1.03.03 - Ciência da Computação / Metodologia e Técnicas da Computação

ANÁLISE DE DECISÕES JUDICIAIS DO STF UTILIZANDO APRENDIZAGEM DE MÁQUINA

Thalyssa de Almeida Monteiro[1], André Lage Freitas[2]
1. Estudante do Instituto de Computação da Universidade Federal de Alagoas
2. Professor do Instituto de Ciências Humanas, Comunicação e Artes da Universidade Federal de Alagoas

Resumo

A jurimetria pode ser definida como a área do direito na qual se aplicam modelos estatísticos com o objetivo de analisar documentos judiciais para extrair informações dos mesmos, geralmente relacionadas à jurisprudência de certos tribunais. As abordagens mais comuns de jurimetria focam em problemas muito específicos que não focam na comparação do conteúdo de decisões judiciais e nas suas comparações com o resultado das respectivas decisões.

O objetivo deste trabalho é responder às seguintes perguntas: *"o conteúdo das Ementas de Acórdãos do STF são semelhantes para as decisões providas e improvidas?" e "agrupando os Acórdãos do STF utilizando seus conteúdos, eles ficarão agrupados também de acordo com suas decisões?"*. A metodologia utiliza técnicas de aprendizado de máquina e de Processamento de Linguagem Natural. Como conjunto de dados, foram utilizados entre 14.000 e 25.000 decisões do Supremo Tribunal Federal (STF), dentre Ementas e Inteiro Teor.

Concluímos que o conteúdo das Ementas de Acórdãos do STF não são necessariamente similares em função do resultado da decisão judicial. Ou seja, houve Ementas cujos conteúdos são similares, porém suas decisões são diferentes.

Palavras-chave: inteligência artificial; direito; processamento de linguagem natural.

Apoio financeiro: Bolsa de Iniciação Científica PIBIC da Pró-Reitoria de Pesquisa e Pós-Graduação da UFAL.

Trabalho selecionado para a JNIC: Universidade Federal de Alagoas - UFAL

Introdução

A jurimetria [1] pode ser definida como a área do direito na qual se aplicam modelos estatísticos com o objetivo de analisar documentos jurídicos, como contratos e decisões judiciais. As informações extraídas desses documentos geralmente estão relacionadas à jurisprudência de determinados tribunais e podem contribuir para estudos que vão desde a construção de um programa capaz de responder parte da prova da OAB [2], até a construção de um modelo capaz de classificar documentos automaticamente utilizando técnicas de mineração de texto e NLP, como mostrado em [3].

Outro uso para esses dados de decisões judiciais envolve a coleta de informações que podem servir para, entre outras coisas, elaborar uma estratégia de defesa com maiores chances de sucesso para um determinado processo, beneficiando o trabalho de agentes do direito. Contudo, a análise de documentos demanda tempo e recursos de tal forma que se torna cada vez mais difícil para o profissional analisá-los em tempo hábil. Neste cenário, a aplicação de algoritmos em bases de dados pode ajudar na criação de modelos capazes de analisar grandes volumes de dados num tempo curto, extraindo as informações necessárias [4].

No estado da arte, as melhores técnicas fazem uso de algoritmos de processamento de linguagem natural (NLP) com aprendizagem de máquina, geralmente para a resolução de problemas muito específicos, o que dificulta a aplicação dessas técnicas em cenários mais genéricos, além do fato de que os tribunais não possuem um padrão compartilhado para organizar sua documentação. Em [5], os autores mencionam esta problemática, usando-a como motivação para o desenvolvimento da LexNLP, uma biblioteca python focada na transformação de documentos legais não-padronizados dos Estados Unidos em arquivos estruturados. No Brasil, existem iniciativas como o CourtsBR [6], um conjunto de pacotes desenvolvidos em R, que servem para o download e extração de informações para alguns tribunais do país.

O objetivo deste trabalho é desenvolver uma ferramenta capaz de classificar decisões judiciais automaticamente. Especificamente, através da identificação automática de classes (ou grupos) através de visualizações gráficas e algoritmos de agrupamento (clustering).

Metodologia

Foram utilizados dois conjuntos de dados, ambos contendo Acórdãos do Supremo Tribunal Federal (STF). Estes acórdãos contém um resumo, chamado de Ementa. Para investigar separadamente as similaridades de ambos, separamos a metodologia em dois cenários, de acordo com os conjuntos de dados utilizados.

