



Universidade Federal do ABC  
Centro de Matemática, Computação e Cognição  
Programa de Pós-Graduação em Ciência da Computação

# **Representação de sentenças jurídicas no contexto de agrupamento automático**

**Cristiano Oliveira Gonçalves**

**Santo André - SP**

**2020**



Cristiano Oliveira Gonçalves

# **Representação de sentenças jurídicas no contexto de agrupamento automático**

**Dissertação de Mestrado** apresentada ao Programa de Pós-Graduação da Universidade Federal do ABC como requisito parcial para obtenção do grau de Mestre em Ciência da Computação

Universidade Federal do ABC – UFABC

Centro de Matemática, Computação e Cognição

Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Thiago Ferreira Covões

Santo André - SP

2020

---

Cristiano Oliveira Gonçalves

Representação de sentenças jurídicas no contexto de agrupamento automático/  
Cristiano Oliveira Gonçalves. – Santo André - SP, 2020-  
58 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Thiago Ferreira Covões

Dissertação (Mestrado) – Universidade Federal do ABC – UFABC  
Centro de Matemática, Computação e Cognição  
Programa de Pós-Graduação em Ciência da Computação, 2020.

1. Agrupamento automático. 2. Documentos de texto. 3. Jurimetria I. Ferreira  
Covões, Thiago. II. Universidade Federal do ABC. III. Programa de Pós-Graduação  
em Ciência da Computação. IV. Representação de sentenças jurídicas no contexto  
de agrupamento automático

---

Cristiano Oliveira Gonçalves

# **Representação de sentenças jurídicas no contexto de agrupamento automático**

**Dissertação de Mestrado** apresentada ao Programa de Pós-Graduação da Universidade Federal do ABC como requisito parcial para obtenção do grau de Mestre em Ciência da Computação

---

**Prof. Dr. Thiago Ferreira Covões**  
Orientador

---

Co-Orientador

---

**Professor**  
Convidado 1

---

**Professor**  
Convidado 2

---

**Professor**  
Convidado 3

Santo André - SP  
2020



*À ciência que facilita o acesso à informação*





# Agradecimentos

Voltarei com tempo à esta seção.



*“Não sei o que,  
não sei o que,  
não sei o que lá.”  
(Autor Desconhecido)*



# Resumo

A digitalização de documentos no setor judiciário brasileiro facilita o acesso a documentos de interesse público. No entanto, para que seja possível levantar métricas de interesse deste crescente repositório informacional, é fundamental que se organizem os documentos de maneira a facilitar a recuperação de informações relevantes, e técnicas de aprendizado de máquina podem diminuir o esforço humano na organização de um grande corpus. Neste trabalho, pretende-se comparar o desempenho do agrupamento de sentenças jurídicas sobre a ótica do assunto sobre o qual tratam, de maneira a contribuir com trabalhos futuros que se propõem a aplicar técnicas computacionais para processamento de textos no domínio jurídico. Para isso, foi estruturada uma base de dados extraída do portal e-Saj composta de mais de três milhões de documentos. Em seguida, diferentes representações textuais serão geradas e técnicas de aprendizado de máquina serão comparadas em cada uma destas representações.

**Palavras-chaves:** Agrupamento textual, representação textual, jurimetria



# Abstract

The digitization of documents in the Brazilian judicial sector facilitates access to documents of public interest. However, in order to be able to gather metrics of interest from this growing informational repository, it is essential that documents be organized in a way that facilitates the retrieval of relevant information, and machine learning techniques can reduce human effort in organizing a large corpus. In this work, we intend to compare the performance of sentence clustering on the subject of documents, in order to contribute to future works that may intend to apply computational techniques for processing long texts in the legal domain. For this, a database extracted from the e-Saj portal composed of more than three million documents was assembled. Then, different textual representations will be generated and machine learning techniques will be compared in each of these representations.

**Keywords:** Text clustering, text representation, jurimetrics





# Lista de ilustrações

Figura 1 – <i>Gráfico dos vetores no VSM.</i> . . . . .	7
Figura 2 – <i>Fluxo do processo de agrupamento.</i> . . . . .	12
Figura 3 – <i>Documentos de textos em espaço euclidiano.</i> . . . . .	15
Figura 4 – <i>Escolha do número de grupos pela análise do cotovelo</i> . . . . .	24



# Lista de tabelas

Tabela 1	–	<i>Frases representadas no VSM</i>	6
Tabela 2	–	<i>Tabela de contingência de <math>\mathcal{U}</math> e <math>\mathcal{V}</math></i>	23
Tabela 3	–	<i>As 10 classes mais frequentes no corpus</i>	34
Tabela 4	–	<i>Os 10 assuntos mais frequentes no corpus</i>	34
Tabela 5	–	<i>Os 10 magistrados mais frequentes no corpus (continua)</i>	34
Tabela 6	–	<i>As 10 comarcas mais frequentes no corpus</i>	35
Tabela 7	–	<i>Os 10 foros mais frequentes no corpus</i>	35
Tabela 8	–	<i>As 10 varas mais frequentes no corpus</i>	36
Tabela 9	–	<i>Quantidade de documentos de acordo com as datas de disponibilização.</i>	36
Tabela 10	–	<i>Quantidade de classes por assunto (continua)</i>	36
Tabela 11	–	<i>Exemplo de termos do TSTF em formato tabular</i>	39
Tabela 12	–	<i>Cronograma estimado de desenvolvimento do trabalho.</i>	42



# Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
abnTeX	Normas para TeX



# Lista de símbolos

$\Gamma$	Letra grega Gama
$\Lambda$	Lambda
$\zeta$	Letra grega minúscula zeta
$\in$	Pertence





# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>1</b>
<b>1.1</b>	<b>Objetivos</b>	<b>2</b>
1.1.1	Objetivos Específicos	3
<b>1.2</b>	<b>Organização do trabalho</b>	<b>4</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>5</b>
<b>2.1</b>	<b>Representação textual no contexto de aprendizado de máquina</b>	<b>5</b>
<b>2.2</b>	<b>Agrupamento</b>	<b>11</b>
2.2.1	Medidas de similaridade	13
2.2.2	Algoritmos de agrupamento	16
2.2.3	Avaliação de agrupamentos	21
<b>2.3</b>	<b>Direito e as ciências exatas</b>	<b>26</b>
<b>3</b>	<b>TRABALHOS RELACIONADOS</b>	<b>29</b>
<b>4</b>	<b>METODOLOGIA</b>	<b>33</b>
<b>4.1</b>	<b>Corpus</b>	<b>33</b>
<b>4.2</b>	<b>Tesouro Jurídico do Supremo Tribunal Federal</b>	<b>37</b>
<b>4.3</b>	<b>Plano de trabalho</b>	<b>38</b>
<b>4.4</b>	<b>Cronograma</b>	<b>42</b>
	<b>REFERÊNCIAS</b>	<b>43</b>
	<b>APÊNDICES</b>	<b>53</b>
	<b>APÊNDICE A – ATIVIDADES DE PREPROCESSAMENTO</b>	<b>55</b>
	<b>ANEXOS</b>	<b>57</b>



# 1 Introdução

A produção de documentos digitais aumentou muito nos últimos anos graças aos adventos da popularização da internet e do barateamento das tecnologias da informação. A digitalização da documentação de diversos processos contemporâneos é uma consequência natural destes eventos, e o setor judiciário brasileiro é um exemplo do exposto: se há vinte anos a consulta de andamentos e decisões em processos jurídicos era restrita às autoridades e partes envolvidas devido à necessidade de acesso físico aos documentos, hoje temos um cenário muito diferente. Portais como o e-Saj<sup>1</sup>, por exemplo, contribuem para tornar acessíveis os dados de interesse público que estão em poder da justiça brasileira. Os registros dos eventos processuais deixaram de ser feitos exclusivamente em arquivos de papel cujo armazenamento se dá em extensas prateleiras e passaram a ocupar também espaço em discos rígidos de servidores *web*.

Esta mudança de paradigma traz novos desafios na organização da informação. Se antigamente as restrições de armazenamento, catalogação e busca de documentos jurídicos eram de naturezas espaciais e logísticas, hoje sua natureza é também computacional. Enquanto que acessar um documento específico com base no número do processo ao qual ele pertence é uma tarefa relativamente simples para o computador, buscar as sentenças mais relevantes de um determinado assunto, por exemplo, é muito mais difícil caso os dados não estejam previamente categorizados. Portanto, o processo de recuperação de informações poderia se beneficiar de uma classificação adequada dos arquivos, e o agrupamento de documentos com conteúdo parecido pode ser útil para auxiliar nesta classificação e facilitar a busca por documentos com certa característica. No entanto, o grande esforço humano necessário para organizar a crescente informação digital justifica o interesse em automatizar estes processos de organização.

