se referem a textos (possivelmente integrais), imagens, planilhas etc. Os dados estruturados, por sua vez, provavelmente se referem aos metadados (classe processual, partes, órgão julgador, assunto segundo a TPU, tudo a exemplo dos metadados processuais do DATAJUD, conforme o art. 6º da Resolução CNJ n. 331/2020). Certamente esses dados são convertidos em uma linguagem padronizada (comumente no formato JSON para esse tipo de aplicação) e posteriormente tokenizados em razão do volume de dados para fins de otimização do processamento. [7] Trata-se, portanto, de um processo padrão de ETL (*extract, transform, load)* [8]. Uma vez carregados os dados na plataforma central, podem ser utilizados mecanismos de mineração de dados, inclusive para posterior utilização em treinamento de modelos de aprendizado de máquina e aplicações afins, aspecto que será analisado adiante.

Em razão da ausência de transparência, as considerações sobre o funcionamento técnico feitas acima são fundadas na praxe observada em bases semelhantes. O funcionamento real do Codex, porém, não é conhecido, o que faz surgir questões referentes à transparência na coleta, tratamento e utilização dos dados, aspectos que serão analisados em tópicos específicos deste relatório.

4. Discussão sobre possíveis impactos na melhoria da recuperação informacional pelo Codex

Feitas as considerações sobre a história, os marcos legais e a estrutura algorítmica do Codex, um *data lake* de informações judiciais não apenas é bem-vindo no contexto da litigância massiva brasileira, como também é absolutamente necessário. Conforme exposto anteriormente por Morais (2021) e Morais (2023), o acesso à informação processual constitui pressuposto pragmático para o funcionamento do sistema de precedentes. Não é possível estruturar um sistema de precedentes se, antes de tudo, o próprio precedente não for passível de conhecimento.

Diversas possibilidades de aplicação surgem a partir de um banco de dados robusto do porte já alcançando pelo Codex. Adiante, destacam-se oportunidades únicas e estruturais para a gestão de dados documentais no contexto brasileiro. Referidas propostas não focam estritamente na técnica computacional utilizada, que podem e efetivamente variam conforme a estratégia de análise de dados utilizada por cada gestor da informação.

4.1. Indexação Nacional Unificada

A consequência imediata do Codex é autorreferencial: o Codex é uma base de dados que retroalimenta o aprimoramento de ferramentas de buscas de seus dados. Isso porque, por

ser um *data lake*, sua imensa base de dados serve para desenvolver modelos de treinamento de máquinas. [9] Em documentos legais, o objetivo principal dos modelos de PLN diz respeito à otimização da busca e recuperação de informações (Bolasco et al., 2005, p. 39). Tanto é assim que a vantagem estratégica de países como a China no campo da inteligência artificial não é propriamente a mão de obra qualificada, mas sim o imenso conjunto de dados à disposição para o treinamento de modelos de aprendizado (Filipova, 2024, p. 48). Nesse mesmo sentido, o fato de o Brasil possuir um banco de dados robusto e unificado, adaptado exclusivamente ao setor jurídico, proporciona o ambiente ideal para a aplicação de técnicas de processamento de linguagem natural (PLN). Katz et al. (2023, p. 10) demonstraram que o aumento do volume de informações disponível para análise é elemento fundamental para o desenvolvimento do PLN.

Por outro lado, o volume de informações não pode ser tido como elemento central isoladamente. Não necessariamente o maior volume de dados repercute em melhor qualidade da análise dos dados. Uma informação de entrada classificada incorretamente, provavelmente gerará um resultado de saída indesejado (Datatilsynet, 2018, p. 11). É no mesmo sentido a consideração de Nunes e Marques (2019, pp. 48 e 50). No campo dos *large language models*, é bem verdade que a quantidade de informação não necessariamente reflete positivamente nos conceitos de revocação e precisão do resgate informacional. Ainda assim, o avanço sistêmico da inteligência artificial proporciona o desenvolvimento de modelos incrementais de resgate informacional. Portanto, a utilização desses modelos tornou-se um imperativo de ordem pragmática.

No que diz respeito à unificação de bases de dados, ao analisar o problema de acessibilidade de dados jurisprudenciais na Europa, Filtz, Kirrane e Polleres (2018) já indicavam a necessidade de uma base unificada sobre processos judiciais no contexto europeu. É o que popularmente têm se denominado, no ramo jurídico, de *legal open data.* Embora o sistema europeu de jurisdição tenha fatores complicadores que inexistem no Brasil (multiplicidade de idiomas, tradições jurídicas e bases normativas), o aspecto da multiplicidade de sistemas de armazenamento de informações jurisprudencial é semelhante em ambos os sistemas jurídicos. Tanto é assim que, entre os fatores complicadores da análise jurisprudencial dos Estados europeus, tem-se a necessidade de criação de *parsers* que estruturam os documentos legais levando-se em consideração as diferenças linguísticas e estruturais dos documentos. Portanto, no que diz respeito à unificação de bases de dados, a simples existência do Codex é um avanço importante para a gestão documental brasileira, capaz de alavancar a popularização do PLN em língua portuguesa, sobretudo em razão da extensão do repositório informacional.

E, sobre a análise da informação em si, há uma série de fatores que, considerados isolada ou cumulativamente, podem impulsionar significativamente a eficiência da mineração de dados e o avanço das técnicas de indexação.

Indexação nada mais é do que uma forma de resumir e representar o conteúdo de documentos (Lancaster, 2004, p. 3-6; Savoy, Gaussier, 2010, p. 456). Por sua vez, a função do indexador é prever as necessidades de determinado usuário em potencial e fornecer os meios para que essa necessidade seja atendida, calibrando-se os índices de revocação e precisão da análise (Lancaster, 2004, p. 17). Dito de forma simplificada, revocação é a capacidade de resgatar os documentos relevantes no universo de todos os documentos relevantes que seriam passíveis de resgate; precisão, por sua vez, é um conceito negativo, relacionado ao *não resgate* de documentos *irrelevantes* (Morais, 2023, p. 116). A relação entre os dois índices tende a ser inversamente proporcional: a melhoria da revocação implica perda de precisão e vice-versa (Lancaster, 2004, p. 4). E, maior a base de dados, menos aceitável a baixa precisão (Lancaster, 2003, p. 4). Essa máxima é particularmente adequada ao Codex, em razão da amplitude dessa base de dados.