Em ambos os cenários, há etapas comuns da metodologia que são:

1. **Pré-processamento**. Remoção de palavras vazias utilizando a lista fornecida pela biblioteca NLTK, retirada de acentuação, espaços duplicados e números no texto, por meio de expressões regulares e a padronização do texto para caixa baixa, com a função *lower()* nativa

do python.

2. **Extração de características.** A segunda etapa consiste em extrair as características dos textos contidos nas decisões e representá-las numericamente. Para isso, foi utilizada a representação numérica da frequência das palavras contidas nas decisões, normalizando-as para diminuir a importância de palavras que aparecem muitas vezes. Essa transformação foi realizada utilizando o algoritmo *Term Frequency-Inverse Document Frequency* (TF-IDF) .

As demais etapas da metodologia são específicas para cada cenário, conforme descrição nas próximas seções.

CENÁRIO 1: Ementas do STF

O conjunto de dados utilizado possui 25.456 ementas com suas respectivas decisões rotuladas entre "provido" e "improvido": <Ementa, decisão>, sendo seus registros publicados desde 07/1950 à 07/2020. Demais tipos de decisões foram descartadas.

Com o conjunto de dados já rotulado, partiu-se para a modelagem dos vetores de características utilizando TF-IDF, pois o conceito por trás do algoritmo garante que termos que aparecem pouco terão sua significância representada, o que é importante para textos técnicos como a base de dados atual.

Com a representação dos vetores de características criada, partiu-se então para a plotagem do gráfico. Para a visualização dos dados, foi utilizado o algoritmo t-SNE (Visualização de dados de altas dimensões com base na distribuição estatística t de Student), além da biblioteca seaborn. Os eixos x e y do gráfico gerado mostram os índices de redução dimensional alcançados pelo algoritmo.

CENÁRIO 2: Acórdãos do STF

A base utilizada para o cenário é composta de 14.000 acórdãos do STF. Estes acórdãos não estão rotulados, o que torna mais conveniente uma abordagem de aprendizado não-supervisionado, utilizando para isso o algoritmo *k-means clustering*, agrupamento de dados com base em “pontos centrais” identificados pelo algoritmo.

Com a base de dados já pré-processada e com a criação dos vetores de características, foi aplicado o algoritmo LSI (Indexação com foco na semântica), e gerada uma matriz na qual estão indexadas a representação textual e numérica de cada entrada da base.

O algoritmo k-means recebeu como entrada a matriz gerada junto com a parametrização para gerar 2 clusters, já que o objetivo atual está focado em dois grupos, que são as possibilidades de decisões judiciais consideradas neste estudo.

Resultados e Discussão

Cenário 1:

O objetivo deste cenário era responder a pergunta: *“o conteúdo das Ementas de Acórdãos do STF são semelhantes para as decisões providas e improvidas?".*

A Figura 1 mostra as Ementas separadas de acordo com a decisão de seus Acórdãos: provido ou improvido. Cada ponto no gráfico representa uma Ementa sendo que os pontos mais próximos entre si são mais semelhantes semanticamente, ou seja, de acordo com o conteúdo (texto) das Ementas. É possível reparar que, mesmo com decisões diferentes, Ementas de decisões providas e improvidas conservam certas semelhanças em sua estruturação textual, como pode ser percebido nas regiões onde há pequenos pontos azuis cercados de grandes volumes de pontos laranjas.

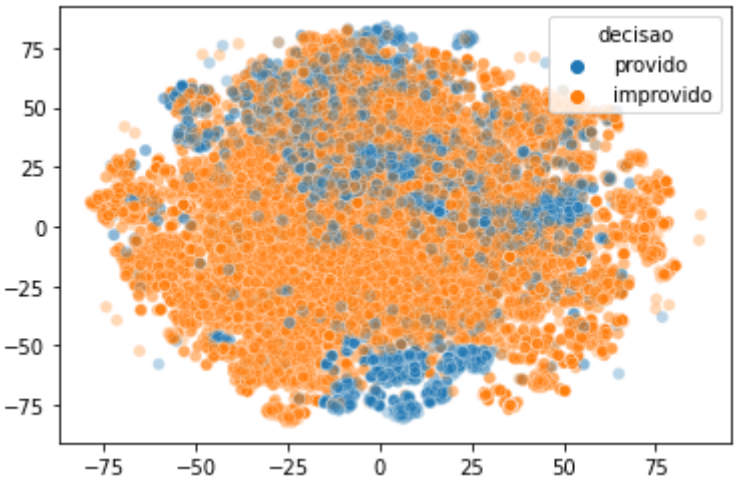

Figura 1: *Visualização de 25.456 Ementas do Supremo Tribunal Federal (STF) utilizando o algoritmo TD-IDF para representação numérica do conteúdo das Ementas e o algoritmo t-SNE para a redução para duas dimensões. Cada ponto representa uma Ementa de forma que quanto mais próximo os pontos, mais próximos são os conteúdos das respectivas Ementas.*