Para auxiliar na automatização dessa organização de textos, técnicas de Processamento de Linguagem Natural (PLN) podem ser relevantes. O processamento da linguagem natural (PLN) trata computacionalmente os diversos aspectos da comunicação humana. Jurafsky e Martin (2000), Gonzalez e Lima (2003), e diversos estudos, como Furquim Luis Otávio de Colla (2011), Mikolov et al. (2013) e Paik (2013), apresentam formas de representar textos como vetores para que possam ser submetidos a operações matemáticas. Esta representação viabiliza a aplicação de algoritmos de aprendizado de máquina em grandes volumes de documentos com o intuito de classificar, agrupar, comparar e buscar informações de maneira automática.

Uma forma simples, porém frequentemente eficaz de representar um documento de

---

<sup>1</sup> <https://esaj.tjsp.jus.br/>

texto como vetor é chamada de *bag of words* (HARRIS, 1954). Nesta representação, cria-se um dicionário com as palavras existentes em um conjunto de documentos, e o número de vezes que cada verbete deste dicionário aparece em um documento é usado como um atributo deste documento. Desta forma, é possível criar uma tabela na qual cada linha representa um conteúdo textual de interesse e cada coluna representa o número de vezes que um verbete do dicionário apareceu no texto.

Quando o volume e o tamanho dos documentos analisados crescem, é natural que o número de verbetes fique ainda maior, e isso leva à geração de tabelas muito esparsas que devido à maldição da dimensionalidade, não são desejáveis em tarefas de aprendizado de máquina (BISHOP, 2006; ALPAYDIN, 2010).

Uma alternativa a representações desta natureza é a criação de *word embeddings*, vetores densos que representam palavras e que são gerados por meio de técnicas como redes neurais (MIKOLOV et al., 2013) e decomposições matriciais (LANDAUER; FOLTZ; LAHAM, 1998). Os documentos de texto são então representados como uma operação nos vetores das palavras que os compõem, de maneira que se todas as palavras possuem dimensão  $M$ , o documento de texto também terá uma representação com dimensão  $M$ .

Embora esta alternativa reduza o problema de dimensionalidade na representação dos documentos, a operação escolhida para representar o texto – normalmente soma ou média dos vetores de suas palavras – pode fazer com que palavras muito relevantes para diferenciação dos documentos entre si sejam *diluídas*. Portanto, identificar aquelas palavras com mais relevância para a tarefa que se deseja desempenhar com os documentos de interesse se faz fundamental, como no caso da representação *bag of terms and law references* proposta em Furquim Luis Otávio de Colla (2011).

Uma vez que documentos jurídicos como sentenças e acórdãos são relativamente longos e gozam de um vocabulário muito particular, estudar representações e técnicas que melhoram a eficiência de sua organização automática pode contribuir para melhor aproveitamento do processo de digitalização do sistema judiciário. A hipótese deste trabalho é a de que é possível aprender características do linguajar jurídico de forma não supervisionada, empregando-as de maneira que o desempenho na tarefa de agrupamento de decisões jurídicas frente ao assunto do qual tratam seja superior àquele observado quando estas características não são empregadas na representação das decisões.

## 1.1 Objetivos

Pretende-se identificar a representação textual e a técnica de agrupamento que apresentam os melhores desempenhos na tarefa de agrupar decisões jurídicas de primeira instância segundo o rótulo *assunto* das jurisprudências. O cumprimento do objetivo será

avaliado conforme documentos disponíveis no sistema e-Saj<sup>2</sup>.

### 1.1.1 Objetivos Específicos

Além do objetivo geral apresentado, intenciona-se:

- Criar o corpus de decisões de primeira instância;
- Desenvolver representações vetoriais das palavras deste corpus;
- Apresentar as similaridades e diferenças das representações de palavras nos domínios geral e específico da língua portuguesa, tendo como referência de associação entre palavras o Tesouro Jurídico do Supremo Tribunal Federal<sup>3</sup>;
- Criar representações para os documentos do corpus usando agregação de vetores de palavras;
- Criar representação para os documentos do corpus baseada no domínio jurídico usando vetores de documentos;
- Agrupar os documentos usando diferentes técnicas de agrupamento em todas as representações geradas e avaliar o desempenho do agrupamento tendo como rótulo o assunto dos documentos;
- Identificar superclasses e subclasses dos assuntos;

Objetiva-se ainda produzir resultados que contribuam nas respostas para as seguintes perguntas:

- Quais elementos linguísticos permitem agrupar os documentos por assunto?
- Qual combinação de algoritmo e de representação apresenta o melhor resultado na identificação do assunto?
- Existe evidência de que alguma ação possa melhorar ainda mais o desempenho dos algoritmos nesta tarefa?
- Quais as implicações computacionais de escalar o trabalho para um número ainda maior de documentos?
- A estrutura dos documentos evidencia superclasses e subclasses de interesse?

---

<sup>2</sup> <<https://bit.ly/36SPXEw>>

<sup>3</sup> <<https://bit.ly/2XEfZbI>>

## 1.2 Organização do trabalho

O restante deste trabalho está organizado da seguinte forma:

- **Capítulo 2 - Fundamentação teórica:**
- **Capítulo 3 - Trabalhos relacionados:**
- **Capítulo 4 - Metodologia:**

## 2 Fundamentação Teórica

Este capítulo apresenta conceitos de representação textual no contexto de Aprendizado de Máquina e Agrupamento de Dados, além de introduzir noções de Jurimetria que são relevantes no escopo deste estudo.

### 2.1 Representação textual no contexto de aprendizado de máquina

Documentos de texto são artefatos de linguagem natural que, por meio de uma estrutura gramatical e da convencionalidade da escrita (MANNING; MANNING; SCHÜTZE, 1999), codificam informação. Podem ser definidos por uma sequência de outros artefatos, como orações, palavras ou letras, ou ainda conforme sua estrutura gramatical e sintática. Cada um destes artefatos componentes pode ter relevância sozinho, mas é principalmente no conjunto de artefatos que reside o grande interesse do processamento de linguagem natural.

A Recuperação da Informação é a disciplina que estuda maneiras eficientes de se recuperar informação relevante em um grande conjunto de informação, de maneira que a informação recuperada satisfaça uma necessidade (SCHÜTZE; MANNING; RAGHAVAN, 2008). Frequentemente, este conjunto de informação está armazenado em computadores.

A tarefa de recuperar de informação relevante em coleções de documentos possui desafios particulares cujas soluções podem ser extrapoladas para outras atividades, como agrupamento ou classificação. Como exemplo, verifica-se no problema de recuperar documentos que podem ser relevantes para uma frase de busca, uma tarefa na qual muitas são as maneiras de escolher quais documentos devem ser retornados como resultado da frase. Pode-se optar por procurar na coleção de documentos a frase de busca de maneira literal, e retornar todos aqueles documentos que a contiverem. Esta solução resolve uma parte do problema de recuperar informação relevante, mas em contextos nos quais o conteúdo da coleção não é de conhecimento de quem busca, pode ser preferível que todos os documentos que tenham todos os termos da frase de busca sejam retornados. Outra opção seria incluir termos associados aos termos da frase de busca para que também sejam considerados no momento de retornar os documentos relevantes. No entanto, determinar a associação correta dos milhares de termos que podem surgir em uma fase de busca é uma tarefa mais complexa do que a busca literal de termos, e diferentes representações dos documentos podem ser mais ou menos efetivas para que esta associação implique em resultados satisfatórios. Schütze, Manning e Raghavan (2008)

O trabalho publicado por Salton, Wong e Yang (1975) propõe um modelo formal de

representação de documentos chamado de *VSM* (do inglês: *Vector Space Model*, ou modelo de espaço vetorial), no qual um documento  $\mathbf{d}$  é representado por um vetor  $\mathbf{t} = (t_1, t_2, \dots, t_m)$ , onde  $t_i$  é o peso do  $i$ -ésimo termo. Assim, cada documento neste espaço possui  $|\mathbf{t}|$  dimensões e  $d_{ij}$  é o peso do  $j$ -ésimo termo no  $i$ -ésimo documento de uma coleção de documentos  $\mathcal{D}$ .

Uma maneira de definir o peso de uma palavra em um documento se chama *bolsa de palavras*, ou *BOW* (do inglês: *bag of words*), técnica na qual um número é atribuído a um termo sem considerar a ordem na qual o termo aparece no documento. Um exemplo do exposto é contar quantas vezes cada termo ocorre em cada documento. Assim, considerando que  $\mathbf{t} = (\text{"audiência"}, \text{"chute"}, \text{"mercado"})$  para facilitar a visualização, a Tabela 1 e a Figura 1 apresentam como quatro frases seriam representadas no *VSM*.

