Assessment of text clustering approaches for legal documents

Ingrid L. A. da Silva[1], **Rafael Ferreira Mello**[1], **Péricles B.C. Miranda**[1], **André C.A. Nascimento**[1], **Isabel W. S. Maldonado**[2], **José L. M. Coelho Filho**[2]

[1] Departamento de Computação, Universidade Federal Rural de Pernambuco (UFRPE), Recife, Brasil
[2] NESS Law, São Paulo, Brasil

Abstract. *The judicial system is composed of numerous court documents. These documents may contain relevant information that supports decision-making in future processes. However, collecting this information is not a trivial task. This article proposes the use of clustering techniques to group similar court lawsuits and facilitate the collection of information. In this way, different approaches were evaluated for identifying the most appropriate to perform this task. The approaches were applied to a database composed of 1515 facts of initial petitions. These approaches were evaluated using internal metrics and texts of the grouped court lawsuits. The results showed that the best approach to grouping court lawsuits is composed of the K-Means algorithm and the TF-IDF representation technique in combination with the PCA technique.*

Resumo. *O sistema judiciário é composto por inúmeros documentos relacionados a processos jurídicos. Esses documentos podem conter informações relevantes que suportem a tomada de decisão em processos futuros. No entanto, a coleta dessas informações não é uma tarefa trivial. Este artigo propõe o uso de agrupamento para reunir processos semelhantes e facilitar a coleta de informações. Dessa forma, diferentes abordagens foram avaliadas com a intenção de identificar a mais adequada para realizar esta tarefa. As abordagens foram aplicadas a uma base de dados composta por 1515 textos de fatos de petições iniciais. Essas abordagens foram avaliadas levando em consideração métricas de avaliação internas e os textos dos processos agrupados. Os resultados apontaram que a melhor abordagem para realizar o agrupamento de processos jurídicos é composta pelo algoritmo K-Means e pela técnica de representação TF-IDF em combinação com a técnica PCA.*

1. Introdução

O sistema judiciário gera uma grande quantidade de documentos relacionados a processos jurídicos que são abertos e julgados diariamente no mundo inteiro [Raghav et al. 2015]. Esses documentos são dados valiosos para todos os agentes envolvidos no meio jurídico, pois processos anteriores que são semelhantes a novos processos podem conter informações que oferecem suporte à tomada de decisão. No entanto, para ter acesso a essas informações, os profissionais precisam procurar os processos semelhantes em uma vasta coleção de documentos que estão em sua maioria no formato de linguagem natural (documentos textuais), o que dificulta a análise e interpretação desses resultados [Kachappilly and Wagh 2018].

Devido ao avanço tecnológico da internet, esses dados estão disponíveis para a consulta através de diversas plataformas. No entanto, não existe um padrão que pode ser

utilizado para identificar casos similares [Lv et al. 2018]. Uma das abordagens para realizar essa tarefa é através das ferramentas de busca, mas estas são limitadas em diversos aspectos. Para realizar uma busca o usuário precisa definir os termos que serão utilizados na consulta e as principais limitações estão relacionadas com a forma com que as consultas são realizadas, por exemplo, algumas dessas ferramentas limitam a busca desses termos ao resumo e as palavras chaves dos documentos [de Colla Furquim and De Lima 2012].

A coleta de informações que sejam relevantes para o andamento de um processo não é uma tarefa trivial e demanda dedicação e esforço dos profissionais da área jurídica. Dessa forma, a inteligência artificial vem sendo utilizada para auxiliar esses profissionais a manterem a produtividade e a eficiência na realização dessas tarefas [Rosca et al. 2020]. O Agrupamento é um técnica da inteligência artificial que busca dividir determinados objetos em grupos, levando em consideração medidas de similaridade. Então, objetos pertencentes a um mesmo grupo devem ser mais semelhantes entre si do que aqueles pertencentes a outros grupos [Raghuveer 2012]. Diversas abordagens de agrupamento estão sendo utilizadas em conjunto com técnicas do processamento de linguagem natural para realizar a tarefa de agrupamento de textos [Amine et al. 2010, Aggarwal and Zhai 2012].

A tarefa do agrupamento de textos consiste em agrupar conjuntamente documentos que sejam semelhantes entre si [Liu et al. 2003]. Para isso, inicialmente é preciso utilizar alguma técnica de representação de textos com o objetivo de transformar os textos dos documentos em atributos que possam ser analisados e aplicados a algoritmos de agrupamento. A abordagem mais utilizada é o modelo de vetor espacial, onde as palavras de cada documento são avaliadas levando em consideração diversos aspectos, como a sua frequência ou o contexto em que está inserida [Fan et al. 2010, Aggarwal and Zhai 2012].

