Iudicium Textum Dataset
Uma Base de Textos Jurídicos para NLP

A. Willian Sousa[1], **Marcos Didonet Del Fabro**[1]

[1]C3SL, Centro de Computação Científica e Software Livre
Depto. de Informática – Universidade Federal do Paraná (UFPR)
CEP: 81530-900 – Curitiba – PR – Brazil

{awsousa,marcos.ddf}@inf.ufpr.br

Abstract. *The automatic text processing of natural language, with the use of probabilistic models and neural networks allows the analysis and classification of large volumes of text, leading the professionals and institutions of legal area to work more efficiently. However, the Natural Language Processing for Portuguese lacks of textual resources to support the creation and training of language models, as more deep studies related. In this article, a dataset of legal texts of the Brazilian Federal Supreme Court is presented to provide such resources. This dataset contains legal documents created by Supreme Court in its integral composition, with the subjects of the forty thousand process discriminated and, information about component sections with their respective author identified, allowing that data be used for studies of text classification, topic modeling, textual composition and others.*

Resumo. *O processamento automático de texto em linguagem natural por meio de modelos probabilísticos e redes neurais, permite a análise e classificação de grandes volumes de texto, tornando áreas como o Direito e os profissionais e instituições que o operam mais eficientes. Contudo, a área Processamento de Linguagem Natural para a Língua Portuguesa carece de recursos textuais, que permitam o treinamento de modelos robustos e o aprofundamento de estudos voltados para ela. Neste artigo, uma base de textos jurídicos obtidos da consulta pública do Supremo Tribunal Federal brasileiro é apresentada como meio de fornecer tais recursos, contando com mais de 40 mil acórdãos, além de 48 mil votos e 39 mil relatórios identificados de acordo com o seu ministro redator, provendo recursos para estudos de classificação de textos, modelagem de tópicos, composição textual e outros.*

1. Introdução

A área de Processamento de Linguagem Natural (PLN) é uma vertente da computação voltada para o estudo e criação de métodos, procedimentos e mecanimos que permitam o entendimento e processamento automático de textos escritos em linguagem natural. Essa tarefa se mostra desafiadora, pois textos em linguagem natural não possuem uma sintaxe rígida, nem tão pouco formalismos construtivos pré-determinados, exceto quando estão englobados dentro de um contexto técnico. De acordo com as tarefas de processamento a serem executadas, são necessários diferentes níveis de conhecimento linguístico, com as dificuldades de implementação proporcionais ao nível de profundidade da análise efetuada [Medeiros 1999].

Por ser a linguagem escrita, um meio utilizado pela mais diversas áreas, para registro e documentação de informações, é de alto grau a sua importância e relevância, em áreas do conhecimento como o Direito, onde o uso da linguagem técnica escrita se funde ao uso da linguagem comum na confecção de peças processuais, por meio de citações e inserções de trechos de documentos utilizados como provas ou recursos explicativos. Assim, não surpreende que o uso de PLN nesta área seja considerado um caminho natural e uma necessidade premente, considerando o grande volume de dados produzido, bem como a necessidade de manter esses dados disponíveis e consultáveis ao longo do tempo, sem causar prejuízos ao andamento processual que possui tempo e prazos definidos.

Ainda que a PLN possa contribuir significativamente para o manuseio e tratamento de informação textual, especialmente na área jurídica e que, nos últimos anos diversos métodos surgiram, facilitando o aprofundamento e aumento dos tipos e quantidades de tarefas automatizadas envolvendo texto, ainda é limitada a quantidade de recursos abertos disponibilizados livremente para a língua portuguesa. Especialmente, no tocante à dados reais que permitam treinamento de modelo probabilísticos e aqueles baseados em redes neurais. Essa escassez se dá pela dificuldade de obtenção destes dados, dos tratamentos necessários para eliminação de ruídos e a complexidade e esforço necessários para torná-los facilmente manuseáveis, sem perda de representatividade da informação original.