Tabela 1 – Frases representadas no *VSM*

i	frase	audiência	chute	mercado
1	A audiência concluiu de maneira decepcionante; o resultado foi equivalente a um chute.	1	1	0
2	Foi convocada uma audiência para avaliar a legitimidade da reação do mercado	1	0	1
3	Depois do chute de Gabriel, o jogo praticamente acabou.	0	1	0
4	O mercado se encontrava fechado. A expectativa é a de abertura do mercado após a audiência.	1	0	2

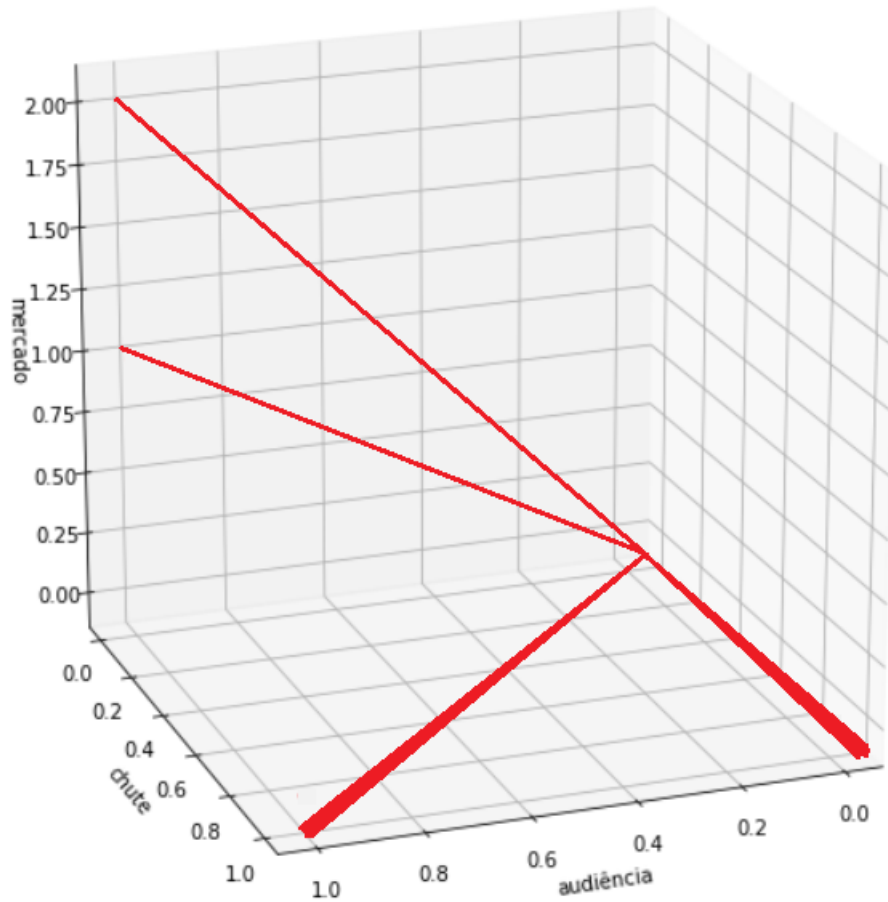
Um problema que precisará ser resolvido no uso do *VSM* consiste em selecionar quais são os termos que devem compor  $\mathbf{t}$ . O uso de todos os verbetes do idioma implicaria em um número muito grande de dimensões, o que dificulta a comparação dos documentos por gerar vetores esparsos<sup>1</sup>. Assim, é comum que sejam empregadas algumas técnicas de processamento de linguagem natural no preparo do texto antes de sua transformação para o *VSM*, como a remoção das palavras mais comuns no idioma e a remoção do sufixo das palavras, tarefa conhecida como *stemming* (PORTER et al., 1980).

Outro aspecto da representação de textos em espaço vetorial é o peso que deve ser usado para os verbetes. No exemplo acima, a contagem de termos foi usada, mas ela pode fazer com que vetores que tenham exatamente as mesmas palavras em quantidades diferentes fiquem numericamente muito diferentes entre si. Se existir a premissa de que documentos de texto que possuem os mesmos termos são mais similares, deseja-se que o mesmo aconteça no espaço em que os documentos são representados. Uma das maneiras de fazê-lo é ponderar as frequências de termos partindo da premissa de que aqueles termos que diferenciam os documentos do resto da coleção são mais relevantes do que aqueles cuja probabilidade de aparecerem nos documentos é maior. Em outras palavras, deseja-se ponderar a frequência do termo em um documento pelo inverso da sua frequência nos

<sup>1</sup> Isso acontece pois, considerando todos os verbetes possíveis, apenas um número pequeno de termos acontece em um documento qualquer



Figura 1 – Gráfico dos vetores no VSM.



documentos do corpus (JONES, 1972). Assim, sendo  $\mathcal{D}$  uma coleção de documentos, então uma forma muito usada desta ponderação é conhecida como *tf-idf* (do inglês: *term frequency-inverse document frequency*):

$$tfidf(t, \mathbf{d}, \mathcal{D}) = tf(t, \mathbf{d}) \cdot idf(t, \mathcal{D})$$

onde  $tf(t, \mathbf{d})$  é a quantidade de vezes que o termo  $t$  aparece no documento  $\mathbf{d}$ ,  $idf(t, \mathcal{D})$  é dado por  $\log_{10} \left( \frac{|\mathcal{D}|}{n_t} \right)$  e  $n_t$  é a quantidade de documentos  $\mathbf{d} \in \mathcal{D}$  nos quais  $t$  aparece.

Uma das críticas relacionadas ao uso de *tf-idf* é a falta de embasamento teórico para justificá-lo (SALTON; BUCKLEY, 1988). Apesar disso, o empirismo mostrou o sucesso desta estatística em diversas aplicações no domínio de Mineração de Dados, como pode ser visto em trabalhos como Hong et al. (2013), Trstenjak, Mikac e Donko (2014), Qaiser e Ali (2018), Zhu et al. (2019). Ramos et al. (2003) examinam seu uso no contexto de recuperação da informação e concluem que indexar documentos usando termos ponderados como *tf-idf* resulta no retorno de resultados adequados às frases de busca usadas. Singh, Tiwari e Garg (2011) concluem que *tf-idf* apresenta melhores resultados na tarefa de agrupamento de documentos quando em comparação ao uso da contagem dos termos.

Apesar do advento de redes neurais artificiais na elaboração de representações de palavras que capturam valor semântico, como será discutido mais adiante, a representação por *tf-idf* continua sendo usada como representação de documentos em diferentes domínios, *e.g.*, como pode ser visto em trabalhos como [Dey et al. \(2020\)](#), [Reshma, Rajagopal e Lajish \(2020\)](#), e [Tummers et al. \(2020\)](#).

Apesar do seu sucesso empírico e popularidade acadêmica, capturar associações entre palavras que possuem forte relação contextual não é uma tarefa trivial. Associações entre palavras como *lâmpada* e *luz* ou *açúcar* e *doce*, e até mesmo entre sinônimos como *buscar* e *procurar*, impõem um importante desafio nas atividades de mineração de textos. Mesmo que cada termo de  $\mathbf{t}$  seja um  $n$ -grama, ou seja, uma sequência de  $n$  palavras observada na coleção e selecionada como um termo  $t$  de  $\mathbf{t}$ , o *tf-idf* tende a não aproximar documentos com base na semelhança semântica dos termos que os compõem ([SALTON; BUCKLEY, 1988](#)), o que induz ao interesse na representação de características textuais mais complexas do que aquelas que a semelhança entre a frequência de termos é capaz de capturar.

[Deerwester et al. \(1990\)](#) discutem de maneira profunda as deficiências dos métodos de representação e indexação de documentos baseados em vetores de termos, e apesar do foco da publicação ser no problema de recuperação de informação, sabe-se que desenvolver representações capazes de associar os documentos semanticamente também contribui para a melhoria de desempenho em tarefas como agrupamento e classificação de documentos. Não obstante, Os autores descrevem um método para resolver os problemas que discutiram que se mostrou muito popular, chamado de *análise semântica latente*, ou apenas *LSA* (do inglês: *latent semantic analysis*).

O objetivo do *LSA* é encontrar a relação de termos e documentos em uma coleção, através da criação de um modelo que permita inferir que um dado termo deveria estar associado a um dado documento mesmo que ele não seja observado no documento em questão. Para tanto, é necessária uma estrutura de dados compatível com a atividade de encontrar os parâmetros deste modelo de inferência. Essa estrutura chama-se *TDM*, acrônimo para *term-document matrix*, ou *matriz termo-documento*. Trata-se de uma matriz cujas linhas representam os termos e as colunas representam os documentos, e os elementos da matriz são frequências de ocorrência dos termos em cada documento. Os elementos da matriz também podem ser um peso *tf-idf*. Em seguida, uma decomposição da matriz em valores singulares ([BLUM; HOPCROFT; KANNAN, 2016](#)) é aplicada, separando a matriz inicial em três matrizes que, espera-se, explicitam a relação semântica entre termos, documentos e tópicos existentes em cada documento.