No contexto jurídico, a petição inicial é um documento judicial extremamente relevante para os processos. Este documento é utilizado para a abertura do processo jurídico e contém informações a respeito das partes envolvidas, dos fatos do caso, do direito que foi utilizado como embasamento, do pedido realizado, entre outras. A seção de fatos da petição inicial consiste de um texto que descreve a situação, os lugares, as evidências, ou seja, a história que conduz ao caso [Polpinij et al. 2020]. De acordo com a teoria do realismo jurídico, a decisão final dos juízes leva em consideração a resposta ao estímulo dos fatos do caso, o que a torna uma seção importante de ser considerada na tarefa de apoio à tomada de decisão [Aletras et al. 2016].

Diante do que foi apresentado, o objetivo deste artigo é avaliar diferentes abordagens de agrupamento na realização de agrupamentos de processos jurídicos com a intenção de apoiar a tomada de decisão por parte dos advogados. Os experimentos foram realizados levando em consideração o texto da seção de fatos pertencentes a petições iniciais. Além disso, a melhor abordagem foi definida levando em consideração diferentes técnicas de representação de texto, diferentes algoritmos de agrupamento.

O restante deste trabalho está organizado da seguinte forma: A Seção 2 apresenta os trabalhos relacionados à problemática trabalhada neste artigo. A Seção 3 apresenta as perguntas de pesquisa que buscamos responder com os experimentos realizados. A Seção 4 descreve a metodologia experimental adotada. A Seção 5 apresenta os resultados obtidos com os experimentos. Por fim, a Seção 6 apresenta as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Na literatura é possível identificar diversos trabalhos que propõem o uso de agrupamento de texto no domínio jurídico para atingir diferentes objetivos. Em [de Colla Furquim and De Lima 2012], o agrupamento de documentos jurídicos foi utilizado com o objetivo de auxiliar na pesquisa jurisprudencial, onde o resultado da pesquisa são os documentos presentes no grupo ao qual um novo documento foi atribuído. O agrupamento é realizado através de um algoritmo de agrupamento *hard* semi-supervisionado que é aplicado a termos do domínio presentes nos documentos e as leis referenciadas nestes mesmos documentos. Algoritmos de agrupamento *hard* são aqueles que atribuem cada objeto a um único grupo. Já nos algoritmos de agrupamento *soft* os objetos podem ser atribuídos a mais de um grupo com diferentes níveis de pertencimento.

Em [Fan et al. 2010], [Raghav et al. 2015] e em [Kachappilly and Wagh 2018], o agrupamento é realizado levando em consideração a abordagem de que "dois objetos são similares se eles estão relacionados por objetos similares". Fan *et al.* [Fan et al. 2010] realiza o agrupamento de textos de leis levando em consideração as relações referenciais entre essas leis, enquanto que *Raghav et al.* [Raghav et al. 2015] e Kachappilly e Wagh [Kachappilly and Wagh 2018] realizam o agrupamento de julgamentos levando em consideração as citações desses documentos. Os três trabalhos utilizam técnicas de agrupamento com o objetivo de auxiliar os profissionais na tarefa de identificação de objetos similares.

[Raghuveer 2012] e [Conrad et al. 2005] utilizaram o agrupamento para auxiliar na organização, análise e recuperação de documentos. Em Raghuveer [Raghuveer 2012] foi realizado agrupamento de textos de julgamentos legais, onde cada grupo representa os documentos mais relevantes para um tópico obtido através da modelagem de tópicos realizada com a técnica de Alocação latente de Dirichlet. A semelhança entre os tópicos e os documentos foi avaliada através da similaridade do cosseno. Já em Conrad *et al.* [Conrad et al. 2005] foram utilizados algoritmos de agrupamento com diversas abordagens com o objetivo de agrupar documentos de um escritório de advocacia. As métricas avaliadas nesse trabalho apontam para a eficiência de agrupamentos com abordagens de agrupamento *soft* e hierárquico.

No geral, os trabalhos analisados utilizam as técnicas de agrupamento na identificação de objetos similares. No entanto, [Poudyal et al. 2019] se distancia um pouco dessa abordagem e faz uso das técnicas de agrupamento para identificar automaticamente argumentos em documentos de jurisprudência. Neste trabalho, as sentenças argumentativas são agrupadas em grupos de potenciais argumentos e como uma sentença pode fazer parte de mais de um argumento, foi utilizada uma abordagem de agrupamento *soft* através da utilização do algoritmo Fuzzy c-means.

Alguns trabalhos da literatura já exploraram o uso dos fatos relacionados a um processo no desenvolvimento de propostas que aplicam a inteligência artificial no domínio jurídico. O trabalho de [Aletras et al. 2016] avalia o uso de diferentes seções de textos de julgamentos na tarefa de predição dos resultados de processos judiciais. Um dos resultados desse trabalho indica que os fatos são o fator preditivo mais importante na realização dessa tarefa. [Chen et al. 2019], [Kowsrihawat et al. 2018] e [Thammaboosadee et al. 2012] também utilizaram os fatos dos casos para predizer resultados das sentenças dos julgamentos através de diferentes modelos baseados em *Deep*

Learning, Redes Neurais e Árvores de Decisão.