Uma indexação eficiente é aquela capaz de ajustar os índices de precisão e revocação conforme a sua finalidade. Além disso, há outros fatores que contribuem para a eficiência da indexação, especialmente a necessidade de delimitar os pontos de interesse do documento, sendo essencial que o indexador *conheça* as necessidades dos usuários potenciais (Lancaster, 2004, p. 15-17 e 24). Ou seja, cada indexação demanda critérios próprios de análise, conforme a finalidade da pesquisa de informação que se empreende.

Feitas essas considerações iniciais sobre a indexação, o *big data* jurisprudencial brasileiro possui elementos facilitadores para sua aplicação consolidados no próprio Codex. O primeiro elemento facilitador é a existência de um *piso* de indexação temática, que são os termos apresentados na TPU. O aprofundamento dos temas que já estão dispostos na TPU, permite economia de tempo no agrupamento de informações e o aprimoramento dos índices de revocação e precisão. Utilizando-se da síntese feita por Puente e Coalla (2023, p. 113), no campo da análise textual, quanto maior o tamanho do *corpus* (banco de dados), menor é o aparecimento de itens inéditos. Os autores explicam essa correlação no sentido de que os *tokens* (entendidos, em sentido amplo, como unidades de análise) aumentam linearmente pelo tamanho da base de dados, mas os *tipos de tokens* tendem a se repetir. [10] Lancaster (2004) também elenca uma série de técnicas que podem contribuir para o aprimoramento desse piso de indexação, tais como, análise conceitual [11] e tradução [12]; técnicas de elaboração de vocabulário controlado [13]; atribuição de pesos e padrões ao vocabulário obtido [14]; e a adequação dessas técnicas à extração automatizada.

O desenvolvimento das técnicas expostas cima, meramente exemplificativas, em razão da sempre possível elaboração de novas técnicas, contribuem para a construção de um sistema de recuperação de informações voltados à necessidade do usuário final. A base de dados pré-estruturada do Codex torna tecnicamente viável a aplicação dessas técnicas, que podem ser potencializadas com a utilização de modelos baseados em inteligência artificial.

O segundo elemento facilitador para a indexação do *big data* jurisprudencial brasileiro diz respeito à uniformização da estrutura das decisões judiciais. A uniformização da estrutura textual permite que os algoritmos de análise documental sejam programados de forma a otimizar sua análise nos pontos que possuem maior relevância documental. Trata-se da primeira diretriz teórica para o desenvolvimento do processamento de linguagem natural no direito, exposta por Morais (2023, pp. 121-126). Palmer (2010, p. 14), considera que o problema de alguns *corpora* de análise documental é, precisamente, a variedade de formatações tipográficas.

Ainda em estágio incipiente, a Recomendação CNJ n. 154/2024, estabelece um modelo de ementas judiciais. Ou seja, há uma tendência de centralização e uniformização que não se restringe à forma de armazenamento (a plataforma em si, o Codex), ou aos *hubs* de aplicações, como o BDPJ. A centralização regulatória do CNJ tende a se estender à própria formatação documental. Por um lado, a uniformização é um aspecto positivo para a análise massificada de dados, porque acelera a análise automatizada. Torna-se desnecessário que o mecanismo extrator e indexador mapeie as estruturas documentais a partir do zero, o que economiza tempo de processamento.[15] Elimina-se parcialmente um dos elementos do pré-processamento, qual seja, a *normalização documental*, que é a conversão de um conjunto de arquivos digitais em um documento formatado (Câmara Júnior, 2013, p. 38).

A partir da análise da Recomendação CNJ n. 154/2024, constata-se uma tendência normativa à uniformização de macro elementos textuais, a começar pelas ementas dos julgados. O avanço da normalização para a uniformização de microelementos textuais, porém, é mais complexo, senão impossível. Seria uma tentativa de uniformizar a própria estrutura de exposição de ideias dentro de unidades menores do texto (parágrafos, frases), eliminando-se também os vieses linguísticos do escritor. Esse tipo de uniformização extinguiria características próprias do escritor, o que não é nem sequer desejável na atividade-fim jurisdicional.

Esses dois aspectos facilitadores (um piso de indexação temático pré-estabelecido e o início das políticas de uniformização de elementos macrotextuais) aliados à robustez do Codex tornam possível a indexação nacional unificada aplicada exclusivamente ao direito. Sua implementação poderá aprimorar o balanceamento entre os índices de precisão e revocação,

otimizando-se o resultado esperado pelo usuário final, consistente no resgate da informação efetivamente desejada.

4.2. Recuperação Semântica Nacional

A segunda vantagem operacional do Codex diz respeito à análise da semântica dos textos judiciais. Originalmente, as técnicas de recuperação de informação focavam em ramos especializados do conhecimento. Ou seja, para cada área do conhecimento, haveria uma forma de recuperação específica. Contudo, em razão do advento da internet, passou-se a utilizar a recuperação de informações em vários documentos simultaneamente (análise *multi-document / open domain*) (Hobbs, Riloff, 2010, pp. 514-515). Contudo, a análise não especializada se mostra problemática para a compreensão semântica, que é eminentemente contextual.

Há décadas já se havia observado que, para fins de análise temática, a mineração de dados textuais deve considerar, invariavelmente, o contexto em que o documento se insere (Tan, 1999, p. 5). A bem da verdade, uma ferramenta de indexação tem por objetivo final *compreender* o sentido da palavra para corretamente classificá-la, o que, de fato, varia conforme o contexto (Nastase, Mihalcea, Radev, 2015, p. 681). Uma base de dados do porte do Codex pode fornecer os subsídios necessários para a criação de um dicionário semântico adaptado ao direito.