Uma decomposição de matriz em valores singulares é a fatoração de uma matriz  $\mathbf{M}$  tal que  $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ . Se  $\mathbf{M}$  possui  $m$  linhas e  $n$  colunas, então  $\mathbf{U}$  e  $\mathbf{V}^*$  são matrizes quadradas de ordem  $m \times m$  e  $n \times n$ , respectivamente, enquanto que  $\mathbf{\Sigma}$  é uma matriz diagonal

de ordem  $m \times n$ . Uma vez que muitos dos componentes resultantes desta fatoração possuem valores muito pequenos, eles são frequentemente ignorados nas matrizes resultantes, o que faz com que esta decomposição também seja útil como maneira de reduzir o número de dimensões da representação original para  $k$  dimensões, onde  $k$  é o número de fatores usado na decomposição.

No contexto de representação de documentos, a matriz  $\mathbf{M}$  é a matriz de termos e documentos, sendo que seu número de linhas é igual ao número de termos e o número de colunas é a quantidade de documentos da coleção. Sua decomposição em valores singulares faz com que a matriz  $\mathbf{U}$  represente os tópicos de cada termo, a matriz  $\mathbf{\Sigma}$  representa a relevância de cada tópico na coleção, e a matriz  $\mathbf{V}^*$  representa a distribuição de tópicos em cada documento. Assim, os vetores resultantes do produto entre  $\mathbf{U}$  e  $\mathbf{\Sigma}$  podem ser usados para representar os termos em  $k$  dimensões. Analogamente, a multiplicação entre

A reconstrução da matriz  $\mathbf{M}$  através da multiplicação de suas componentes de menor dimensão gera uma matriz  $\mathbf{M}'$  na qual tanto documentos quanto termos são representados pela combinação linear dos termos originais, de forma que termos que aparecem frequentemente juntos sejam reduzidos ao mesmo componente. Assim, os novos vetores são capazes de incorporar, até certa medida, a associação entre termos e incorporá-la na representação dos documentos.

A proximidade destes vetores, sejam eles dos termos ou dos documentos, neste novo espaço vetorial  $k$  dimensional, pode então representar a proximidade do significado destes vetores de maneira mais indireta e abstrata do que se consegue com a representação original. Assim, para um documento  $\mathbf{d}$ , uma nova representação  $\hat{\mathbf{d}}$  de  $k$  dimensões se dá por  $\hat{\mathbf{d}} = \mathbf{\Sigma}_k^{-1} \mathbf{U}_k^T \mathbf{d}$ .

Zhang, Yoshida e Tang (2011) comparam *tf-idf* e *LSA* nas tarefas de recuperação de informação e classificação de documentos, concluindo que esta última é a que apresenta o melhor desempenho. Tang et al. (2005) comparam diferentes métodos de redução de dimensionalidade no contexto de agrupamento de documentos e também conclui que *LSA* apresenta os melhores resultados. Portanto, além de oferecer representações tanto para termos quanto para documentos, *LSA* é capaz de ser útil em diversos desafios na representação de documentos para a tarefa de aprendizado de máquina. Depende, no entanto, na determinação de um valor para  $k$  e é agnóstica à ordem de aparecimento das palavras nos documentos. Uma vez que a ordenação é uma importante característica do sentido que damos às frases, métodos que consigam representá-la são desejáveis.

Uma abordagem que ganhou muita relevância e popularidade (LEVY; GOLDBERG, 2014; PAPAKYRIAKOPOULOS et al., 2020) devido aos resultados encorajadores na tarefa de capturar sentido semântico e sintático, ao mesmo tempo que consegue incorporar a ordem de ocorrência das palavras no texto, é usar algoritmos de redes neurais que têm vetores de palavras tanto como entrada quanto como saída. Nesta abordagem, cada palavra

é representada por um vetor único de  $k$  dimensões, e objetiva-se treinar um algoritmo capaz de identificar o vetor de palavras de saída mais provável dado um conjunto ou uma sequência de vetores de palavras de entrada. Os pesos associados a cada uma das  $k$  dimensões são então usados como representações de cada uma das palavras. Comumente refere-se a esse vetor de pesos como *embedding*. Quando o objetivo do algoritmo é descobrir o vetor da palavra central de uma sequência com número ímpares de palavras, refere-se ao processo de treinamento como *CBOW*, ou *continuous bag of words*. Quando deseja-se descobrir todos os antecessores e sucessores de uma palavra de entrada, denomina-se o processo de treinamento como *skip-gram* (MIKOLOV et al., 2013).

Diferentes arquiteturas de redes neurais têm sido introduzidas na literatura, e métodos de representação vetorial de palavras vêm sendo adaptados para uso em diferentes tarefas de processamento de linguagem natural com notável sucesso (PENNINGTON; SOCHER; MANNING, 2014; LING et al., 2015; BOJANOWSKI et al., 2017). Um problema que precisa ser resolvido no uso destes métodos é o de representação dos documentos, uma vez que os modelos de palavras associam cada palavra a um vetor, e não cada documento a um vetor. Para este fim, e considerando que um documento é qualquer sequência de palavras, trabalhos como Le e Mikolov (2014a) e Dai, Olah e Le (2015) usam arquiteturas de redes neurais que expandem aquelas usadas no treinamento de vetores de palavras para gerar vetores que representam frases, parágrafos ou documentos inteiros. Esta representação consiste em concatenar vetores cuja função é representar documentos a algum tipo de agregação dos vetores de palavras que compõem aquele documento, frase ou parágrafo, e o algoritmo de treinamento aprende ambos os vetores. Esta abordagem é particularmente usada na resolução de problemas supervisionados ou semi-supervisionados envolvendo a classificação de documentos e análise de sentimentos (KIM et al., 2019; BILGIN; ŞENTÜRK, 2017; LEE; YOON, 2018; LEE; JIN; KIM, 2016; TRIEU; TRAN; TRAN, 2017).

Uma maneira mais simples de contornar este problema é representar os documentos apenas com a agregação dos vetores de palavras que os compõem (FERRERO et al., 2017). Um dos desafios desta abordagem é que quando o documento representado possui muitas palavras, o peso de cada palavra no vetor final é diluído. Portanto, é razoável usar a soma ou a média dos vetores de palavras ponderados pelo *idf* delas. É possível ainda diminuir esta diluição por meio de da remoção do sufixo das palavras nos documentos, o que faz com que termos como *reagiu* e *reagindo* sejam reduzidos ao seu radical *reag*. Essa redução pode ser feita mediante aplicação de *stemming* ou *lematização* (LOVINS, 1968).

O uso destas diferentes técnicas de representação no mesmo contexto de agrupamento de documentos jurídicos é interessante porque o linguajar jurídico possui particularidades que cada uma destas técnicas de representação pode capturar de maneiras diferentes. Não foram encontrados trabalhos que analisam o efeito destas diferentes técnicas no

agrupamento de documentos do domínio jurídico da língua portuguesa durante a redação desta dissertação, e a aplicabilidade das conclusões deste estudo podem contribuir com a organização do crescente acervo digital da justiça brasileira.

## 2.2 Agrupamento

Intuitivamente, agrupamento refere-se à atividade de agrupar, ou seja, ao ato ou à ação de dividir em grupos. Dentre as diversas definições encontradas para a palavra *grupo* no dicionário Michaelis Online<sup>2</sup>, destaca-se abaixo aquelas que não possuem uso em domínio específico do conhecimento e aquela cujo uso se dá na biologia:

### grupo

#### gru-po

sm

1 Conjunto de pessoas ou coisas que formam um todo: “*No canto da sala, havia um grupo de carteiras amontoadas. Na entrada da escola, um grupo de meninas conversava.*”

2 Agrupamento de diversas pessoas: “*De quando em quando, de entre o grosso e macho vozear dos homens, esguichava um falsete feminino, tão estridente que provocava réplica aos papagaios e aos perus da vizinhança. E, daqui e dali, iam rebentando novas algazarras em grupos formados cá e lá pela estalagem*” (AA1).

3 Conjunto de seres ou coisas previamente estabelecidos e para fins específicos: “*A chamada ainda durou algum tempo, porque Amâncio era dos primeiros; afinal, o bedel mastigou o último nome; fechou-se a porta da sala; e um silêncio formalista espalhou-se entre a turma dos estudantes e o grupo dos examinadores*” (AA2).

[...]

8 BIOL Conjunto de seres com características comuns, organizados em categorias sistemáticas.

[...]