Com o intuito de reduzir a escassez de recursos textuais abertos para a nossa língua materna, foi construída a base apresentada neste artigo, contendo mais de 50 mil documentos jurídicos, produzidos no intervalos de 09 anos. Mais de 22 gigabytes de arquivos em formato PDF foram recuperados e processados na geração dos dados para compor a base. Para isso foram utilizados documentos que estivessem publicamente disponíveis e acessíveis, permitindo a evolução e expansão da área, fomentando a criação de modelos textuais específicos da área jurídica, tendo em vista a importância da justiça para uma sociedade democrática.

O artigo está organizado da seguinte forma. A Seção 2 descreve os documentos que compõem a base, sua finalidade, estrutura e composição. A Base de Textos Jurídicos é apresentada na Seção 3, com detalhamento de sua criação e do resultado final obtido. A Seção 4 descreve alguns desafios e limitações e, por fim, a Seção 5 exibe a conclusão do artigo.

2. Acordão

Acórdãos são documentos que resultam do julgamento por instâncias superiores do Judiciário Brasileiro, sendo um documento de estrutura rígida e bem definida. Entretanto, cada Tribunal, pela ausência de uma legislação que defina a estrutura e o conteúdo de um acórdão, ressalvado na sua independência funcional e na proposição de seu regimento interno, define a composição destes documentos, que no caso do STF, estão estruturados em seções.

A primeira seção do acórdão do STF contém uma descrição do tipo de processo a ser julgado, o nome do ministro relator do processo e as partes envolvidas. A segunda parte apresenta de modo sintético e resumido as matérias às quais o processo está relacionado, os fundamentos da decisão e uma breve descrição do próprio processo. A terceira parte, apresenta o texto do acordão propriamente dito - o resultado da votação. A próxima seção apresenta o relatório emitido pelo relator do processo, trazendo os fatos e as cir-

cunstâncias do caso julgado. A próxima seção engloba os votos dos ministros, com o primeiro voto sendo do relator e os demais ordenados de acordo com o tempo de atividade dos ministros no Tribunal, do mais novo ao mais antigo e o voto do presidente vem por último, encerrando a seção de votos.

Há ainda uma última seção, o Extrato de Ata, que repete algumas das seções como as partes, a decisão do acórdão e a indicação daqueles que estiveram presentes e ausentes ao julgamento.

3. A ITD - Iudicium Textum Dataset

Esta seção aborda a base de textos jurídicos, disponível para download no link: `http://dadosabertos.c3sl.ufpr.br/acordaos` e explicita a forma como a mesma foi concebida, detalhando os procedimentos necesários para sua criação, bem como as ferramentas utilizadas. Apresentando ainda, o resultado produzido e informações de acesso e obtenção da base.

A ITD, de acordo com a sua concepção inicial, deve englobar uma variedade de documentos que permita sua utilização em todas as esferas jurídicas e aplicação nas mais diversas tarefas aplicadas sobre distintos tipos de documentos. Para este momento de concepção da base, foram escolhidos apenas os acórdãos do Supremo Tribunal Federal (STF) publicados entre os anos de 2010 a 2018. Algumas avaliações foram feitas com documentos de anos anteriores a esse período, porém a quantidade de documentos digitalizados com baixa qualidade, fez com que optassêmos apenas por aqueles que permitissem recuperar informação de maneira precisa.

3.1. Etapas de Criação da Base

Aqui descrevemos cada uma das etapas necessárias para a criação da base, desde a captação dos dados brutos até o resultado final.

3.1.1. Recuperação dos Documentos

Os acórdãos do STF, dada a sua importância e natureza pública são disponibilizados para a sociedade através de uma página de internet que permite a consulta à jurisprudência do tribunal através do endereço `http://www.stf.jus.br/portal/jurisprudencia/`. Nesta página é possível definir filtros para a pesquisa, entre eles a data na qual foram publicados, o ministro relator, o tipo de documento e outros. Assim, a página de pesquisa gera uma requisição HTTP GET e recebe como resultado uma lista paginada de todos os documentos, cada um deles em separado, com informações básicas do processo e links para recuperar um arquivo no formato PDF contendo a integra da decisão. Esta lista, a depender do tamanho do resultado da pesquisa, torna muito lento a avaliação de cada um dos processos por um ser humano, sendo necessário um meio automático de recuperá-los.