Como exposto por Yarowsky (2010, p. 315), a desambiguização é um problema, essencialmente, de classificação. Uma das formas de se obter uma classificação robusta e eficiente é pela criação de um inventário semântico (Yarowsky, 2010, p. 315; Hobbs, Riloff, 2010, p. 526). Referido dicionário nada mais é do que um aprimoramento da indexação exposta no subcapítulo anterior, com relevante impacto na eficiência do resgate informacional. Para sua construção, é possível a utilização de técnicas de elaboração de tesauros (*bag-of-words)* [16], tagueamento [17], hierarquia de conceitos [18], atribuição de pesos [19], entre tantas outras.

Independentemente da técnica a ser utilizada, a criação de um dicionário semântico adaptado à análise de decisões judiciais, à semelhança da criação do indexador nacional unificado, é uma aplicação avançada do PLN possivelmente viabilizada pela utilização do Codex. O desenvolvimento de ambas as técnicas tem potencial para aprimorar substancialmente o resgate da informação jurisprudencial brasileira.

5. Desafios regulatórios relacionados à implementação do Codex e sua integração a modelos de IA

Expostas as vantagens operacionais oferecidas pelo Codex, convém analisar os desafios relacionados à sua implementação e a regulação de modelos de processamento de

linguagem natural que fazem uso desse formato de base de dados massiva, parcialmente estruturada, e potencialmente portadora de dados sensíveis.

Entre os elementos que demandam atenção regulatória contextualizada e setorial para o desenvolvimento do PLN com dados do Codex, foram consideradas questões como acesso aos dados; segurança no armazenamento das informações (segurança informacional); e proteção do direito fundamental à proteção de dados pessoais.

5.1. Acesso aos dados e APIs

O acesso aos dados do Codex perpassa três questões: quem acessa? Sob qual fundamento? Por qual meio?

Por exigência constitucional, presume-se que a publicidade é a regra, e o sigilo é a exceção no caso dos atos processuais (art. 5º, inciso LX; e art. 93, inciso IX da CF). Por se tratar de garantia constitucional, a publicidade dos atos processuais é cláusula pétrea nos termos do art. 60, § 4º, inciso IV da Constituição. É no mesmo sentido o art. 189 do Código de Processo Civil (CPC): é excepcionalíssima a decretação do segredo de justiça, seja sobre atos específicos, seja sobre o processo como um todo. Referida disposição infraconstitucional se aplica por extensão a aos processos eleitorais, trabalhistas e administrativos, em razão da aplicação subsidiária e supletiva do art. 15 do CPC. Dito isso, presume-se que a maior parte dos dados que compõem o Codex são públicos, e a menor parte é sigilosa, em razão da correlação entre a regra da publicidade e a exceção do sigilo.

Por se tratar de garantia constitucional fundamental, qualquer brasileiro tem direito subjetivo a acessar as informações processuais armazenadas no Codex, excepcionadas as de natureza sigilosa. É nesse mesmo sentido a Resolução n. 121/2010 do CNJ, que restringe o acesso público a determinadas peças processuais, fato questionável do ponto de vista constitucional. Contudo, é necessário esclarecer *por qual meio* se acessam os dados.

Até o momento de elaboração deste artigo, os únicos dados da plataforma Codex disponíveis publicamente são aqueles do painel de monitoramento ao qual foi feita a referência na nota de fim de texto número 1. Ou seja, trata-se de uma informação estatística, e não de acesso aos dados em si. Contudo, a Resolução n. 574/2024 do CNJ indicou que provavelmente será disponibilizado uma *Application Programming Interface* (API) para acesso ao repositório do Codex, mediante uma contrapartida (financeira ou não) específica. Por sua vez, a Portaria n. 316/2024 regulamenta elementos técnicos para viabilizar esse acesso, entre os quais o cumprimento de protocolos de segurança.

A viabilização das APIs parece ser o objetivo principal do Codex, porque é esperado que *data lakes* desse porte sirvam ao acesso massivo das informações. Como forma de implementar efetivamente os princípios da *open justice* e, ao mesmo tempo, resguardar a

proteção informacional, a utilização das APIs viabilizaria forma específica de acesso aos dados, o que reduziria a sobrecarga de extratores sobre os sistemas de outros tribunais, ao mesmo tempo em que se asseguraria o acesso eficiente aos dados para posterior tratamento, mineração, correlação de parâmetros, entre outras aplicações jurimétricas.

Embora referido acesso por API não esteja plenamente regulamentado e disponibilizado publicamente, é esperado que toda a sociedade tenha acesso ao Codex para análise massiva de dados processuais (públicos).

5.2. Segurança informacional

Segundo a clássica tríade de Saltzer e Schroeder (1975, p. 5), as violações informacionais são de três tipos: vazamento não autorizado de dados; modificação não autorizada de informações; e negativa não autorizada de acesso. No caso do Codex, essas três vulnerabilidades analogamente seriam o vazamento de dados com sigilo protegido por lei; a modificação de documentos processuais; e a negativa de acesso a usuário ou grupo de usuários de forma ilegítima.

O vazamento não autorizado de dados é o desafio mais evidente do Codex. Isso porque princípio básico da segurança informacional é colocação de informações em repositórios distintos. Trata-se do princípio da segregação informacional do *least common mechanism*, que é um dos oito princípios apresentados por Saltzer e Schroeder (1975). Esse princípio preconiza que, se houver um único mecanismo/componente compartilhado e usado por todos, basta vulnerar esse único ponto de falha.

O segundo tipo de vulnerabilidade informacional, consistente na modificação dos dados, evoca a recente experiência de *ransomware* do acervo do STJ (STJ, 2020). Em 3 de novembro de 2020, o sistema de gestão processual do STJ foi alvo de um ataque cibernético em que os invasores impediram o tribunal de acessar o seu acervo processual, criptografaram seus dados e exigiram recompensa para a sua devolução. Não se tem notícia de nenhum suspeito formalmente indiciado e os dados não foram recuperados. A solução encontrada pelo STJ foi utilizar uma cópia de segurança. Esse exemplo é didático no sentido de que, apesar dos avanços em termos de segurança informacional em grandes bancos de dados, sobretudo os estatais, *sempre* existem vulnerabilidades. A segurança informacional consiste em identificar justamente as vulnerabilidades e saná-las, ou mitigá-las, sobretudo com estratégias de redundância.