Porém, é possível notar também a necessidade de um balanceamento na base de dados, uma vez que há uma discrepância entre o volume de decisões rotuladas como providas e o de decisões rotuladas como improvidas. É importante também a validação deste modelo com conjuntos de dados de outros tribunais superiores, como bases de decisões do Superior Tribunal de Justiça (STJ), Superior Tribunal Militar (STM), Tribunal Superior do Trabalho (TST), Tribunal Superior Eleitoral (TSE).

Cenário 2:

O objetivo deste cenário era responder a pergunta: *"agrupando os Acórdãos do STF utilizando seus conteúdos, eles ficarão agrupados também de acordo com suas decisões?".*

O gráfico a seguir apresenta o resultado da clusterização (clustering) dos acórdãos do STF utilizando o algoritmo k-means. Após o pré-processamento e a modelagem por meio de TF-IDF, optou-se por utilizar o algoritmo Latent Semantic Index (LSI) para criar mais uma representação. O LSI foi escolhido por se tratar de textos maiores (os Acórdãos completos), já que este algoritmo se utiliza de álgebra linear e da estrutura semântica de um documento para resolver problemas como o dos sinônimos, onde duas ou mais palavras representam um mesmo significado, ou palavras cujo significado pode variar dependendo de sua posição na frase.

Em seguida, aplicou-se o algoritmo de agrupamento k-means, configurado com k=2, para agrupar os Acórdãos em dois grupos, conforme mostra a Figura 2. Foram escolhidos dois grupos para testar a hipótese de que a característica que separaria os Acórdãos em 2 grupos distintos seria por decisão, provido ou improvido. Considerando o desbalanceamento entre acórdãos de decisões providas e improvidas, era esperado que um dos clusters agrupasse mais itens que o outro, porém, os pontos mais dispersos no gráfico podem indicar que, ainda que pertençam ao mesmo grupo, nem todos os itens possuem grandes semelhanças entre si. Isso levanta a possibilidade destes itens mais distantes serem, na verdade, representações de acórdãos cujos rótulos não foram considerados no escopo deste trabalho, como por exemplo, "Prejudicada".

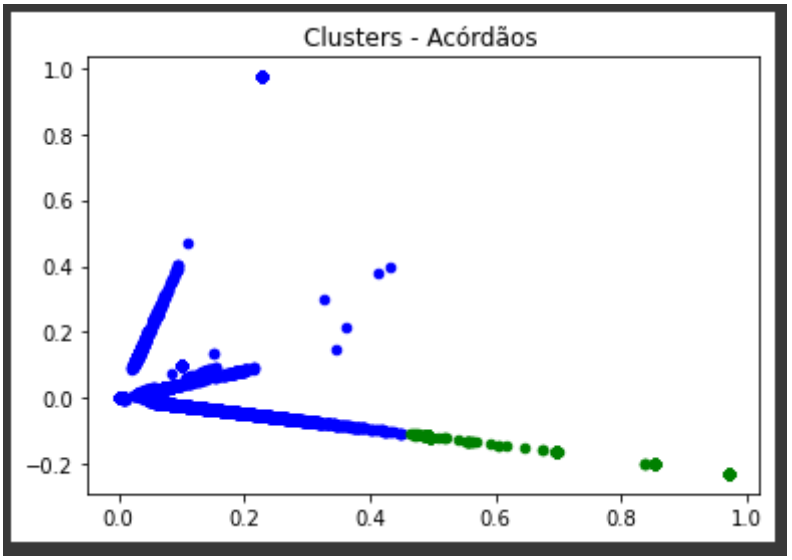

Figura 2: *Visualização de 14.000 Acórdãos do Supremo Tribunal Federal (STF) utilizando os algoritmos TD-IDF e LSI para representação numérica do conteúdo dos Acórdãos e o algoritmo k-means para agrupar os Acórdãos em 2 grupos.*

Conclusões

Este trabalho tem como objetivo desenvolver ferramentas capazes de auxiliar o operador do Direito na extração de informação através da classificação de decisões judiciais automaticamente. A metodologia inclui algoritmos de modelagem de Processamento de Linguagem Natural (TF-IDF e LSI), o algoritmo de redução de dimensão t-SNE e o algoritmo de aprendizado de máquina *k-means* que agrupa dados em função das suas semelhanças.