Apesar dos trabalhos anteriores utilizarem algoritmos de agrupamento em diferentes contextos jurídicos, eles não apresentam uma avaliação de diferentes abordagens para identificar a melhor combinação de técnicas de processamento de linguagem natural e agrupamentos no contexto jurídico.

3. Perguntas de Pesquisa

Diante do que foi apresentado na seção anterior, o objetivo deste trabalho é avaliar diferentes abordagens para a realização do agrupamento de processos jurídicos, utilizando os textos dos fatos de petições iniciais. Nesse sentido, esse trabalho se propõem a responder às seguintes perguntas de pesquisa:

PERGUNTA DE PESQUISA 1: *Qual a abordagem mais eficiente para realizar o agrupamento de processos jurídicos, utilizando a seção de fatos da petição inicial?*

PERGUNTA DE PESQUISA 2: *A abordagem de agrupamento proposta é eficiente o suficiente para ser utilizada como ferramenta de apoio a tomada de decisão em aplicações reais do mundo jurídico?*

4. Metodologia Experimental

4.1. Base de Dados

A base de dados utilizada neste trabalho consiste de um conjunto de petições iniciais de processos judiciais de tribunais brasileiro. A petição inicial é o documento que inicia o processo judicial e costuma apresentar uma estrutura semelhante à mostrada na Figura 1.

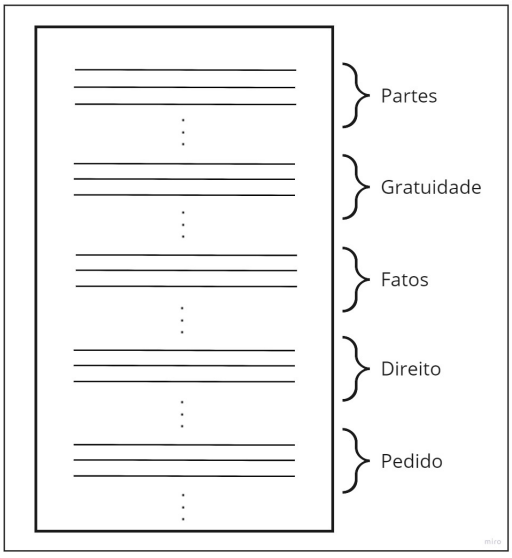


Figura 1. Estrutura de uma Petição Inicial

A seção 'Partes' apresenta os dados do autor e do réu envolvidos no processo. A 'Gratuidade' é uma seção opcional e é utilizada quando o autor não possui condições financeiras de arcar com as custas processuais e necessita solicitar justiça gratuita. Na seção de 'Fatos' deve-se relatar toda a história que levou a abertura do caso jurídico,

deve ter informações a respeito do que aconteceu, quando aconteceu e as consequências resultantes. A seção 'Direito' deve apresentar as regras jurídicas que foram violadas. Já a seção de 'Pedido' deve levar em consideração os fatos e fundamentos apresentados e deve conter o que se busca como solução para o caso judicial.

Neste trabalho, o agrupamento foi realizado levando em consideração o texto da seção de 'Fatos', pois a intenção foi agrupar processos semelhantes levando em consideração a situação envolvida no caso jurídico. Ao todo foram utilizados 1515 textos de fatos, que passaram pelas seguintes etapas de pré-processamento:

1. Remoção de pontuação;
2. Remoção de numerais;
3. Remoção de palavras com tamanho menor ou igual a 3;
4. Remoção de palavras frequentes que não apresentam significado relevante para o texto, chamadas de *stopwords*;
5. Aplicação do processo de lematização, onde cada termo é substituído por sua forma padrão, evitando que diferentes inflexões da palavra resultem em valores que sejam considerados distintamente no algoritmo.

4.2. Representação de Textos

Na tarefa de agrupamento de documentos, uma das etapas que deve ser realizada é a representação dos textos na forma de atributos que podem ser utilizados nos algoritmos de agrupamento. Neste trabalho, foram avaliadas as abordagens de TF-IDF [Amine et al. 2010] em conjunto com técnicas de extração e seleção de atributos e a técnica de *Word Embedding* através do algoritmo word2vec [Xiao et al. 2017]. Essas duas técnicas foram escolhidas para realizar a representação textual dos documentos por apresentarem duas estratégias de funcionamento distintas. Enquanto o TF-IDF leva em consideração a frequência das palavras no documento e no corpus, a técnica de *Word Embedding* captura informação do contexto semântico no qual a palavra está inserida.