Uma semana após o ataque cibernético ao STJ, no dia 10 de novembro, o CNJ editou a Portaria n. 242/2020, e instituiu o Comitê de Segurança Cibernética do Poder Judiciário, que tinha por objetivo organizar quatro documentos estratégicos principais: o Protocolo de Prevenção a Incidentes Cibernéticos [20]; o Protocolo de Gerenciamento de Crise Cibernéticas [21]; o Protocolo de Investigação desses ilícitos [22]; e a Estratégia de Segurança

Cibernética e da Informação do Judiciário [23]. Atualmente, a Portaria CNJ n. 162 de 10 de junho de 2021 congrega os três protocolos de segurança informacional em seus anexos.

O primeiro protocolo, qual seja, o de Prevenção de Incidentes Cibernéticos do Poder Judiciário, consta do Anexo I da Portaria CNJ n. 162/2021. O foco desse protocolo consiste nas diretrizes de capacitação de equipes responsáveis pela segurança cibernética do Judiciário; e na identificação e contenção de ameaças. Entre as diversas diretrizes, destaca-se a declaração da boa prática de segurança no sentido de que a segurança cibernética é um *empreendimento coletivo*.

O segundo protocolo, qual seja, o de Gerenciamento de Crise Cibernética, consta do Anexo II da Portaria do CNJ n. 162/2021. Diferentemente do primeiro protocolo, esse possui procedimentos concretos e estruturados para lidar com as situações pré-crise, de crise e pós-crise. No que diz respeito à pré-crise, tem-se a definição de atividades críticas e o estabelecimento de protocolos específicos. Na crise, aplicam-se protocolos pré-definidos (com adaptações conforme o tipo de crise, vide item 5.8). E, no pós-crise, a identificação das falhas de segurança que ocasionaram a crise, elaboração de relatório com estratégias para mitigar novas crises, e elaboração de estratégias de recuperação.

O terceiro e último protocolo, qual seja, o de Investigação de Ilícitos Cibernéticos do Poder Judiciário, consta do Anexo III da Portaria n. 162/2021 do CNJ, no qual são estabelecidas orientações para a preservação de evidências de autoria e materialidade do ataque cibernético, e o itinerário de comunicação do ocorrido aos órgãos de polícia judiciária e ao Ministério Público, especificamente por meio de um *Relatório de Comunicação de Incidente de Segurança em Redes Computacionais*.

Encerram a Estratégia Nacional de Segurança Cibernética do Poder Judiciário, os anexos IV a VIII da Portaria CNJ n. 162/2021. Os Anexos IV a VI são manuais de referência com normas técnicas de proteção de infraestruturas, prevenção, mitigação de ameaças gestão de identidade e controle de acessos; o anexo VI é um manual de referência de política de educação e cultura cibernética; e o Anexo VIII é um glossário.

No que se refere especificamente ao Codex, sua opacidade impede a verificação, pelo público externo, do atendimento de sua estrutura à Estratégia Nacional de Segurança Cibernética do Poder Judiciário. Não existe um formato dialógico entre o CNJ, universidades, órgãos de polícia judiciária, a Ordem dos Advogados do Brasil, entre outros atores diretamente interessados na tutela dos dados jurídicos do Codex. Esse processo eminentemente fechado de gestão informacional viola a boa prática de segurança informacional do Anexo I (Protocolo de Prevenção de Incidentes Cibernéticos) da Portaria CNJ n. 162/2021, e a própria diretriz de transparência do Anexo V da mesma portaria (manual de referência de prevenção e mitigação de ameaças), especificamente nos itens 13.1, f, e 12.1, f. Ou seja, o CNJ reconhece a

transparência como elemento importante para a gestão de risco informacional, e, ao mesmo tempo, limita essa transparência com quase total opacidade sobre o Codex.

Por consequência, a opacidade impede que as vulnerabilidades, que certamente existem em maior ou menor grau, sejam objeto de escrutínio público para fins de aperfeiçoamento do sistema. Se, de um lado, o acesso à sua estrutura poderia torná-la mais vulnerável a ataques perniciosos, a sua abertura também viabilizaria uma política de aprimoramento conjunta. Essa, inclusive, é a proposta do *open design* há muito apresentada por Saltzer e Schroeder (1975).

Assim, em termos de governança de dados, há lacunas que devem ser objeto de melhorias governamentais pelo CNJ, especialmente quanto à política de segurança informacional do Codex.

5.3. Proteção de dados pessoais

Sob a perspectiva da infraestrutura, a segurança informacional tutela o receptáculo: sistemas que armazenam e processam os dados. Sob a perspectiva dos dados em si, a proteção de dados tutela o próprio *conteúdo* armazenado, ou seja, os dados. E, em se tratando de dados pessoais, esses integram os direitos de personalidade da pessoa natural (por interpretação sistêmica entre o art. 5º, incisos, X, XII e LXXIX da Constituição Federal, arts. 11 e ss. do Código Civil, e 2º, incisos I, IV e VII da Lei Federal n. 13.709/2018, LGPD, tal qual decidido pelo Supremo Tribunal Federal na ADI n.º 6.387/DF), e, por extensão, da pessoa jurídica (art. 52 do Código Civil).

No que diz respeito à tutela dos dados do Codex, o art. 9º da Resolução CNJ n. 446/2022 fez referência direta à LGPD. A referência expressa à aplicação da LGPD supera qualquer discussão a respeito de sua aplicação. E, embora não fosse necessária essa referência expressa, as disposições da LGPD seriam necessariamente aplicadas em razão do seu art. 3º, *caput*, e da não incidência de nenhuma hipótese do art. 4º, ambos dessa mesma lei.

Assim, é necessário diferenciar os tipos de dados que constam de um processo judicial para, só então, verificar o tratamento legal que deve ser atribuído a cada um deles. Por exigência legal, um processo se inicia com a apresentação dos dados pessoais do autor e do réu, salvo o desconhecimento dessas informações, que deverá ser objeto de diligências judiciais para sua obtenção (art. 319, inciso II e § 1º do CPC). E, a depender da natureza da discussão objeto do processo, é possível que os autos contenham dados pessoais sensíveis. É o caso, por exemplo, de demandas de saúde, nas quais é necessária a apresentação de exames médicos.