Para recuperar todas as integras dos acórdãos de 2010 a 2018, desenvolvemos um *crawler* na linguagem Python[1], utilizando as bibliotecas *BeautifulSoup*[2], *Lxml*[3] e *Re-*

[1]https://www.python.org
[2]https://pypi.org/project/beautifulsoup4/
[3]https://lxml.de

quests[4], executando um conjunto de requisições HTTP para o serviço de consulta do STF, recebendo o resultado dessas requisições, isolando apenas as informações de interesse, executando novas requisições HTTP, recuperando os arquivos das integras e gravando-os localmente, totalizando 50.928 acórdãos. Os arquivos originais e aqueles deles derivados, foram nomeados conforme o número do acórdão a que se referem, obedecendo uma notação específica. Sobre estes documentos, utilizando a biblioteca Apache PDFBox[5] foi feita a extração de forma integral de seus textos no formatos de texto puro e Hyper Text Markup Language (HTML), sem divisão das seções componentes.

3.1.2. Separação da Seções

Inicialmente os acórdãos foram separados em partes distintas, a saber, dados do acórdão, relatório, votos e extrato da ata. Os relatórios ao serem desmembrados do documento original, produziram um único arquivo por acórdão, já os votos, produziram de um a vários arquivos, pois o voto de cada ministro foi gravado em um arquivo em separado. Já a parte referente aos dados do acórdão, gerou um único arquivo englobando as seções de Partes, Ementa e Acórdão, sendo este último a descrição do acórdão propriamente dito.

O processo de separação das seções executado sobre os arquivos PDF, foi possível graças a análise da estrutura de alguns exemplares de acórdãos por meio da qual se percebeu que cada um dos exemplares possuia um conjunto de marcadores que apontam para a página inicial de cada seção. Essa informação foi validada em todos os documentos baixados e grande parte apresentava a mesma estrutura. Exceto, documentos muito antigos - anteriores a 2010 - e aqueles compostos apenas por imagens obtidas da digitalização de documentos físicos.

A separação das seções e armazenamento das mesmas em arquivos, exigiu o desenvolvimento de um programa em Java, utilizando a biblioteca Apache PDFBox, para ler cada uma das seções e gerar as chamadas da própria biblioteca, necessárias para extração dos textos em formato HTML e seu consequente armazenamento. Cada arquivo gerado foi nomeado de forma que permita identificar o acórdão ao qual pertence, a seção a que se refere, o intervalo de páginas correspondentes no documento original e a sua posição relativa às demais seções. Para seções como Votos e Relatório, é incluído o nome do redator do texto daquela seção. Por exemplo, o arquivo `1_Voto_MIN_ROSA_WEBER_pg_6_13_seq_3.html` refere-se à um dos votos do acórdão número 01, emitido pela Ministra Rosa Weber e está localizado entre as páginas 06 e 13 no documento original, sendo a 3ª parte do documento. Mais formalmente, a nomenclatura dos arquivos obedece a seguinte notação:

```
<NUM_ACORDAO>_<SECAO>_pg_<PGINI>_<PGFIM>_seq_<SEQ>.<EXT>
<NUM_ACORDAO>::=<NUMERO>
<SECAO>::=<IDSECAO>|<ID_MINISTRO>
<IDSECAO>::=Ementa_e_Acordao|Relatorio_<ETIQUETA_MINISTRO>|
            Voto_<ETIQUETA_MINISTRO>
<ETIQUETA_MINISTRO>::=<TIPOID>|<TIPOID>_<MINISTRO>
<TIPOID>::=MINISTRO_PRESIDENTE|VICE_PRESIDENTE|MIN
<MINISTRO>::=GILMAR_MENDES|MARCO_AURELIO|CARLOS_VELLOSO|
```