Na técnica de TF-IDF, cada documento é representado através de um vetor numérico de palavras, onde o peso de cada palavra é calculado levando em consideração a frequência da palavra em cada documento e a frequência da palavra no corpus inteiro. Dessa forma, a medida atribui um peso maior para os termos que aparecem frequentemente em um documento e raramente no corpus completo [Amine et al. 2010]. Abaixo temos a fórmula do TF-IDF, onde *TF(t,d)* corresponde ao número de ocorrências de um termo *t* em um documento *d*, *N* é a quantidade de documentos presentes no corpus e *N(t)* é a quantidade de documentos que contém o termo *t*.

$$TF - IDF(t, d) = TF(t, d) \times \log(\frac{N}{N(t)}) \tag{1}$$

Na extração de atributos, um conjunto de novos atributos são obtidos através de transformações nos atributos originais e nesse trabalho a extração de atributos foi realizada utilizando a técnica de *Principal Component Analysis* (PCA) [Liu et al. 2003]. Já na seleção de atributos, um subconjunto dos atributos originais são selecionados levando em consideração algum critério. O critério utilizado neste trabalho foi a seleção dos 1000 atributos com maior valor para a medida TF-IDF.

O *Word embedding* é uma forma de representação que pode ser aprendida a partir de um corpus e essa representação permite que palavras com significados similares tenham representações similares. Dessa forma, é possível capturar informações do contexto em que a palavra está inserida [Xiao et al. 2017]. Para este trabalho foi utilizado o algoritmo word2vec em sua versão *skip-grams*, onde para cada palavra do documento, foi obtido um vetor com 100 dimensões composto por palavras associadas.

4.3. Algoritmos de Agrupamento

Neste trabalho foram selecionados três diferentes algoritmos para realizar o agrupamento dos documentos: *K-Means, Agglomerative clustering e Spectral clustering*. A seleção destes algoritmos levou em consideração o objetivo de analisar diferentes abordagens de agrupamento, com a intenção de identificar a mais adequada para realizar o agrupamento dos textos dos fatos de petições iniciais.

O algoritmo *K-Means* utiliza a abordagem de particionamento ou de agrupamento plano, onde cada documento é atribuído ao grupo em que a distância entre o vetor de atributos do documento e o centro do grupo é a menor [Kachappilly and Wagh 2018]. Em seguida, os centros dos grupos são atualizados e o mesmo processo de atribuição dos grupos é aplicado. Essas etapas são aplicadas repetidamente com o objetivo de minimizar a soma das distâncias quadradas dos grupos.

O *Agglomerative clustering* utiliza a abordagem hierárquica, que consiste em gerar uma árvore hierárquica de grupos. Neste algoritmo, cada documento inicia no seu próprio grupo e depois os grupos são sucessivamente fundidos uns com os outros de acordo com alguma medida de similaridade, até atingir a quantidade de grupos definida [Berkhin 2006].

Já o *Spectral clustering* utiliza a abordagem baseada em grafos, onde é criado um grafo de similaridade dos documentos, seguido pelo cálculo dos autovetores da matriz laplaciana [Von Luxburg 2007]. Esses autovetores contém informações que indicam como agrupar os nós do grafo. Por último, o algoritmo *k-means* é utilizado nesses vetores com o objetivo de obter os rótulos dos grupos de cada nó do grafo, ou seja, de cada documento.

4.4. Medidas de Avaliação

Os agrupamentos no contexto do domínio jurídico costumam ser avaliados através de métricas de avaliação internas e externas e através de especialista humano [Conrad et al. 2005]. As métricas de avaliação internas utilizam informações obtidas do próprio agrupamento e costumam avaliar o quão bem definido foi o agrupamento realizado. As métricas externas são utilizadas quando as classes verdadeiras dos objetos são conhecidas e utilizam informações de acertos realizados com o agrupamento. Já os especialistas humanos costumam avaliar a coerência e a utilidade entre os objetos dos grupos formados.

Para a avaliação deste trabalho foram utilizadas métricas internas como coeficiente de silhueta, o índice de Calinski-Harabasz e o índice de Davies-Bouldin [Wang and Xu 2019]. Essas métricas podem ser utilizadas sem o conhecimento de algum conjunto de classes verdadeiras, pois levam em consideração as distâncias dos elementos nos grupos, com o objetivo de avaliar o quanto os objetos de um mesmo grupo são similares entre si e dissimilares entre grupos diferentes.