Da análise da Resolução CNJ n. 446/2022, o art. 2º, § 3º dispõe genericamente que compete ao CNJ a proteção de confidencialidade, controle de acesso e níveis de sigilo dos documentos. Porém, são desconhecidos os procedimentos técnicos aplicáveis pelo CNJ para

alcançar esse objetivo. Tal como exposto no subcapítulo anterior, a opacidade sobre a forma efetiva de gestão dos dados impede que pessoas ou entidades não integrantes do CNJ contribuam na indicação de pontos efetivos de melhora. E, mais ainda, impede que se verifique o nível de conformidade do Codex da própria LGPD.

6. Conclusões e recomendações

O Codex se revela como importante, senão o principal, repositório de documentos judiciais brasileiros e, possivelmente se consolidará como um dos maiores do mundo em razão do contexto de litigância massiva em que se insere.

Em razão da amplitude do seu banco de dados, tem potencial de contribuir para a criação de sistemas de indexação nacional unificada e de recuperação semântica nacional, ambos adaptados ao contexto jurídico. Referidas técnicas de tratamento de dados podem incrementar significativamente os índices de eficiência da recuperação e análise jurisprudencial. Deve-se destacar a governança de acesso dos dados via API, notadamente pela Portaria CNJ n. 316 de 2024, meio que viabilizará a utilização efetiva dessas técnicas.

Contudo, a opacidade atual do Codex impede a análise pormenorizada de aspectos fundamentais da plataforma. Em síntese: **a)** não há transparência e auditabilidade púbica sobre o seu desenvolvimento, com pouquíssima informação sobre sua evolução ao longo do tempo; **b)** não há informação sobre a arquitetura de seu algoritmo, ainda que em termos genéricos; **c)** não há um programa de segurança informacional pautado pela transparência e integração de diversos setores da sociedade, tais como universidades, órgãos de polícia judiciária, Ordem dos Advogados do Brasil, Ministério Público, entre outros; entre outras lacunas. Convém quem essas lacunas sejam supridas pelo CNJ ao longo da implementação do Codex, sobretudo para viabilizar a atuação dialógica de distintos setores sociais.

Apresentados todos esses elementos, no que diz respeito ao desenvolvimento dos modelos de processamento de linguagem natural, a regulação de seu desenvolvimento perpassa pela análise de elementos centrais do banco de dados que lhe servem de subsídio. Assim, os aspectos fundamentais da regulação do Codex se espelham, em grande medida, nos modelos de PLN aplicados à plataforma. São questões centrais as formas lícitas de acesso aos dados processuais *públicos*; a finalidade da análise dos dados; e o tratamento adequado dos dados conforme a legislação de proteção de dados em vigor no Brasil.

Não há dúvidas de que o processamento de linguagem natural é uma, senão a mais eficiente, das formas de lidar com a análise do volume massivo de informações processuais do Brasil, especialmente para a consolidação do sistema de precedentes. O desenvolvimento do processamento de linguagem natural, porém, demanda transparência, auditabilidade e implementação de princípios éticos na gestão de dados, o que inclui, invariavelmente, a observância dessas diretrizes pelo próprio repositório dos dados que lhe serve de subsídio.

A exemplo da Biblioteca de Babel, um repositório virtualmente infinito de informações deve possuir mecanismos de busca adequados para o resgate eficiente da informação. Essa biblioteca, por congregar todo o conhecimento existente, tem índice de revocação tendente ao infinito. Sem mecanismos que refinem a precisão da busca, a sua eficiência é irrelevante e a informação se degenera em ruído.

Referências

Basile, J. (n.d.). *Library of Babel*. https://libraryofbabel.info

Bolasco, S., et al. (2005). Understanding Text Mining: A Pragmatic Approach. In S. Sirmakessis (Ed.), *Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference (Studies in Fuzziness and Soft Computing)* (pp. 31-50). Springer.

Borges, J. L. (1999). A biblioteca de Babel. In *O jardim de veredas que se bifurcam*. Editora Globo S.A.

Brasil. (1988). *Constituição da República Federativa do Brasil, de 5 de outubro de 1988*. http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm

Brasil. (2002). *Lei nº 10.406, de 10 de janeiro de 2002*. Institui o Código Civil. http://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm

Brasil. (2015). *Lei nº 13.105, de 16 de março de 2015*. Código de Processo Civil. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm

Brasil. (2018). *Lei nº 13.709, de 14 de agosto de 2018*. Lei Geral de Proteção de Dados Pessoais (LGPD). http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm

Brasil. Superior Tribunal de Justiça. (2020, 19 de novembro). *Comunicado da Presidência do STJ*. https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/19112020-Comunicad o-da-Presidencia-do-STJ.aspx

Câmara Júnior, A. T. da. (2013). *Processamento de linguagem natural para indexação automática semântico-ontológica* [Tese de Doutorado, Universidade de Brasília].

Brasil. Conselho Nacional de Justiça. (2007). *Resolução nº 46, de 18 de dezembro de 2007*. Cria as Tabelas Processuais Unificadas do Poder Judiciário e dá outras providências. https://atos.cnj.jus.br/atos/detalhar/167

Brasil. Conselho Nacional de Justiça. (2008). *Resolução nº 65, de 16 de dezembro de 2008*. Dispõe sobre a uniformização do número dos processos nos órgãos do Poder Judiciário e dá outras providências. https://atos.cnj.jus.br/atos/detalhar/119

Brasil. Conselho Nacional de Justiça. (2010). *Resolução nº 121, de 5 de outubro de 2010*. Dispõe sobre a divulgação de dados processuais eletrônicos na rede mundial de computadores, expedição de certidões judiciais e dá outras providências. https://atos.cnj.jus.br/atos/detalhar/92

Brasil. Conselho Nacional de Justiça. (2013). *Resolução nº 185, de 18 de dezembro de 2013*. Institui o Sistema Processo Judicial Eletrônico - PJe como sistema de processamento de informações e prática de atos processuais e estabelece os parâmetros para sua implementação e funcionamento. https://atos.cnj.jus.br/atos/detalhar/1933