Em aprendizado de máquina, o termo *grupo* pode referir-se a um conjunto de objetos que são mais similares entre si do que aos objetos de outros conjuntos, estejam estes objetos organizadas em categorias sistemáticas ou não. Wolkind e Everitt (1974 apud JAIN; DUBES, 1988) documenta as seguintes definições para grupo:

1 Um grupo é um conjunto de entidades que são *parecidas*, e entidades que pertencem a grupos diferentes não o são.

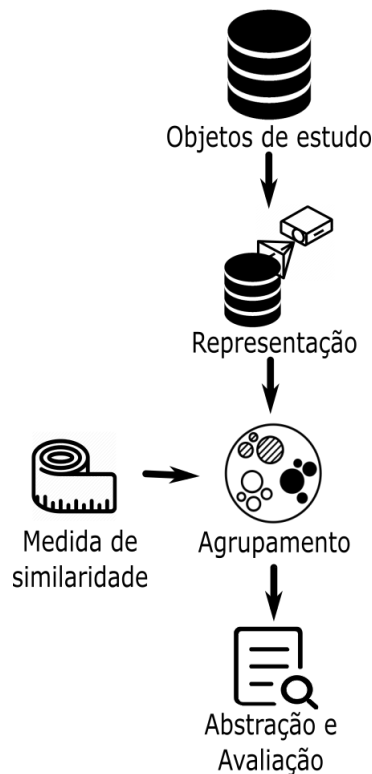
2 Um grupo é uma agregação de pontos no espaço de teste tais quais a *distância* entre quaisquer dois pontos do grupo é menor do que a distância entre qualquer ponto do grupo e qualquer ponto fora dele.

3 Grupos podem ser descritos como regiões conectadas de um espaço multidimensional que contém uma densidade relativamente *alta* de pontos, separadas de outras regiões por uma região contendo uma densidade relativamente baixa de pontos.

<sup>2</sup> <https://michaelis.uol.com.br/>, acessado em março de 2020

O termo *agrupamento* refere-se ao processo por meio do qual se encontra grupos em conjuntos de dados, e trata-se de uma área do conhecimento muito ampla que é estudada por diferentes comunidades científicas e que está fortemente associada ao aprendizado não supervisionado. Neste tipo de aprendizado, não existem dados rotulados disponíveis. Jain e Dubes (1988) apresentam uma sequência de 5 atividades em alto nível para definir o processo de agrupamento: representação de padrões, definição de medida de similaridade de padrões, agrupamento, abstração de dados e avaliação de resultados. A Figura 2 ilustra o processo.

Figura 2 – Fluxo do processo de agrupamento.



Na atividade *representação de padrões*, define-se o número de grupos, a quantidade de objetos de estudo e as características destes objetos que serão empregadas para representá-los. Normalmente os objetos são representados como pontos em um espaço  $n$ -dimensional cujas dimensões armazenam informações de interesse, conforme discutido na seção 2.1. Podem haver restrições que impeçam a verificação destas informações, como nos casos nos quais não se conhece o número de grupos mais adequado ao problema estudado.

A atividade *definição de medida de similaridade de padrões* consiste na escolha de uma função que tem como entrada um par de objetos e que retorna um número que representa sua similaridade. A escolha desta função não é trivial pois o conceito de similaridade pode ser específico ao problema que se deseja resolver. Se analisarmos como exemplo a tarefa de agrupamento de documentos de texto, é perceptível que agrupar diagnósticos médicos frente à gravidade de doenças pode implicar em uma escolha de

medida de similaridade muito diferente do problema de agrupar as páginas *web* retornadas em um buscador. Outro fator que contribui para a pouca trivialidade desta atividade é a grande variedade de medidas existentes e a incerteza sobre a eficiência de cada uma delas. Estes e outros desafios motivaram trabalhos como [Ghosh e Strehl \(2006\)](#), [Metzler, Dumais e Meek \(2007\)](#), [Yih e Meek \(2007\)](#) e [Huang \(2008\)](#).

A atividade *agrupamento* consiste na atribuição de um objeto a um ou mais grupos. A tarefa de atribuir cada objeto a apenas um grupo denomina-se *hard clustering*, enquanto que a tarefa de atribuir um objeto a mais que um grupo ou de encontrar a probabilidade de um objeto pertencer a cada grupo recebe o nome de *soft clustering*.

As duas atividades restantes definidas por [Jain e Dubes \(1988\)](#), *abstração de dados* e *avaliação de resultados*, são opcionais. A primeira consiste em qualificar os agrupamentos encontrados usando avaliação humana ou automática das características dos agrupamentos obtidos, enquanto que a última refere-se a maneiras de verificar se o processo de agrupamento atingiu o objetivo esperado. Uma vez que os dados podem não estar rotulados, não é incomum que se verifique a probabilidade de as condições de agrupamento encontradas, como a quantidade ideal de grupos e a distribuição dos objetos nos grupos, serem obtidas aleatoriamente. Assim, quanto menor esta probabilidade, mais consistentes os resultados.

Para que seja possível aplicar o processo de agrupamento descrito acima, é necessário o aprofundamento em conceitos fundamentais. A [subseção 2.2.1](#) aprofunda o assunto de medidas de similaridade e discute o impacto da escolha de uma determinada medida para o agrupamento de documentos de texto. A [subseção 2.2.2](#) aborda algumas maneiras de desempenhar a tarefa de agrupamento. Finalmente, a [subseção 2.2.3](#) apresenta formas objetivas de avaliar a qualidade dos agrupamentos obtidos tanto quando existe uma agrupamento de referência quanto quando uma não está disponível.

### 2.2.1 Medidas de similaridade

A escolha de uma medida de similaridade adequada é um dos primeiros passos para a execução do processo de agrupamento. Uma vez que não existe uma medida que é sempre melhor para qualquer problema, uma das maneiras de determinar qual a mais adequada é o empirismo. No entanto, executar diversos processos de agrupamento pode ser computacionalmente inviável quando o volume de dados for muito grande, o que faz com que seja de interesse o estudo do comportamento de diferentes medidas em diferentes problemas e algoritmos de agrupamento.

[Strehl, Ghosh e Mooney \(2000\)](#) e [Huang \(2008\)](#) estudaram o efeito de diferentes métricas de similaridade em documentos de texto, e ambos concluíram que elas possuem especificidades frente à qualidade dos grupos que geram. É de particular interesse a



comparação entre a distância euclidiana e a distância de cosseno, na qual conclui-se que a última é mais adequada devido a maneira como os documentos de textos são dispostos quando representados como vetores de características extraídas das palavras que os compõem. Estes estudos concluem ainda que os resultados obtidos com a distância de cosseno são comparáveis àqueles obtidos pelas medidas *distância de jaccard* (JACCARD, 1901), *correlação de pearson*, como aplicada em Sedgwick (2012), e *entropia relativa*, como definida em Bigi (2003), e apresenta propriedades desejáveis como ser invariante à escala das características e retornar um valor no intervalo  $[0,1]$ . Sendo assim, justificaremos o uso da distância de cosseno comparando seu funcionamento com o funcionamento da distância euclidiana como medidas de similaridade de documentos de texto. Para fazê-lo, é necessário compreender as definições formais de métrica, da distância euclidiana e da distância de cosseno.

Sejam  $\mathbf{a}$  e  $\mathbf{b}$  dois vetores, e seja  $d(\mathbf{a}, \mathbf{b})$  a distância entre eles. Diz-se que  $d(\mathbf{a}, \mathbf{b})$  é uma métrica se, e somente se  $d(\mathbf{a}, \mathbf{b})$  respeitar as seguintes condições (HUANG, 2008):

1.  $d(\mathbf{a}, \mathbf{b}) \geq 0$
2.  $d(\mathbf{a}, \mathbf{b}) = 0 \iff \mathbf{a} = \mathbf{b}$
3.  $d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$
4.  $d(\mathbf{a}, \mathbf{c}) \leq d(\mathbf{a}, \mathbf{b}) + d(\mathbf{b}, \mathbf{c})$

A distância euclidiana é uma métrica que mede o quão distantes entre si dois objetos estão no espaço euclidiano por meio do segmento de linha reta entre eles. Se  $\mathbf{a} = [a_1, a_2] = [7, 10]$ ,  $\mathbf{b} = [b_1, b_2] = [2, 4]$  e  $\mathbf{c} = [c_1, c_2] = [10, 8]$  forem documentos de texto em espaço euclidiano representados com base na quantidade de palavras que os compõem e que estão associadas ao assunto *esporte* ou ao assunto *cinema*, então a distância euclidiana entre eles é ilustrada pela linha azul na Figura 3.