[4]https://2.python-requests.org

[5]https://pdfbox.apache.org

```
                CEZAR_PELUSO|ELLEN_GRACIE|ALEXANDRE_DE_MORAES|
                EDSON_FACHIN|ROSA_WEBER|CARMEN_LUCIA|LUIZ_FUX|
                TEORI_ZAVASCKI|DIAS_TOFFOLI|SEPULVEDA_PERTENCE|
                CELSO_DE_MELLO|ROBERTO_BARROSO|JOAQUIM_BARBOSA|
                AYRES_BRITTO|EROS_GRAU|RICARDO_LEWANDOWSKI
<PGINI>::=<PG>
<PGFIM>::=<PG>
<PG>::=<NUMERO>
<SEQ>::=<NUMERO>
EXTENSAO::=HTML|TXT
<NUMERO>::=<NUMERO><DIGITO>|<DIGITO>
<DIGITO>::=0|1|...|9
```

Essa nomenclatura permite a obtenção da integra do acordão por meio de suas partes, além de possibilitar a utilização das seções com identificação do redator em processos de aprendizagem, comparação, agrupamento e outras.

3.1.3. Pré-processamento de Texto

Concluído o processo de separação das seções dos acórdãos, seguiu-se uma limpeza de textos desnecessários, como dados de assinatura digital dos documentos, numeração de páginas e outros inseridos no cabeçalho e rodapé dos textos. Apesar dos dados terem sido extraídos em dois formatos distintos, o HTML foi escolhido para esta etapa, dada a facilidade de identificação e separação das estruturas sintáticas do texto por meio de suas tags. A figura 1 ilustra as seções da primeira parte do documento. Há ainda, outros textos como assinaturas, locais e datas, indicação final e inicial do relator, nome do ministro no voto e outros que foram avaliados e eliminados dos documentos por não serem considerados relevantes para os objetivos de construção da base. Entretanto, essas informações ainda podem ser recuperadas na própria base, pois mantivemos uma cópia do texto integral para cada documento.

Para cada elemento da seção das partes do processo, foi necessária uma separação entre a indicação de seu tipo e o nome da parte. Da ementa e do texto do acórdão, foram retirados apenas os textos identificadores da seção e dados como assinaturas e datas. A separação desses textos foi feita por meio de expressões regulares e análise da estrutura de marcação dos documentos HTML. Deve-se ressaltar que a maior parte dos procedimentos de limpeza e organização dos textos foi feita baseando-se no conteúdo da extração em HTML e a mesma não pode ser aplicada diretamente sobre os arquivos em texto puro.

Mesmo o acórdão sendo uma síntese do julgamento e da decisão, em algumas situações pode apresentar uma quantidade de páginas expressiva, com a seção de Votos ocupando a maior parte, por conta das justificativas dos ministros e debates que podem ocorrer em plenário e que são registrados no documento. Estas informações e outras que figurem na seção de votos, são extraídas e processadas, porém no arquivo gerado, os mesmos não tem uma identificação do redator, pelo fato daquele texto ser um registro de um debate técnico e não pertencer a um único ministro. Assim, essa informação se encontra disponível na ITD, porém sem identificação, sendo possível consultá-la em separado para outras atividades de pesquisa.