O coeficiente de silhueta leva em consideração a distância entre os elementos de um mesmo grupo e a distância entre os elementos desse grupo com os elementos do grupo mais próximo. Essa métrica varia entre -1 e 1 e quanto maior o seu valor, mais separados e densos são os grupos. O índice de Calinski-Harabasz corresponde a razão entre a dispersão presente em um mesmo grupo e a dispersão presente entre os diferentes grupos. Para essa métrica, altos valores também indicam melhor separação dos grupos. Já o índice de Davies-Bouldin leva em consideração a similaridade média entre os grupos, que é calculada utilizando o tamanho dos grupos e a distância entre os seus centróides. Nessa métrica um menor valor indica uma melhor partição entre os grupos.

As abordagens de agrupamento propostas neste trabalho também foram avaliadas levando em consideração os textos de diferentes documentos e os grupos aos quais eles pertencem, utilizando uma abordagem semelhante à utilizada no trabalho [Raghav et al. 2015]. Dessa forma, foram selecionados aleatoriamente 30 pares de documentos, que foram lidos com o objetivo de atribuir uma pontuação de 0 a 10 para a similaridade dos documentos. Uma pontuação igual a 0 indica que não existe similaridade entre os documentos e uma pontuação igual a 10 indica uma alta similaridade. Em seguida, os pares foram avaliados como verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN). Por último, foi realizado o cálculo das métricas de precisão, cobertura e F1.

5. Resultados

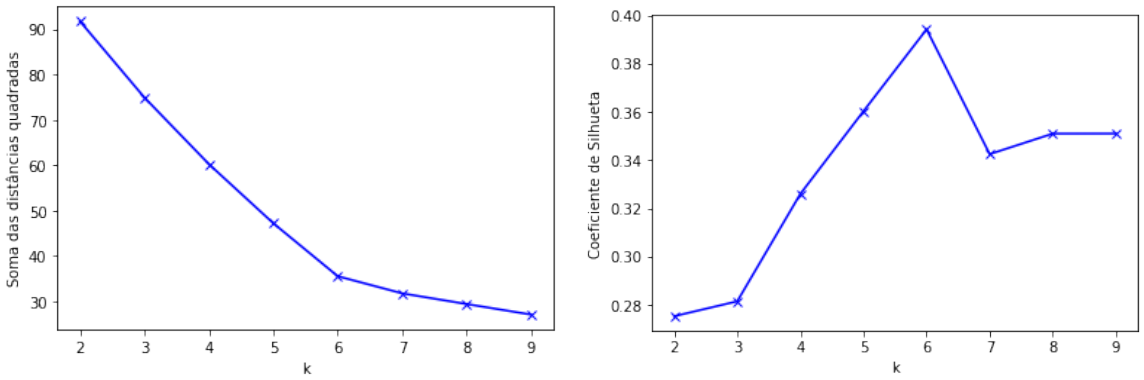
Nesta seção serão apresentados os resultados obtidos com o objetivo de responder às perguntas de pesquisa propostas.

5.1. Pergunta de Pesquisa 1

Com a intenção de identificar a abordagem mais eficiente para realizar o agrupamento de processos judiciais, os diferentes algoritmos de agrupamento foram avaliados em conjunto com diferentes técnicas de representação de textos. Primeiramente, foi necessário utilizar os métodos do cotovelo e da silhueta para identificar a quantidade de grupos ideal [Kodinariya and Makwana 2013]. Esses métodos foram avaliados com as diferentes abordagens de representações de textos com a intenção de verificar a quantidade de grupos mais adequada para ser utilizada nos experimentos. A técnica de TF-IDF + PCA apresentou características que tornaram possível avaliar a quantidade de grupos ideal através desses métodos. Nas Figuras 2(a) e 2(b) é possível verificar que a quantidade de grupos indicada é 6, pois é o ponto em que se encontra o "cotovelo"e o ponto que apresenta o maior valor para o coeficiente de silhueta.

Após a identificação da quantidade de grupos ideal, os três algoritmos de agrupamento descritos (*K-Means, Agglomerative Clustering e Spectral Clustering*) foram aplicados as três representações do texto descritas (TF-IDF+Seleção de Atributos, TF-IDF + PCA e *Word Embeddings*). Na Figura 3 é possível observar os coeficientes de silhueta resultantes dos experimentos. Utilizando a representação de texto TF-IDF+Seleção de atributos, os algoritmos *K-Means* e *Spectral clustering* apresentaram os melhores resultados. Com a representação de texto baseada em TF-IDF+PCA, o algoritmo *K-mean*s apresentou o melhor resultado, sendo este semelhante ao apresentado com o algoritmo *Spectral Clustering*. Já para a representação baseada em *Word embeddings* o algoritmo *Spectral clustering* obteve o melhor resultado.

Figura 2. Identificação da quantidade de grupos



(a) Método do Cotovelo

(b) Método da Silhueta

Como descrito na seção 4.4, o coeficiente de silhueta varia entre -1 e 1 e quanto maior o coeficiente, mais bem definidos são os grupos. Levando isso em consideração, a abordagem que apresentou o melhor resultado foi a combinação da representação de texto baseada em TF-IDF+PCA e o algoritmo de agrupamento *K-Means*.