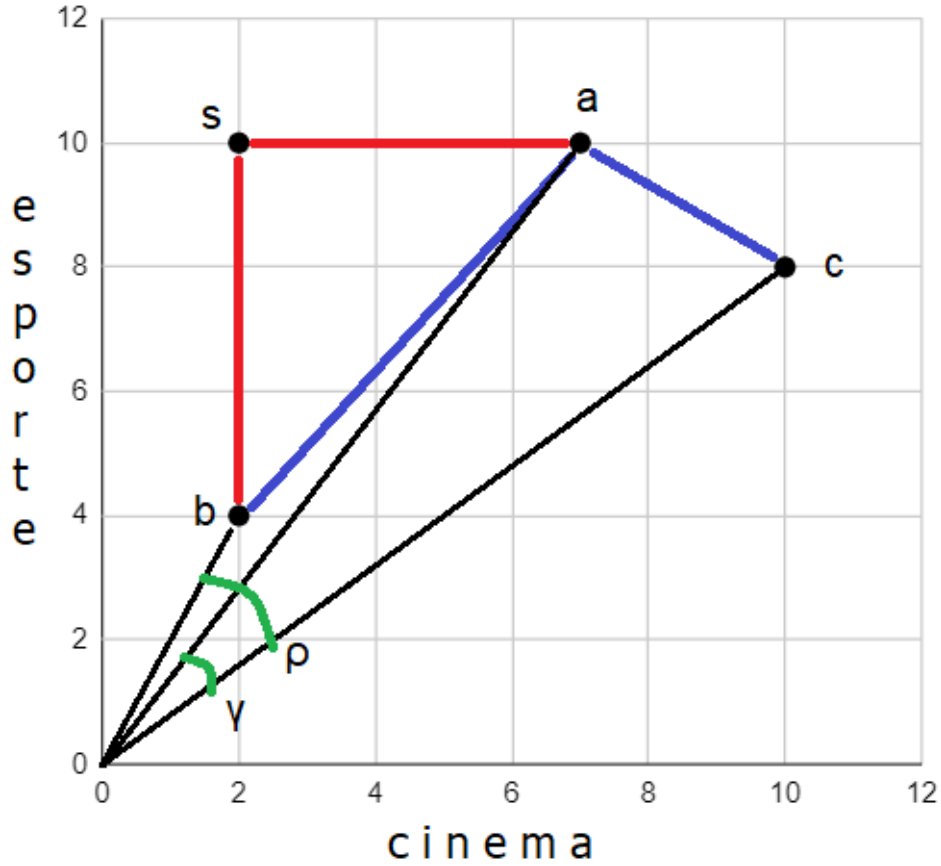
Os pontos  $\mathbf{a}$  e  $\mathbf{b}$  são vértices do triângulo  $abs$ . O espaço entre  $\mathbf{a}$  e  $\mathbf{b}$  é dado pela hipotenusa  $ab$ , cujo comprimento é  $\sqrt{as^2 + bs^2}$ . Sabendo que  $as = |a_1 - b_1|$  e que  $bs = |a_2 - b_2|$ , temos que  $ab = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$ . É possível generalizar esta fórmula para calcular a distância entre dois pontos num espaço  $n$ -dimensional:

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2} = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$$

Uma característica importante da distância euclidiana é que a escala das dimensões afeta o quanto cada uma influencia no valor final calculado, o que demanda cuidado no tratamento dos dados caso ela seja escolhida como medida de similaridade.



Figura 3 – Documentos de textos em espaço euclidiano.



A distância de cosseno é o cosseno do ângulo entre dois vetores de mesma base. A magnitude  $\|\mathbf{v}\|$  de um vetor  $\mathbf{v}$  é o módulo da soma de seus  $n$  componentes, ou seja,  $\sqrt{\sum_{i=1}^n v_i^2}$ . O produto interno de dois vetores  $\mathbf{a}$  e  $\mathbf{b}$  é denotado por  $\mathbf{a} \cdot \mathbf{b}$  e é calculado como a sua magnitude multiplicada pelo cosseno do ângulo entre eles, ou seja,  $\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos(\theta)$ . Portanto, a similaridade de cosseno pode ser escrita como:

$$s(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^n a_i b_i}{\sqrt{\sum_{i=1}^n a_i^2} \sqrt{\sum_{i=1}^n b_i^2}}$$

Dado que  $\cos(0) = 1$  e  $\cos(\theta \geq 0) \leq 1$ , dois vetores são considerados idênticos se o ângulo entre eles for 0, o que faz com que caso os vetores tenham magnitudes diferentes, a distância de cosseno não possa ser considerada uma métrica por desrespeitar a segunda condição. A distância de cosseno possui a vantagem de desconsiderar a magnitude dos vetores ao computar o quão distantes eles são, e essa é uma característica desejável quando se compara documentos de texto.

Ao observar a similaridade entre os pontos pela ótica da distância euclidiana (evidenciada pelas linhas azuis na Figura 3), fica evidente que  $\mathbf{a}$  e  $\mathbf{c}$  formam o par mais similar. Já pela ótica da distância de cosseno (evidenciada pelos arcos verdes),  $\mathbf{a}$  e  $\mathbf{b}$

são mais similares, já que o ângulo entre eles é  $\rho - \gamma$  e que  $\cos(\rho - \gamma) > \cos(\rho)$  e  $\cos(\rho - \gamma) > \cos(\gamma)$ . A distância euclidiana indica que **a** e **c** estão mais próximos entre si porque a magnitude destes vetores é mais parecida dado que a quantidade de palavras dos documentos é similar. Já a distância de cosseno aproxima mais aqueles documentos cuja *proporção* entre os assuntos de palavras é mais similar, mesmo que eles tenham tamanhos diferentes. No entanto, caso a magnitude dos vetores seja a mesma, as medidas possuem uma relação linear.

A escolha da medida de distância ideal pode variar a depender do domínio do problema que se deseja resolver, mas conforme o exposto nesta subseção, a literatura aponta para o uso da distância de cosseno como uma medida eficiente para computar a similaridade entre documentos de texto (STREHL; GHOSH; MOONEY, 2000), (HUANG, 2008), e trabalhos recentes nos quais são usadas representações textuais vetoriais e densas continuam a fazê-lo, como visto em Mikolov et al. (2019), Dai, Olah e Le (2015), Wang et al. (2017) e Hartmann et al. (2017).

## 2.2.2 Algoritmos de agrupamento

A intuição por trás dos algoritmos de agrupamento é relativamente simples: basta comparar os objetos de estudo com os outros objetos, e se eles forem próximos o suficiente, eles devem ficar no mesmo grupo. A prática do processo, no entanto, implica em administrar diversos fatores, como a quantidade ideal de grupos ou o que são bons e maus grupos. Outros fatores que devem ser levados em consideração na escolha do método são a escalabilidade, o formato dos grupos que ele gera, a possibilidade de ter um objeto em mais que um grupo e a flexibilidade do método frente a diferentes medidas de similaridade.

Os algoritmos de agrupamento hierárquico consistem em métodos para transformar a matriz de distância de um grupo de objetos em um conjunto de grupos aninhados (JAIN; DUBES, 1988). Esse aninhamento fornece uma estrutura de hierarquia entre os grupos, de maneira que informações relevantes sobre a estrutura dos dados agrupados possam ser evidenciadas. A intuição por trás destes algoritmos é a de que ao se calcular a distância entre um objeto e todos os outros, aquele par de objetos mais próximos são então considerados um único objeto, e então o processo se repete até que um critério de parada seja atingido. Este critério de parada depende do sentido pelo qual se inicia o cálculo das distâncias. Na abordagem aglomerativa, calcula-se a distância entre cada par de objetos e aqueles mais similares são colocados no mesmo grupo. Os objetos deste grupo são então representados por alguma métrica que represente o grupo inteiro, e o processo se repete até que todos os objetos estejam no mesmo grupo. Karypis, Kumar e Steinbach (2000) apresentam os passos abaixo para ilustrar o agrupamento hierárquico aglomerativo:

1. Compute a similaridade entre todos os pares de grupos, ou seja, calcule a matriz

de similaridade cuja  $ij$ -ésima entrada dá a similaridade entre o  $i$ -ésimo e o  $j$ -ésimo grupo

2. Agrupe os dois grupos mais similares
3. Atualize a matriz de similaridade
4. Repita os passos 2 e 3 até que haja apenas um grupo

Já na abordagem divisiva, assume-se que todos os objetos estão no mesmo grupo, e estes grupos são divididos sucessivamente até que cada grupo seja composto por apenas um objeto.