Brasil. Conselho Nacional de Justiça. (2018). *Inteligência artificial: parceria com tribunal de Rondônia aproxima o futuro*. https://www.cnj.jus.br/inteligencia-artificial-parceria-com-tribunal-de-rondonia-aproxima-o-futuro/

Brasil. Conselho Nacional de Justiça. (2020a). *Portaria nº 242, de 10 de novembro de 2020*. Institui o Comitê de Segurança Cibernética do Poder Judiciário. https://atos.cnj.jus.br/atos/detalhar/3566

Brasil. Conselho Nacional de Justiça. (2020b). *Portaria nº 290, de 17 de dezembro de 2020*. Institui o Protocolo de Gerenciamento de Crises Cibernéticas no âmbito do Poder Judiciário (PGCC/ PJ). https://atos.cnj.jus.br/atos/detalhar/3640

Brasil. Conselho Nacional de Justiça. (2020c). *Portaria nº 291, de 17 de dezembro de 2020*. Institui o Protocolo de Investigação para Ilícitos Cibernéticos no âmbito do Poder Judiciário. https://atos.cnj.jus.br/atos/detalhar/3641

Brasil. Conselho Nacional de Justiça. (2020d). *Portaria nº 292, de 17 de dezembro de 2020*. Determina a adoção de Protocolo de Prevenção a Incidentes Cibernéticos no âmbito do Poder Judiciário (PPICiber/PJ). https://atos.cnj.jus.br/atos/detalhar/3651

Brasil. Conselho Nacional de Justiça. (2020e). *Resolução nº 331, de 20 de agosto de 2020*. Institui a Base Nacional de Dados do Poder Judiciário – DataJud como fonte primária de dados do Sistema de Estatística do Poder Judiciário – SIESPJ para os tribunais indicados nos incisos II a VII do art. 92 da Constituição Federal. https://atos.cnj.jus.br/atos/detalhar/3428

Brasil. Conselho Nacional de Justiça. (2020f). *Resolução nº 335, de 29 de setembro de 2020*. Institui política pública para a governança e a gestão de processo judicial eletrônico. Integra os tribunais do país com a criação da Plataforma Digital do Poder Judiciário Brasileiro – PDPJ-Br. Mantém o sistema PJe como sistema de Processo Eletrônico prioritário do Conselho Nacional de Justiça. https://atos.cnj.jus.br/atos/detalhar/3496

Brasil. Conselho Nacional de Justiça. (2021a). *Portaria nº 118, de 13 de abril de 2021.* Dispõe sobre o portfólio de soluções de tecnologia da informação e comunicação e serviços digitais do Conselho Nacional de Justiça. https://atos.cnj.jus.br/atos/detalhar/3866

Brasil. Conselho Nacional de Justiça. (2021b). *Portaria nº 162, de 10 de junho de 2021*. Aprova Protocolos e Manuais criados pela Resolução CNJ nº 396/2021, que instituiu a Estratégia Nacional de Segurança Cibernética do Poder Judiciário (ENSEC-PJ). https://atos.cnj.jus.br/atos/detalhar/3982

Brasil. Conselho Nacional de Justiça. (2021c). *Resolução nº 396, de 7 de junho de 2021*. Institui a Estratégia Nacional de Segurança Cibernética do Poder Judiciário (ENSEC-PJ). https://atos.cnj.jus.br/atos/detalhar/3975

Brasil. Conselho Nacional de Justiça. (2022a). *Justice 4.0 Program*. https://www.cnj.jus.br/wp-content/uploads/2022/05/justice-4-0-program.pdf

Brasil. Conselho Nacional de Justiça. (2022b). *Portaria nº 170, de 20 de maio de 2022*. https://atos.cnj.jus.br/atos/detalhar/4549

Brasil. Conselho Nacional de Justiça. (2022c). *Portaria nº 183, de 2 de junho de 2022*. Fixa prazo para a integração dos sistemas judiciais eletrônicos em funcionamento nos Tribunais à Plataforma Codex. https://atos.cnj.jus.br/atos/detalhar/4570

Brasil. Conselho Nacional de Justiça. (2022d). *Relatório de transição SG v. 5: 2022 09 05*. https://bibliotecadigital.cnj.jus.br/jspui/bitstream/123456789/768/1/relatorio-de-transicao-sg-v5-2022-09-05.pdf

Brasil. Conselho Nacional de Justiça. (2022e). *Resolução nº 446, de 14 de março de 2022*. Institui a plataforma Codex como ferramenta oficial de extração de dados estruturados e não estruturados dos processos judiciais eletrônicos em tramitação no Poder Judiciário Nacional e dá outras providências. https://atos.cnj.jus.br/atos/detalhar/4417

Brasil. Conselho Nacional de Justiça. (2023a). *Portaria nº 82, de 31 de março de 2023*. Institui o regulamento do Prêmio CNJ de Qualidade, ano 2023. https://atos.cnj.jus.br/atos/detalhar/5019

Brasil. Conselho Nacional de Justiça. (2023b). *Portaria nº 353, de 4 de dezembro de 2023*. Institui o Regulamento do Prêmio CNJ de Qualidade, ano 2024. https://atos.cnj.jus.br/atos/detalhar/5366

Brasil. Conselho Nacional de Justiça. (2024a). *Resolução nº 574, de 26 de agosto de 2024*. Dispõe sobre o acesso a dados judiciais públicos consolidados pelo Conselho Nacional de Justiça, prevê a possibilidade de depósito de serviços privados na PDPJ-Br e institui o portal unificado para usuários internos. https://atos.cnj.jus.br/atos/detalhar/5708

Brasil. Conselho Nacional de Justiça. (2024b). Portaria nº 316, de 20 de setembro de 2024. Regulamenta o acesso a dados judiciais públicos consolidados pelo Conselho Nacional de Justiça. https://atos.cnj.jus.br/atos/detalhar/5792

Brasil. Conselho Nacional de Justiça. (2024c). *Portaria nº 411, de 2 de dezembro de 2024*. Institui o Regulamento do Prêmio CNJ de Qualidade, ano 2025. https://atos.cnj.jus.br/atos/detalhar/5880

Brasil. Conselho Nacional de Justiça. (2024d). *Recomendação nº 154, de 13 de agosto de 2024*. Recomenda a todos os tribunais do país a adoção de modelo padronizado de elaboração de ementas (ementa-padrão). https://atos.cnj.jus.br/atos/detalhar/5693

Brasil. Conselho Nacional de Justiça. (n.d.-a). *Painel de acompanhamento da Plataforma Codex*. https://paineisanalytics.cnj.jus.br/single/?appid=de543e38-1c79-412e-b7b2-51624b55349d&sheet=2c946c67-9efe-4e23-a74c-d1c4792fb0bb

Brasil. Conselho Nacional de Justiça. (n.d.-b). *Plataforma Codex*. https://www.cnj.jus.br/sistemas/plataforma-codex/

Datatilsynet. (2018). *Artificial intelligence and privacy*.