A primeira conclusão é que o conteúdo das Ementas de Acórdãos do Supremo Tribunal Federal (STF) não são necessariamente similares em função do resultado da decisão judicial. Ou seja, houve Ementas cujos conteúdos são similares, porém suas decisões são diferentes.

A segunda conclusão é que não foi possível utilizar a metodologia deste trabalho para saber se, agrupando os Acórdãos do STF utilizando seus conteúdos, eles ficariam agrupados também de acordo com suas decisões. A metodologia mostrou-se limitada, pois o conjunto de dados não estava rotulado. Dessa maneira, um trabalho futuro nesta linha de investigação consiste em rotular os acórdãos do conjunto de dados. Isso possibilitaria estudar as possíveis relações entre os textos das Acórdãos e os respectivos grupos encontrados pelo algoritmo de agrupamento. Além disso, rotulando-se os acórdãos do conjunto de dados, poderíamos investigar as características dos dois grupos de acórdãos a fim de facilitar a compreensão das características que melhor representam cada grupo, por exemplo, analisar de um verbo ou jargão estão mais

presentes em acórdãos cujas decisões são providas. Outra possibilidade inclui aperfeiçoamentos no próprio processo de clusterização, como testes com o k igual ao número de tipos de decisões possíveis dos Acórdãos, e testes com k igual a 2, porém descartando Acórdãos com decisões diferentes de provido e improvido.

Outro trabalho futuro é equilibrar as amostras dos conjuntos de dados em relação às suas decisões, por exemplo, de forma que a quantidade de decisões providas seja aproximada à quantidade de decisões improvidas, evitando um viés na construção do modelo de aprendizado de máquina. Além disso, a execução de testes com dimensão 3D e outros algoritmos de redução, como *Principal Component Analysis* (PCA), permitem a análise de outras possibilidades que podem levar a novas conclusões.

E outra hipótese para pesquisas futuras é que, separando conjuntos de dados de decisões por Ramo ou Área do Direito, consegue-se resultados mais precisos. Ademais, uma outra contribuição deste trabalho inclui o desenvolvimento de uma ferramenta capaz de permitir a visualização de documentos jurídicos relacionados a decisões judiciais de acordo com suas decisões. Essa ferramenta também permite agrupar (cluster) decisões judiciais com base no conteúdo das decisões. Por último, essa ferramenta pode ser utilizada gratuitamente e possui código-fonte aberto, o que permite melhorias e contribuições por demais membros da comunidade científica.

Referências bibliográficas

[1] Loevinger, Lee. 1963. Jurimetrics: The Methodology of Legal Inquiry. Law and Contemporary Problems 28 (1): 5. https://doi.org/10.2307/1190721

[2] DELFINO, Pedro; CUCONATO, Bruno; HAEUSLER, Edward Hermann; RADEMAKER, Alexandre. Passing the Brazilian OAB exam: Data preparation and some experiments. Frontiers in Artificial Intelligence and Applications, [S. l.], p. 89-94, 2017.

[3] BARROS, Rhuan et al. Case Law Analysis with Machine Learning in Brazilian Court. Recent Trends and Future Technology in Applied Intelligence, [s. l.], p. 857-868, 2018. E-book.

[4] Lage-Freitas A, Allende-Cid H, Santana O, Oliveira-Lage L. 2022. Predicting Brazilian Court Decisions. PeerJ Computer Science 8:e904 https://doi.org/10.7717/peerj-cs.904

[5] Bommarito, Michael James and Katz, Daniel Martin and Detterman, Eric, LexNLP: Natural Language Processing and Information Extraction For Legal and Regulatory Texts (June 6, 2018). Available at SSRN: https://ssrn.com/abstract=3192101 or http://dx.doi.org/10.2139/ssrn.3192101

[6] (Vários autores). CourtsBr: pacote de ferramentas para processamento de dados jurídicos. Disponível em <https://github.com/courtsbr>. Último acesso em: 11/04/2022.