Como mencionado anteriormente, é necessário criar uma maneira de calcular a distância entre grupos de objetos, e é aqui que reside uma grande diferenciação entre os métodos de agrupamento hierárquico [Aggarwal e Zhai \(2012\)](#). No método de *ligação única* (do inglês: *single-linkage*), a distância entre dois grupos é a menor distância entre quaisquer pares de objetos nestes grupos. Já no método de *ligação completa* (do inglês: *complete-linkage*), a distância entre dois grupos é a menor dentre as maiores distâncias encontradas entre dois objetos de grupos diferentes. Existem diversas outras maneiras de computar a distância entre grupos de objetos, e [Szekely e Rizzo \(2005\)](#) apresentam uma extensão ao método de mínima variância de Ward ([JR, 1963](#)), que pode ser visto como uma fórmula que generaliza diversos métodos para este cálculo. A escolha do método pode afetar sensivelmente os resultados do processo de agrupamento de documentos de texto ([AGGARWAL; ZHAI, 2012](#)), e tanto a alta complexidade computacional de alguns métodos de distância quanto a baixa qualidade de agrupamento de documentos de texto demonstrada por [Karypis, Kumar e Steinbach \(2000\)](#) fomentam o estudo de outras abordagens para o agrupamento de documentos. [Nassif e Hruschka \(2011\)](#) concluem que apesar de os métodos hierárquicos apresentarem bons resultados no agrupamento de documentos, o  $k$ -médiãs inicializado adequadamente apresenta resultados melhores. O algoritmo  $k$ -médiãs é um método de particionamento que consiste em representar um grupo de objetos como o ponto médio, ou centroide, dos os objetos de um determinado grupo. Diferente do agrupamento hierárquico, o uso do  $k$ -médiãs implica em definir *a priori* o número de grupos que se deseja obter. O algoritmo divide o conjunto de dados nestes  $k$  grupos minimizando o erro quadrático entre os elementos do grupo e seu centroide. Devido à facilidade de implementação, à sua eficiência computacional e à relativa interpretabilidade, este algoritmo é muito presente na literatura de descoberta de conhecimento em bancos de dados e de análise multivariada ([AGGARWAL; ZHAI, 2012](#)). Formalmente, seja  $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$  um conjunto de  $n$  objetos, seja  $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$  um conjunto de  $k$  grupos, e seja  $\mu_k = \frac{1}{|\mathcal{C}_k|} \sum_{\mathbf{x}_i \in \mathcal{C}_k} \mathbf{x}_i$  a média das dimensões do grupo  $c_k$ . O  $k$ -médiãs objetiva minimizar o erro quadrático para todos os grupos:

$$J(C) = \sum_{k=1}^K \sum_{a_i \in c_k} \|a_i - \mu_k\|^2$$

Seu processo de otimização pode ser definido por três etapas: *inicialização*, *atribuição* e *atualização*. Na etapa de *inicialização*, são definidos os  $k$  centroides no espaço multidimensional no qual  $\mathbf{a}$  foi definido. Uma vez que o  $k$ -médias faz otimizações locais, é sabido que os resultados do algoritmo são muito sensíveis às condições iniciais. Diferentes estudos apresentam soluções para este problema, como o uso do agrupamento hierárquico para a determinação dos centroides (MILLIGAN, 1980), a criação de um processo global de otimização para definição dos centroides iniciais (LIKAS; VLASSIS; VERBEEK, 2003) ou sucessivas execuções da inicialização aleatória. Pena, Lozano e Larranaga (1999) discutem quatro métodos de inicialização, e sugerem que o método de escolha aleatória e o método de Kauffman (ROUSSEEUW; KAUFMAN, 1990) apresentam os melhores resultados, sendo que este último se mostra o mais adequado em termos de qualidade dos grupos resultantes. Na etapa de atribuição, a similaridade entre cada objeto e cada centroide é calculada, e então os objetos são atribuídos ao grupo cuja similaridade com o centroide é a maior. Embora a métrica de similaridade do algoritmo seja a distância euclidiana, algumas tarefas de agrupamento podem se beneficiar do uso de outras medidas. Bora et al. (2014) compara o algoritmo tradicional com implementações que usam a distância de Manhattan e a distância de cosseno e concluem que a implementação com a distância de Manhattan apresenta os melhores resultados, conclusão essa que também é compartilhada por Loochach e Garg (2012). Dado que os experimentos discutidos foram conduzidos em bases de dados diferentes e com no máximo 207 objetos, pode-se assumir que para trabalhos envolvendo volumes maiores de dados em um contexto de alta dimensionalidade, as contribuições destes estudos são principalmente:

1. Os estudos evidenciam a relevância de se investigar os efeitos de diferentes métricas de distância a depender do contexto dos dados.
2. Os estudos evidenciam a possibilidade de usar uma métrica de distância cuja complexidade computacional é menor do que aquela da distância euclidiana.

Karypis, Kumar e Steinbach (2000) comparam o método hierárquico aglomerativo com duas variantes do algoritmo  $k$ -médias e concluem que a variante deste último que implementa bisseção dos dados apresenta resultados iguais ou superiores ao método hierárquico para agrupar documentos. Além disso, afirmam que este apresenta complexidade computacional quadrática enquanto o  $k$ -médias apresenta complexidade computacional linear. Assim, é comum que os algoritmos hierárquicos sejam usados para contornar limitações ou melhorar os resultados do  $k$ -médias e suas variantes.

No contexto de agrupamento de documentos de texto representados como vetores de características extraídas de suas palavras, o uso da distância euclidiana pode parecer uma escolha evidente, mas como discutido na subseção 2.2.1, documentos com muitas palavras podem gerar vetores com magnitude muito grande, o que os afasta, no espaço euclidiano, de documentos com conteúdo parecido porém com quantidades de palavras diferentes. A fim de mitigar este problema, Dhillon e Modha (2001) sugerem o uso da distância euclidiana das projeções dos vetores que representam os documentos com magnitude normalizada, de maneira que a magnitude de todos os vetores seja 1. Comumente, refere-se a esta transformação como esfera unitária (do inglês: *unit sphere*), e aplicar distância euclidiana como medida de distância é equivalente a aplicar a distância de cosseno (BUCHTA et al., 2012). À esta modificação do  $k$ -médias refere-se como  $k$ -médias esférico (do inglês: *spherical k-means*), cujo objetivo é minimizar  $\sum_{c_k \in C} \sum_{a_i \in c_k} (1 - |\cos(a_i, \mu_k)|)$ .

A literatura aponta que o uso do  $k$ -médias esférico apresenta melhores resultados que a versão clássica do algoritmo na tarefa de agrupamento de documentos de texto, como visto em Lakshmi e Balakrishna (2016) e Zhong e Ghosh (2003), e isso se dá devido às propriedades da distância de cosseno que são desejáveis nesta tarefa. Pesquisadores seguem estudando maneiras de melhorar o desempenho computacional e de desenvolver outras propriedades desejáveis para o algoritmo, como visto em Kim, Kim e Cho (2020), Li et al. (2019) e Tunali, Bilgin e Camurcu (2016), o que justifica seu uso em trabalhos no contexto de recuperação de informação e organização automática de documentos.

O  $k$ -médias é um caso específico de uma classe de modelos mais abrangente chamada de modelos mistura, ou *mixture models*, entre os quais um método muito popular no agrupamento de documentos é o *Expectation-Maximization* (EM), que consiste em encontrar distribuições de probabilidade que mais se aproximem dos dados sendo agrupados através da estimação de máxima verossimilhança. Na etapa **Expectation**, o algoritmo calcula as probabilidades de os dados ocorrerem dada uma distribuição de probabilidades com parâmetros  $\beta$ , e na etapa **Maximization**, o algoritmo encontra parâmetros  $\beta$  melhores que os anteriores através da estimação de máxima verossimilhança. O algoritmo repete as duas etapas até que as distribuições em cada uma delas não mudem mais.

A estimação de máxima verossimilhança é uma maneira de estimar a distribuição de probabilidades conjuntas para um conjunto de observações. Sendo  $\mathbf{x} = [x_1, x_2, \dots, x_n]$  um conjunto de  $n$  objetos independentes pertencentes à mesma distribuição, estimar a máxima verossimilhança para estes conjuntos implica em estimar a distribuição à qual os dados pertencem, e quais são os parâmetros que fazem com que esta distribuição mais se assemelhe à probabilidade dos objetos. Formalmente, procura-se estimar a probabilidade de  $\mathbf{a}$  ocorrer dados a distribuição e seus parâmetros  $\beta$ , e  $P(\mathbf{x}; \beta) = \prod_i^n P(x_i; \beta)$ . A ocorrência de probabilidade igual a zero em um dos componentes de  $\mathbf{a}$  faz com que todo o produto tenha valor zero, e mesmo na ausência desta ocorrência, a multiplicação de diversas

probabilidades com valores muito próximos de zero pode gerar instabilidade na estimação da probabilidade conjunta. Assim, é comum reescrever o como o somatório do logaritmo das probabilidades condicionais, ou seja,  $L(\mathbf{x}; \beta) = \sum_i^n \log(P(x_i; \beta))$ , onde  $L$  vem do inglês, *likelihood*, ou verossimilhança. Deseja-se, portanto, maximizar  $L(\mathbf{x}; \beta)$  ou minimizar  $-L(\mathbf{x}; \beta)$ . Diferente do  $k$ -médiãs, o EM é capaz de gerar grupos com formatos elipsoidais nos quais cada objeto possui uma probabilidade de pertencimento. Assim, um objeto pode em teoria ser igualmente provável nos  $k$  grupos. Ambas as abordagens, no entanto, requerem a determinação do número  $k$  de grupos nos quais os dados devem ser divididos, o que fomenta o interesse no uso de um algoritmo computacionalmente escalável e que determine o número de grupos automaticamente sem limitar o formato que estes grupos podem assumir. É de interesse também o desenvolvimento de algoritmos que não associam *todos* os objetos a algum grupo, ou seja, que sejam capazes de lidar com ruídos nos dados.