A seção de Extrato da Ata, documentos extras, textos documentais e os debates

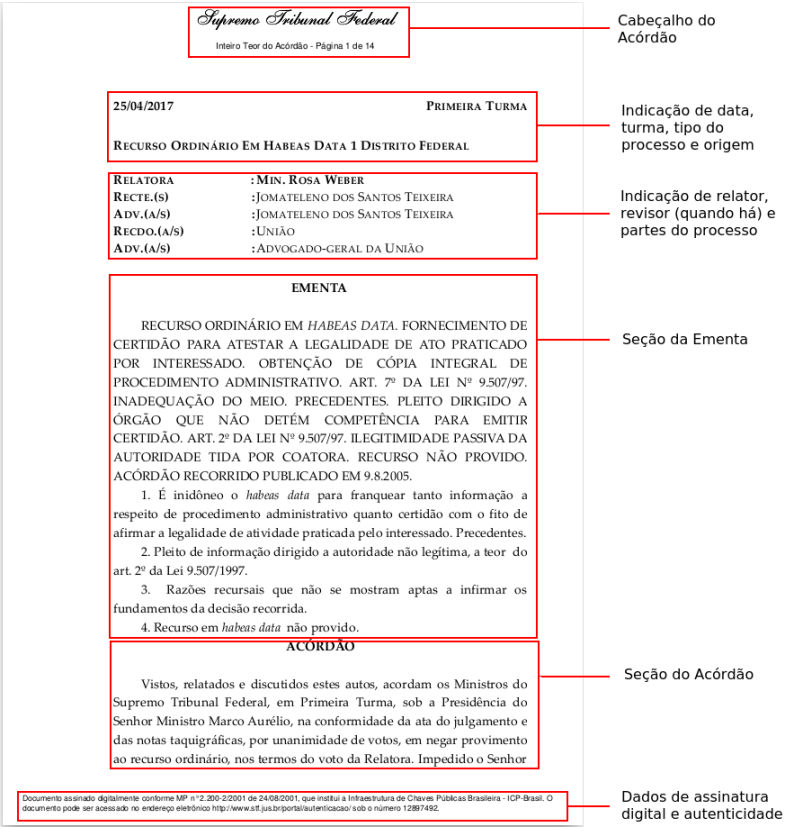

Figura 1. Partes do documento do acórdão

não sofreram nenhum tipo de pré-processamento, exceto remoção de cabeçalho e dados de autenticidade do documento.

3.1.4. Base de Documentos NoSQL

O JavaScript Object Notation (JSON)[6] é uma notação que permite compartilhamento e publicação de dados em diversas áreas e para os mais diversos propósitos. Sendo comum a sua utilização para interação remota entre aplicações de Internet, intercâmbio de dados científicos e para disponibilização de dados públicos abertos [Baazizi et al. 2019]. O formato JSON foi escolhido para armazenamento e estruturação da nossa base de textos jurídicos, por conta da sua flexibilidade, interoperabilidade, difusão e uso tanto em aplicações da indústria, quanto acadêmicas. A estrutura de documentos neste formato, permite expansão futura sem perda de compatibilidade com versões anteriores dos dados.

Para facilitar o armazenamento e manipulação, os dados em formato JSON foram inseridos em uma base de documentos MongoDB[7], um sistema gerenciador de documentos NoSQL que permite utilização sem o uso de um esquema rígido de descrição da estrutura dos seus documentos. Neste sistema os dados são armazenados em um formato chamado BSON (Binary JSON), uma representação binária da serialização de documentos JSON. A inserção dos dados textuais na base NoSQL, após as etapas de pré-

[6]https://www.json.org
[7]https://www.mongodb.com

processamento, foi feita por meio de um programa desenvolvido na linguagem Python, estando o mesmo disponível juntamente com a base.

Além de armazenar os documentos de forma acessível, também é necessário manter a representação do documento do acórdão em JSON o mais fiel possível ao seu original. Com esse intuito, a estrutura do documento foi concebida da seguinte forma:

```
{
  _id : <Identificação interna do MongoDB>,
  caminho : <Caminho do diretorio da base de texto>,
  arquivohtml : <Caminho do arquivo da integra em HTML>,
  arquivopdf : <Caminho do arquivo da integra em PDF>,
  arquivotxt : <Caminho do arquivo da integra em TXT>,
  arquivossecoes : {
    ATA : <Caminho do arquivo do Extrato da Ata>,
    EMENTA_ACORDAO : <Caminho da seção Ementa e Acórdão>,
    OUTROS : [ <Caminho dos outros arquivos do Acórdão> ],
    RELATORIO : <Caminho do arquivo do Relatório>,
    VOTOS : {
        <Nome arquivo do voto> : <Caminho do arquivo do Voto>
    }
  },
  numero : <Número do Acórdão>,
  partes : [
     {
       nome : <Nome da parte>,
       sigla : <Sigla da parte>,
       tipo : <Identificação de tipo da parte>
     }
  ],
  ementa : { texto : <Texto da Ementa>   },
  acordao : { texto : <Texto do acórdão> },
  extratoata : { texto : <Texto do Extrato da Ata> },
  formato : <Formato escolhido para pré-processamento>,
  integrahtml : <Integra em formato HTML>,
  integratxt : [  <Integra em TXT separada por linhas> ],
  relatorio : {
        relator : <Ministro relator>,
        texto : <Texto do relatório>
  },
  votos : [
    {
          texto : <Texto do voto>,
          votante : <Ministro votante>
        }
  ]
}
```