Filipova, I. A. (2024). Legal regulation of artificial intelligence: Experience of China. *Journal of Digital Technologies and Law*, *2*(1), pp.46–73.

Filtz, E., et al. (2018). Interlinking Legal Data. In *Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systems (SEMANTiCS 2018)* (pp. 1–4). CEUR-WS.org. https://www.researchgate.net/publication/344378847_Interlinking_Legal_Data

Fundação Getulio Vargas. (2024, outubro). *Inteligência artificial: tecnologia aplicada à gestão dos conflitos no âmbito do Poder Judiciário brasileiro*. FGV Justiça. https://justica.fgv.br/sites/default/files/2024-10/ia_2fase_0.pdf

Hobbs, J. R., & Riloff, E. (2010). Information extraction. In N. Indurkhya & F. J. Damerau (Eds.), *Natural Language Processing* (2nd ed., pp. 511–532). Chapman & Hall/CRC.

Jenkins, J. (2008). What can information technology do for law? *Harvard Journal of Law & Technology*, *21*(2), pp. 589-607.

Katz, D. M., et al. (2023, 23 de fevereiro). *Natural language processing in the legal domain* [Preprint]. arXiv. https://arxiv.org/abs/2302.12039

Lancaster, F. W. (2004). *Indexação e resumos: teoria e prática* (2ª ed.). Briquet de Lemos.

Migalhas. (2019, 31 de outubro). PJe ou e-Proc? Tribunais contestam resolução do CNJ sobre suspensão imediata de e-Proc. *Migalhas Quentes*. https://www.migalhas.com.br/quentes/314284/pje-ou-e-proc--tribunais-contestam-resolucao-do-cnj-sobre-suspensao-imediata-de-e-proc

Morais, G. R. de. (2021). Why natural language processing is indispensable to entrech Brazil's precedents system. In A. de O. Gomes, E. R. Guarido Filho, L. P. de M. Tavares, P. M. A. Correia, & T. de A. Guimarães (Eds.), *Encontro de Administração da Justiça: Anais do ENAJUS 2021*. IBEPES.

Morais, G. R. de. (2023). *Pressupostos e diretrizes para a aplicação do processamento de linguagem natural à recuperação de informações jurisprudenciais: uma aplicação da AI4SG*. Editora Dialética.

Nastase, V., Mihalcea, R., & Radev, D. R. (2015). A survey of graphs in natural language processing. *Natural Language Engineering*, *21*(5), pp. 665–698.

Nunes, D., & Marques, A. L. P. C. (2019). Algoritmo: o risco da decisão das máquinas. *Bonijuris*, *31*(659), pp. 44–58.

Palmer, D. D. (2010). Text preprocessing. In N. Indurkhya & F. J. Damerau (Eds.), *Natural Language Processing* (2nd ed., pp. 9–30). Chapman & Hall/CRC.

Puente, M., & Coalla, L. (2023). *Mineração de Dados Textuais: Conceitos e Técnicas*. Blucher.

Saltzer, J., & Schroeder, M. (1975). *The protection of information in computer systems*. University of Virginia. https://www.cs.virginia.edu/~evans/cs551/saltzer/

Savoy, J., & Gaussier, E. (2010). Information retrieval. In N. Indurkhya & F. J. Damerau (Eds.), *Natural Language Processing* (2nd ed., pp. 455–484). Chapman & Hall/CRC.

Silva, N. C. da, et al. (2018). Document type classification for Brazil's supreme court using a Convolutional Neural Network. In *The Tenth International Conference on Forensic Computer Science and Cyber Law* (pp. 7–11). ICoFCS.

Tan, A.-H. (1999). Text mining: The state of the art and the challenges. In *Annals of Workshop on Knowledge discovery from Advanced Databases* (pp. 65–70).

Winkels, R. (2018). Automatic classification of civil law cases. In R. V. C. Fernandes & A. G. P. de Carvalho (Eds.), *Tecnologia Jurídica & Direito Digital: II Congresso Internacional de Direito, Governo e Tecnologia* (pp. 71–81). Fórum.

Yarowsky, D. (2010). Word sense disambiguation. In N. Indurkhya & F. J. Damerau (Eds.), *Natural Language Processing* (2nd ed., pp. 315–338). Chapman & Hall/CRC.

Notas

1. Informação disponível no Painel de Acompanhamento da Plataforma Codex (Brasil, s. d.-a).

2. Informação disponível na página do CNJ sobre o Codex (Brasil, s. d.-b) em: https://www.cnj.jus.br/sistemas/plataforma-codex/. Acesso em 14 jun. 2025. Referência completa na lista de referências.

3. Informação disponível em: https://www.cnj.jus.br/inteligencia-artificial-parceria-com-tribunal-de-rondonia-aproxima-o-futuro/. Acesso em 14 jun. 2025. Referência completa na lista de referências.

4. Lê-se o seguinte no Programa de Justiça 4.0 do CNJ: "The Codex platform is also part of the AI ecosystem. It was created in 2022 to be the official framework for collecting and storing all data from electronic lawsuits in Brazil" (Brasil, 2022ª, p. 11).

5. A exceção consta do art. 1º da resolução, que obriga a implantação apenas nos tribunais indicados nos incisos II a VII do art. 92 da Constituição Federal. A exceção ao CNJ é compreensível, pois os seus processos têm natureza eminentemente administrativa, não judicial.