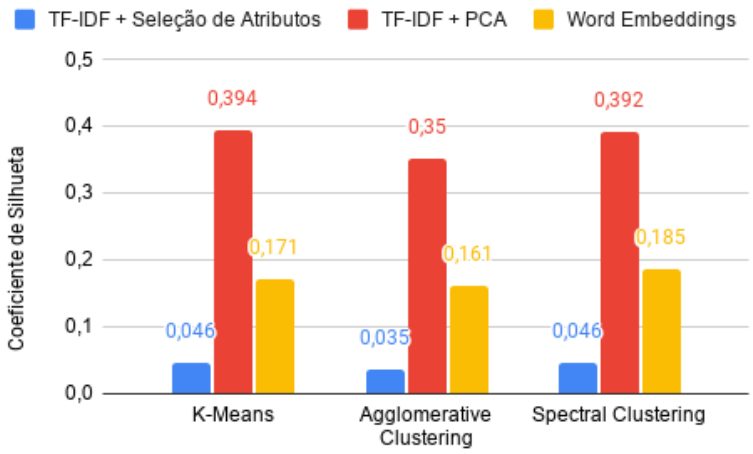


Figura 3. Coeficientes de Silhueta dos Experimentos

A Figura 4 apresenta os resultados dos experimentos levando em consideração o índice de Calinski-Harabasz. Para esse índice devemos avaliar levando em consideração o maior valor obtido, pois valores altos indicam uma melhor separação entre os grupos. Nesse caso, o melhor resultado para todos os tipos de representação de texto, foi obtido com o algoritmo *K-Means*. Já a Figura 5 apresenta os resultados do índice de Davies-Bouldin e esse índice aponta que quanto menor o valor, melhor é a partição entre os grupos. Dessa forma, a melhor abordagem foi a combinação da representação de texto TF-IDF+PCA com o algoritmo de agrupamento *K-Means*.

Observando as Figuras 3, 4 e 5 é possível identificar que a forma de representação textual apresentou um impacto evidente nos resultados. Já os diferentes algoritmos de agrupamento não apresentam diferenças tão acentuadas entre si. Dessa forma, podemos

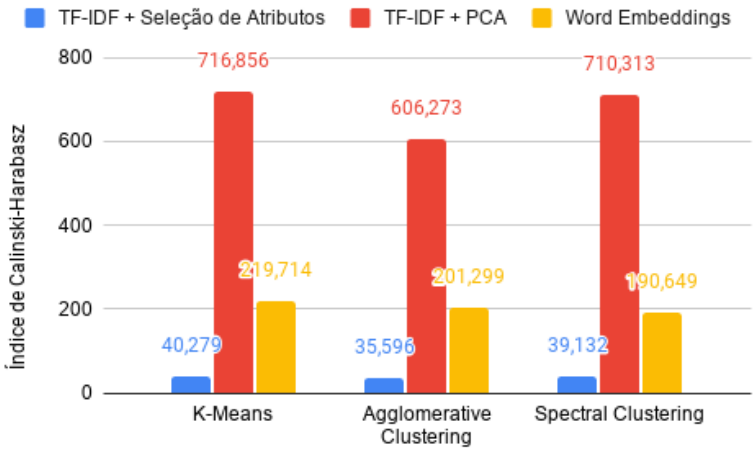


Figura 4. Índices de Calinski-Harabasz dos Experimentos

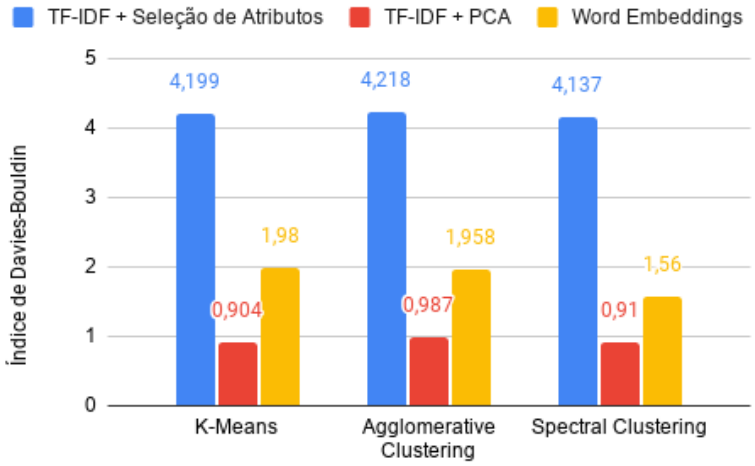


Figura 5. Índices de Davies-Bouldin dos Experimentos

responder a **pergunta de pesquisa 1** indicando que a melhor abordagem para realizar o agrupamento de processos jurídicos utilizando o texto de fatos da petição inicial, utiliza a técnica de representação de texto TF-IDF+PCA. Além disso, apesar de não apresentar resultados impactantes em relação aos outros algoritmos de agrupamento, o algoritmo *K-Means*, quando combinado com a técnica TF-IDF+PCA, apresentou os melhores valores para as métricas analisadas.