Após os devidos tratamentos dos dados textuais, a base de documentos conta com 41353 documentos, aproximadamente 70% da quantidade de arquivos recuperados inicialmente. Essa diferença se deve ao fato de alguns desses documentos, apesar de estarem em formato PDF, não permitirem a extração de seus textos sem um prévio processamento de reconhecimento ótico de caracteres. Como nosso interesse era a obtenção de textos de forma precisa, esses documentos foram ignorados durante os processos de conversão.

3.2. Estatísticas da Base

Apresentamos aqui, algumas informações sobre os quantitivos da ITD considerando apenas as informações armazenadas no MongoDB.

A base conta com um total de 41.353 documentos, com igual quantidade de relatórios e 56.250 votos. A maioria dos acórdãos possui apenas um texto de voto, como mostra a figura 2, na qual são apresentados dados dos acórdãos pela sua quantidade de votos. Já nas figuras 3 e 4 são apresentados, respectivamente os quantitativos de votos e relatórios dos últimos 12 ministros em atividade no STF. Essas quantidades não cobrem todos os votos da base, nem tão pouco os relatórios.

As quantidades de votos e relatórios emitidos pelos ministros dependem das funções que estes acumulam dentro do tribunal e as incumbências que lhes são determinadas pelo regimento interno do órgão. Assim, quando na condição de presidente de turma ou do próprio STF, um ministro poderá ter processos que são exclusivos de sua competência. Os votos e relatórios emitidos sob essas condições estão identificados na base com a etiqueta `MINISTRO_PRESIDENTE`. A identificação dos redatores destes votos e relatórios não foi executada, pois os mesmos totalizam apenas 1180 votos e relatórios, o que representa apenas 2,85% do total de relatórios e 2,09% do total de votos. Entretanto, caso os mesmos sejam necessários para alguma atividade, há como identificá-los com um trabalho de extração de informação textual, pois no corpo do texto do voto e do relatório há uma indicação do nome do ministro redator.