6. Atualmente, a implantação do Codex é elemento avaliador do desempenho dos tribunais (Portaria CNJ n. 411/2024), sendo que esse elemento constava das avaliações desde 2022 (ver Portaria CNJ n. 170/2022; Portaria n. 82/2023; Portaria n. 353/2023). Ou seja, a integração do tribunal à plataforma é incentivada mediante o aumento do escore da avaliação de desempenho para fins estatísticos.

7. Corrobora esse fato a pesquisa da FGV (2024, p. 55), que utilizou dados fornecidos pelo CNJ. Referida pesquisa afirma textualmente que de fato se trata de aplicação JSON em todo o processo.

8. Sobre procedimentos ETL, Bolasco et al. (2005, p. 39) assim o explicam: *Extraction Transformation Loading are aimed at filing non-structured textual material into categories and structured fields. The search engines are usually associated with ETL that guarantee the retrieval of information, generally by systems foreseeing conceptual browsing and questioning in natural language.*

9. Tanto é assim que existe ferramenta específica do CNJ (JUMP-CNJ) para a mineração de dados do Codex; conforme o relatório de transição de gestão de 2022 (CNJ, 2022, p. 75).

10. A codificação da língua portuguesa está prevista na ISO-8859-1. Conforme Palmer (2010, p. 12), referida codificação é apenas *uma* forma de se tokenizar o português.

11. Trata-se de definir o assunto do documento, os motivos de ter sido incorporado ao acervo, e os pontos de interesse do usuário naquele documento (Lancaster, 2004, p. 9).

12. A tradução é a conversão de um documento para um conjunto de termos de indexação, o que pode ser feito por extração, que é a obtenção do termo de indexação através de termos que já constam do documento; ou por atribuição, na qual os termos não estão no documento (Lancaster, 2004, p. 18).

13. O vocabulário controlado é, basicamente, um dicionário de termos relevantes para a análise; pode dizer respeito tanto ao conteúdo material do texto, ou seja, as palavras que são relevantes para identificar do que trata o documento; quanto ao conteúdo formal do texto, ou seja, as palavras que compõem a sintática do texto. Nesse sentido, veja-se Lancaster (2004, p. 19) e Savoy e Gaussier (2010, p. 460).

14. A atribuição de pesos é uma entre várias técnicas que podem ser usadas para a superação do problema (semântico) de polissemia e da própria exaustividade da indexação, vez que, se uma única palavra com peso maior for capaz de indexar adequadamente o conteúdo do texto, elimina-se a necessidade de outras palavras no índice. Nesse sentido, veja-se Câmara Júnior (2013, p. 90); Savoy e Gaussier (2010, p. 458; que, inclusive, propõem a mescla entre fatores de peso qualitativo referentes à eficiência do termo enquanto indexador, e quantitativos, referentes ao número de ocorrência dos termos) e Lancaster (2004, p. 186). Há décadas a atribuição de pesos é vista como uma forma promissora de aprimoramento da classificação documental, inclusive com a utilização de indexadores automatizados por IA, tal qual já havia sido demonstrado por Jenkins em 2008 (p. 598).

15. Veja-se que a identificação das estruturas textuais compõe a etapa de pré-processamento (Nastase, Mihalcea, Radev, 2015, p. 670). Embora o mapeamento de estruturas textuais pareça trivial, o Projeto Victor, em sua primeira etapa, consistia na classificação das peças processuais por tipo de petição. Não só foi altíssimo o ganho de eficiência dos setores que faziam essa classificação, como também serviu de elemento básico para a estruturação das próximas etapas de seu desenvolvimento (Silva et al, 2018, p. 1).

16. Criação de um inventário de palavras associadas a um tema específico. O tema é, posteriormente, dividido em subtemas com palavras específicas dessa subclassificação. O número de subclassificações variará conforme o nível de granularidade desejado pelo indexador (Yarowsky, 2010, p. 318). Há quem, como Winkels (2018, p. 72), que considere esse método excessivamente simplista para a obtenção de bons resultados. Contudo, a simplicidade do método, por si só, não é fato determinante para subestimá-lo, dado que o nível de acurácia varia conforme a finalidade da indexação, aspecto já exposto neste artigo.
17. O tagueamento é a vinculação de um atributo temático ao texto por uma etiqueta de classificação. Geralmente, é uma técnica complementar a outras formas de indexação. É comum que haja tagueamento em tesauros, por exemplo.

18. Técnica que corresponde à criação de estruturas hierárquicas de organização das palavras, identificando-se o seu sentido do termo mais específico ao mais geral. Yarowsky (2010, p. 317) exemplifica com o exemplo da atribuição de sentido à palavra *crane* (que pode significar guindaste ou uma espécie de pássaro). Para essas palavras, no inglês, haveria uma hierarquia como guindaste (*crane*) > objeto inanimado > entidade física; e a hierarquia pássaro (*crane*) > ser vivo > entidade física. Ambos se assemelhariam na análise mais generalizada (entidade física), mas se diferenciariam em nível intermediário pelo fato de um ser vivo e o outro não.

19. A atribuição de pesos, assim como o tagueamento, geralmente é uma técnica complementar. Conforme o contexto em que a palavra se insere, atribui-se um peso maior ao seu significado mais provável. Pode-se usar essa técnica, inclusive, como um atributo vinculado às *etiquetas* da técnica de tagueamento.

20. Inicialmente, regulado pela Portaria n. 292 de 17 de dezembro de 2020. Esse protocolo foi revogado pela portaria n. 162 de 10 de junho de 2021, que aprovou os protocolos da Resolução n. 396 de 7 de junho de 2021.

21. Inicialmente, regulado pela Portaria n. 290 de 17 de dezembro de 2020. Esse protocolo foi revogado pela portaria n. 162 de 10 de junho de 2021, que aprovou os protocolos da Resolução n. 396 de 7 de junho de 2021.

22. Inicialmente, regulado pela Portaria n. 291 de 17 de dezembro de 2020. Esse protocolo foi revogado pela portaria n. 162 de 10 de junho de 2021, que aprovou os protocolos da Resolução n. 396 de 7 de junho de 2021.

23. Instituída pela Resolução n. 396 de 7 de junho de 2021, que revogou os protocolos anteriores.