5.2. Pergunta de Pesquisa 2

Para responder se a abordagem de agrupamento selecionada é eficiente o suficiente para ser utilizada como ferramenta de apoio para a tomada de decisão, o agrupamento obtido através dessa abordagem foi avaliado levando em consideração os textos de diferentes documentos e os grupos aos quais foram atribuídos. Inicialmente, foram selecionados aleatoriamente 30 pares de documentos que receberam uma pontuação de 0 a 10 que deveria levar em consideração a similaridade entre os pares selecionados e a possibilidade de um dos documentos apoiar a tomada de decisão relacionada ao outro. Essa pontuação

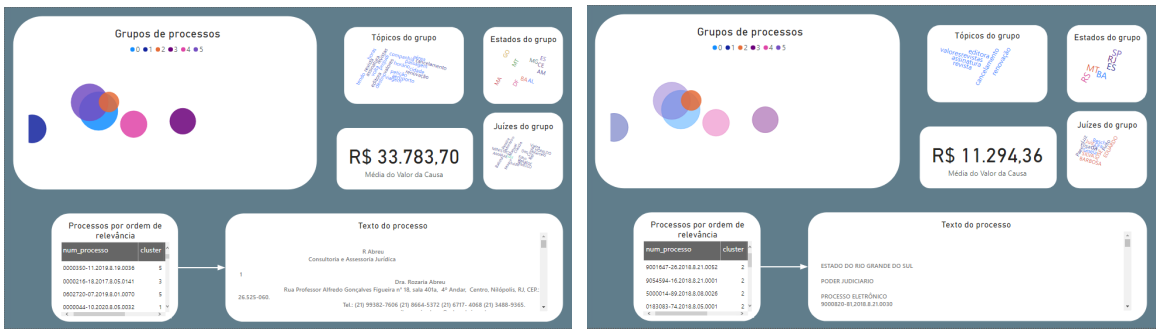
foi atribuída de forma manual, através da leitura dos pares de documentos selecionados. Após isso, cada par foi classificado levando em consideração as seguintes regras:

- **Verdadeiro Positivo (VP)** - Pontuação >= 5 e foram atribuídos ao mesmo grupo
- **Verdadeiro Negativo (VN)** - Pontuação < 5 e foram atribuídos a grupos diferentes
- **Falso Positivo (FP)** - Pontuação < 5 e foram atribuídos ao mesmo grupo
- **Falso Negativo (FN)** - Pontuação >= 5 e foram atribuídos a grupos diferentes

Dos 30 pares selecionados, a abordagem de agrupamento *K-Means* e TF-IDF+PCA apresentou 27 acertos e apenas 3 erros, sendo eles 8 classificados como VP, 19 classificados como VN, 2 classificados como FP e 1 classificado como FN. Além disso, também foi realizado o cálculo das métricas de precisão, cobertura e F1 que resultaram em 0.80, 0.89 e 0.84, respectivamente. As métricas avaliadas para responder essa pergunta de pesquisa apresentaram bons resultados e através delas podemos concluir que a abordagem que combina o algoritmo de agrupamento *K-Means* e a técnica de representação de texto TF-IDF+PCA pode ser utilizada como ferramenta para apoiar a tomada de decisão em aplicações reais do mundo jurídico.

Por fim, uma ferramenta que utiliza os resultados do agrupamento foi planejada e implementada com o objetivo de apoiar a tomada de decisão por parte dos advogados. Essa versão pode ser visualizada na Figura 6(a), que contém os grupos resultantes do agrupamento, algumas informações associadas aos grupos e os textos dos processos do grupo ordenados por ordem de relevância. Essa ferramenta é interativa e o usuário pode selecionar um determinado grupo e as informações serão atualizadas levando em consideração os processos desse grupo. Na Figura 6(b) é possível visualizar as informações atualizadas ao selecionar um determinado grupo.

Figura 6. Ferramenta com resultados do agrupamento



(a) Visão Geral

(b) Grupo selecionado

6. Conclusão

Neste trabalho foram investigadas diferentes abordagens para realizar o agrupamento de processos jurídicos. O agrupamento foi realizado com a intenção de agrupar processos semelhantes e facilitar a coleta de informações que sejam valiosas para o andamento de um processo e que sirvam como apoio à tomada de decisão por parte dos advogados. Para identificar a melhor abordagem de agrupamento, diferentes algoritmos e diferentes representações de texto foram avaliadas em uma base de dados composta por 1515 textos de fatos de petições iniciais.