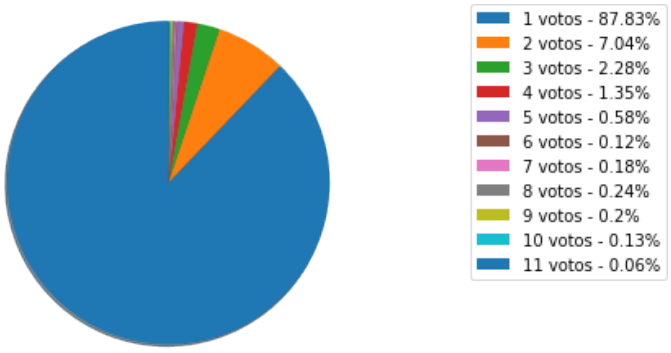

Figura 2. Acórdãos por Número de Votos

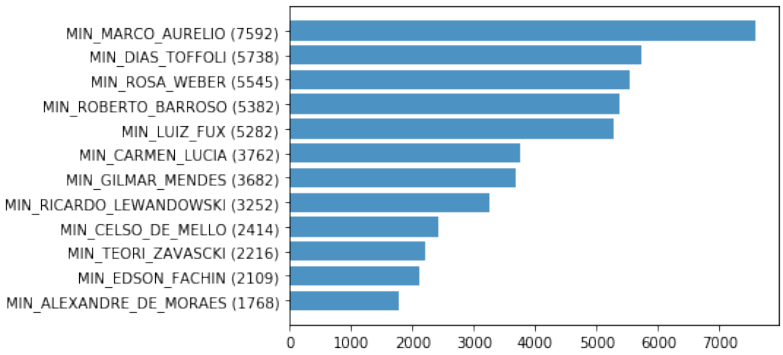

Figura 3. Votos por Ministro

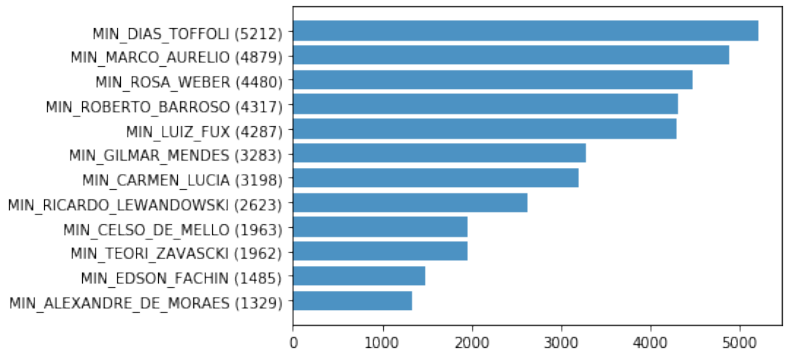

Figura 4. Relatórios por Ministro

3.3. Disponibilização da Base

A ITD pode ser obtida através do endereço `http://dadosabertos.c3sl.ufpr.br/acordaos`, onde estão disponibilizados os arquivos originais das integras e das seções nos formatos PDF, HTML e TXT. Também estão disponíveis os dados em formato JSON e os mesmos podem ser utilizados para importação em qualquer gerenciador de documentos que ofereça suporte ao formato, ainda que os mesmos tenham sido exportados do MongoDB, versão 4.0.9, a qual se recomenda o uso por questões de facilitação do uso dos programas desenvolvidos.

Os arquivos estão disponibilizados conforme a estrutura de diretórios detalhada na tabela 1, com a ressalva que no diretório *BaseITD/ITD* do site há apenas um arquivo compactado que comporta a estrutura aqui descrita. Já no diretório *BaseITD/tools* estão todos os programas desenvolvidos para criação da base, além de exemplos de como utilizá-los e no diretório *BaseITD/json* estão os arquivos prontos para importação em bases NoSQL.

Estão disponíveis 03 arquivos em formato JSON, onde o primeiro contém toda a base de acórdãos, o segundo apenas os votos e o terceiro apenas os relatórios. As informações destes dois últimos arquivos podem ser obtidas a partir da base completa, tendo sido separados apenas para facilitar o uso em tarefas de PLN, como treinamento de modelos de representação vetorial de palavras e classificação de textos, nos quais estamos trabalhando e que serão futuramente disponibilizados de forma aberta.

Diretório	**Conteúdo**
`BaseITD/ITD`	Diretórios nomeados de acordo com os números dos acórdãos
`BaseITD/ITD/<NNN>/html`	Arquivos html extraídos do acórdão
`BaseITD/ITD/<NNN>/txt`	Arquivos txt extraídos do acórdão
`BaseITD/tools`	Scripts e programas para extração e geração da base, além de notebooks de uso da base
`BaseITD/json`	Arquivos json contendo documentos da base

Tabela 1. Estrutura de diretórios da ITD

4. Desafios e Limitações

Os principais desafios encontrados na criação da base estão relacionados à falta de padronização e aos meios e formatos como os documentos se encontram disponíveis, exigindo o desenvolvimento de ferramentas específicas e obrigando a uma validação e análise contínuas dos dados obtidos, como meio de assegurar a sua qualidade.