As avaliações realizadas indicam que a melhor abordagem para realizar o agrupamento de processos jurídicos é composta pelo algoritmo de agrupamento *K-Means* e pela técnica de representação TF-IDF em combinação com a técnica de extração de atributos PCA. Uma outra avaliação que leva em consideração os textos dos processos agrupados também indica que essa abordagem pode ser utilizada como ferramenta para apoiar a tomada de decisão em aplicações reais do mundo jurídico. Além disto, foi apresentado um protótipo de ferramenta que foi desenvolvida para apresentação das principais informações que foram extraídas dos grupos de documentos. Assim, a proposta descrita neste trabalho foi levada até o usuário final.

Como trabalhos futuros, pretende-se aperfeiçoar os resultados do agrupamento através da adição de atributos não textuais. Além disso, pretende-se expandir as funcionalidades e informações apresentadas na ferramenta de apoio a tomada de decisão desenvolvida com os resultados do agrupamento. Por fim, é necessário realizar um estudo da efetividade das informações dos grupos em um contexto real, com um estudo de usabilidade da aplicação numa ferramenta real.

Referências

Aggarwal, C. C. and Zhai, C. (2012). A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer.

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., and Lampos, V. (2016). Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Amine, A., Elberrichi, Z., and Simonet, M. (2010). Evaluation of text clustering methods using wordnet. *Int. Arab J. Inf. Technol.*, 7(4):349–357.

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer.

Chen, B., Li, Y., Zhang, S., Lian, H., and He, T. (2019). A deep learning method for judicial decision support. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 145–149. IEEE.

Conrad, J. G., Al-Kofahi, K., Zhao, Y., and Karypis, G. (2005). Effective document clustering for large heterogeneous law firm collections. In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 177–187.

de Colla Furquim, L. O. and De Lima, V. L. S. (2012). Clustering and categorization of brazilian portuguese legal documents. In *International Conference on Computational Processing of the Portuguese Language*, pages 272–283. Springer.

Fan, B., Liu, T., Hu, H., and Du, X. (2010). Law text clustering based on referential relations. In *2010 Fifth Annual ChinaGrid Conference*, pages 60–66. IEEE.

Kachappilly, D. and Wagh, R. (2018). Similarity analysis of court judgments usingclustering of case citation data: a study. *International Journal of Engineering & Technology*, 7(2):855–858.

Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.

Kowsrihawat, K., Vateekul, P., and Boonkwan, P. (2018). Predicting judicial decisions of criminal cases from thai supreme court using bi-directional gru with attention mechanism. In *2018 5th Asian Conference on Defense Technology (ACDT)*, pages 50–55. IEEE.

Liu, T., Liu, S., Chen, Z., and Ma, W.-Y. (2003). An evaluation on feature selection for text clustering. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 488–495.

Lv, B., Hou, W., Liu, G., Gao, J., Yuan, X., Li, P., and Chen, Z. (2018). A deep cfs model for text clustering. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 132–137. IEEE.

Polpinij, J., Bheganan, P., Luaphol, B., Sibunruang, C., and Namee, K. (2020). Identifying of decision components in thai civil case decision by text classification technique. In *International Conference on Computing and Information Technology*, pages 11–20. Springer.

Poudyal, P., Gonçalves, T., and Quaresma, P. (2019). Using clustering techniques to identify arguments in legal documents. In *ASAIL@ ICAIL*.

Raghav, K., Reddy, P. B., Reddy, V. B., and Reddy, P. K. (2015). Text and citations based cluster analysis of legal judgments. In *International conference on mining intelligence and knowledge exploration*, pages 449–459. Springer.

Raghuveer, K. (2012). Legal documents clustering using latent dirichlet allocation. *IAES Int. J. Artif. Intell*, 2(1):34–37.

Rosca, C., Covrig, B., Goanta, C., van Dijck, G., and Spanakis, G. (2020). *Return of the AI: An Analysis of Legal Research on Artificial Intelligence Using Topic Modeling*. CEUR-WS. org.

Thammaboosadee, S., Watanapa, B., and Charoenkitkarn, N. (2012). A framework of multi-stage classifier for identifying criminal law sentences. *Procedia Computer Science*, 13:53–59.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wang, X. and Xu, Y. (2019). An improved index for clustering validation based on silhouette index and calinski-harabasz index. In *IOP Conference Series: Materials Science and Engineering*, volume 569, page 052024. IOP Publishing.

Xiao, G., Chow, E., Chen, H., Mo, J., Guo, J., and Gong, Z. (2017). Chinese questions classification in the law domain. In *2017 IEEE 14th International Conference on e-Business Engineering (ICEBE)*, pages 214–219. IEEE.