Uma vez que formatos como PDF não são pensados para estruturação de texto, mas sim para a sua apresentação, analisar a estrutura do texto formatado e definir uma maneira de, a partir dos marcadores de formatação, obter a estrutura desejada e permitir que a mesma pudesse ser reconstruída tomando-se como base os dados já extraídos e processados, foi certamente a barreira mais importante a ser transposta durante o desenvolvimento da base.

A definição de métodos e estruturas de dados que permitissem extrair os dados, possibilitando a reconstrução da estrutura inicial do documento, foi feita por meio da geração de um formato intermediário, no caso o HTML, que apesar de também ser uma linguagem de formatação, facilitou o processo de extração. Assim, a conversão de PDF para HTML foi responsável pela manutenção da macro-estrutura do documento e a extração a partir do HTML, pela geração dos dados componentes da base. Esta última parte, exigiu um trabalho de garantia da qualidade dos resultados através da análise contínua e, por vezes, obrigou a reorganização de toda a cadeia de procedimentos de conversão e extração dos documentos para garantir a qualidade desejada.

No tocante às limitações da base, o fato da mesma conter apenas um único tipo de documento - acordãos emitidos pelo STF - e não permitir a sua utilização para análise de acórdãos de outros tribunais sem adaptações, pode reduzir a velocidade da sua expansão. Entretanto, como a mesma foi pensada e estruturada para não aderir a uma definição rígida dos dados, poderá abarcar uma grande variedade de tipos de documentos e comportar novas informações. Aumentando as suas possibilidades de uso, que, atualmente, já não são limitadas apenas às áreas tecnológicas, pois o seu conjunto de textos pode subsidiar, por exemplo, pesquisas nas áreas de estudos jurídicos e da linguagem escrita e falada, cobrindo um espectro de estudo que vai desde a forma como os ministros escrevem e estruturam os seus votos e relatórios, até as motivações e valorações utilizadas para embasar as suas decisões.

5. Conclusão

Neste artigo apresentamos a *Iudicium Textum Dataset*, uma base de textos jurídicos em Língua Portuguesa composta por documentos dos acórdãos do Supremo Tribunal Federal, que até onde sabemos se trata da primeira base deste tipo, tratada e disponibilizada abertamente. Na literatura relacionada é possível encontrar trabalhos como [Braz et al. 2018], [Da Silva et al. 2018] e [de Araujo et al. 2018] nos quais a criação de bases de textos jurídicos é citada, porém apenas a última está disponível publicamente e é voltada especificamente para a tarefa de reconhecimento de entidades nomeadas, contendo uma pequena quantidade de documentos, ainda que os mesmos apresentem considerável variabilidade.

Esperamos que a publicação dos dados consolidados e em formato aberto auxilie na ampliação da área de PLN, fomente o desenvolvimento de novas aplicações, a criação de novas bases e a melhoria e expansão da própria ITD, visando uma justiça mais eficiente e, cada vez mais, acessível a todos.

Referências

Baazizi, M.-A., Colazzo, D., Ghelli, G., and Sartiani, C. (2019). Schemas and types for json data: From theory to practice. In *Proceedings of the 2019 International Conference on Management of Data*, pages 2060–2063. ACM.

Braz, F. A., da Silva, N. C., de Campos, T. E., Chaves, F. B. S., Ferreira, M. H. S., Inazawa, P. H., Coelho, V. H. D., Sukiennik, B. P., de Almeida, A. P. G. S., de Barros Vidal, F., Bezerra, D. A., Gusmao, D. B., Ziegler, G. G., Fernandes, R. V. C., Zumblick, R., and Peixoto, F. H. (2018). Document classification using a bi-lstm to unclog brazil's supreme court. *CoRR*, abs/1811.11569.

Da Silva, N. C., Braz, F., de Campos, T., Gusmao, D., Chaves, F., Mendes, D., Bezerra, D., Ziegler, G., Horinouchi, L., Ferreira, M., 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-ICoFCS*, pages 7–11.

de Araujo, P. H. L., de Campos, T. E., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). Lener-br: A dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer.

Medeiros, M. B. B. (1999). Tratamento automático de ambigüidades na recuperação da informação.