A fim de endereçar essas limitações, o DBSCAN (*density-based spatial clustering of applications with noise*, ou agrupamento espacial baseado em densidade de aplicações com ruído) foi publicado por Ester et al. (1996). Seus autores afirmam que a capacidade humana de reconhecer grupos de objetos reside na capacidade de identificar a densidade nestes objetos. Assim, grupos diferentes são separados no espaço por áreas de menor densidade do DBSCAN se baseia na ideia de que o raio de cada objeto deve ter um número mínimo de pontos para que estes objetos sejam considerados como pertencentes ao mesmo grupo. Caso contrário, o ponto é considerado um ruído e não é atribuído a nenhum grupo.

Formalmente, seja  $\mathcal{X}$  o conjunto dos objetos  $n$ -dimensionais de estudo, e seja  $N_{Eps}(\mathbf{a})$  a vizinhança de um ponto  $\mathbf{a} \in \mathcal{X}$  que é dada por  $N_{Eps}(\mathbf{a}) = \{\mathbf{b} \in \mathcal{X} | d(\mathbf{a}, \mathbf{b}) \leq Eps\}$ , sendo que  $d(\mathbf{a}, \mathbf{b})$  é a distância euclidiana entre  $\mathbf{a}$  e  $\mathbf{b}$  e  $Eps$  é o raio. Sempre que o número de pontos de  $N_{Eps}(\mathbf{a})$  for maior ou igual a  $MinPts$ , o ponto  $\mathbf{a}$  é rotulado como *core point*. Caso  $\mathbf{a}$  não seja um *core point* mas pertença à vizinhança de algum *core point*, então  $\mathbf{a}$  é rotulado como *border point*. Caso contrário,  $\mathbf{a}$  é rotulado como *noise point*. Todos os *core points* que pertencerem à vizinhança de outros *core points*, bem como seus respectivos *border points*, fazem parte do mesmo grupo. Os passos abaixo ilustram o funcionamento do algoritmo:

1. Selecione um ponto  $\mathbf{p}$  não visitado qualquer para iniciar o algoritmo.
2. Se  $N_{Eps}(\mathbf{p}) \geq MinPts$ ,  $\mathbf{p}$  representa um grupo, e todos os pontos de  $N_{Eps}(\mathbf{p})$  fazem parte do grupo. A vizinhança  $N_{Eps}(\mathbf{q})$  de qualquer ponto  $\mathbf{q} \in N_{Eps}(\mathbf{p})$  para o qual  $N_{Eps}(\mathbf{q}) \geq MinPts$  também é adicionada ao grupo. Caso  $N_{Eps}(\mathbf{p}) < MinPts$ ,  $\mathbf{p}$  é rotulado como ruído (que pode vir a fazer parte de um grupo posteriormente caso pertença à vizinhança de um outro *core point*).
3. Repita os passos 1 e 2 até que todos os pontos tenham sido rotulados.

O DBSCAN requer a determinação de valores adequados para *MinPts* e *Eps*, e os autores propõem uma heurística capaz de determinar os valores mais adequados através da análise da distribuição da densidade dos objetos, que é calculada com base nas diferentes quantidades de objetos entre um objeto qualquer e seu  $k$ -ésimo vizinho mais próximo. [Arlia e Coppola \(2001\)](#) e [Gaonkar e Sawant \(2013\)](#) apresentam maneiras de selecionar o valor de *Eps* automaticamente, esforço este que é expandido por trabalhos mais abrangentes, que objetivam escolher dinamicamente tanto *Eps* quanto *MinPts*, como visto em [Zhou, Wang e Li \(2012\)](#), [Karami e Johansson \(2014\)](#) e [Lai et al. \(2019\)](#).

Na implementação apresentada por [Ester et al. \(1996\)](#), os autores afirmam que a complexidade de tempo de execução do algoritmo é  $O(n \cdot \log(n))$ . [Gan e Tao \(2015\)](#), no entanto, afirmam que na realidade o tempo de execução do algoritmo é da ordem de  $O(n^2)$ . [Schubert et al. \(2017\)](#) visitam os argumentos apresentados por [Gan e Tao \(2015\)](#) e clarificam que o cálculo da complexidade do algoritmo não é trivial pois fatores como o parâmetro *Eps*, a função de distância e a técnica de indexação e recuperação dos objetos de  $\mathcal{X}$  têm um fator relevante na complexidade, e concluem que o pior caso do algoritmo é de fato  $O(n^2)$  mas que seus experimentos mostraram que o DBSCAN é tão bom quanto os métodos apresentados por [Gan e Tao \(2015\)](#). Já [Kriegel et al. \(2011\)](#) afirmam que o DBSCAN pode ser implementado com complexidade  $O(n \cdot \log(n))$  caso estruturas de indexação apropriadas sejam implementadas.

### 2.2.3 Avaliação de agrupamentos

Considera-se que um processo de agrupamento foi bem sucedido quando os grupos formados por ele fazem com que os objetos dentro de um grupo sejam mais similares entre si do que quando comparados com objetos de outros grupos. No entanto, além das questões técnicas já discutidas ao longo da subseção 2.2.1, o objetivo da tarefa de agrupamento também deve ser levado em conta. A interpretação humana dos resultados é importante, mas ela pode não ser viável se depender da inspeção minuciosa de cada objeto em cada grupo, especialmente no caso de um volume de dados muito grande. A inspeção visual dos resultados auxilia na análise humana, mas é particularmente desafiadora em conjuntos de dados com muitas dimensões. Neste cenário, técnicas de redução de dimensionalidade podem ser úteis, mas elas não endereçam os agravantes derivados do fato de que pessoas diferentes podem interpretar o mesmo resultado de maneiras muito distintas, além da possível ausência de capital humano qualificado para a avaliação subjetiva. Finalmente, a definição dos melhores parâmetros para os algoritmos de agrupamento em um determinado conjunto de dados também requer uma análise objetiva, já que comumente é baseada em uma estratégia de tentativa e erro, como discutido na subseção 2.2.2. Portanto, premissas abrangentes são desejáveis para a produção de resultados comparáveis entre si e agnósticos aos algoritmos usados. Estes desafios fazem com que medidas genéricas e objetivas da



qualidade de agrupamentos sejam empregadas no estudo.

[Rand \(1971\)](#) argumenta que, de maneira geral, um método objetivo de avaliação de agrupamento leva em conta que em um processo de agrupamento cada ponto é atribuído a algum grupo, que os grupos são definidos tanto pelos pontos que os compõem quanto pelos pontos que não fazem parte deles, e que todos os pontos são de igual importância para a determinação dos grupos. Apesar da existência de métodos que atribuem um objeto a mais que um grupo ao mesmo tempo ou que consideram uma parte dos objetos como ruído nos dados e não os atribuem a nenhum grupo, as duas últimas premissas estão presentes em muitos métodos de agrupamento empregados por pesquisadores de diferentes áreas do conhecimento. De maneira geral, existem dois grandes cenários de avaliação: um quando os dados estão rotulados e outro quando não estão ([ARBELAITZ et al., 2013](#)).

Quando existem rótulos para os objetos de estudo o processo de avaliação é também chamado de *validação extrínseca*, e consiste em determinar se os grupos formados são compostos por objetos de mesmo rótulo, cenário no qual também espera-se que o número de grupos encontrados seja próximo ou idêntico ao número de rótulos. Ou seja, dado o conjunto de objetos  $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ , e seja  $\mathcal{U} = \{\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_p\}$  uma partição de  $\mathcal{X}$  em  $p$  grupos gerados através dos rótulos dos elementos de  $\mathcal{X}$ , e  $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_q\}$  uma partição de  $\mathcal{X}$  em  $q$  grupos gerados através de um processo de agrupamento dos objetos, o processo teve êxito completo caso para todo par  $x_i, x_j \in \mathcal{U}_p$  exista  $\mathcal{V}_q$  tal que  $\mathcal{U}_p = \mathcal{V}_q$ . Considerando as definições de  $\mathcal{X}$ ,  $\mathcal{U}$  e  $\mathcal{V}$ , as possibilidades de se permutar  $x_i$  e  $x_j$  nos diferentes grupos destas duas partições são:

- (a)  $x_i$  e  $x_j$  fazem parte do mesmo grupo tanto em  $\mathcal{U}$  quanto em  $\mathcal{V}$ .
- (b)  $x_i$  e  $x_j$  fazem parte de grupos diferentes tanto em  $\mathcal{U}$  quanto em  $\mathcal{V}$ .
- (c)  $x_i$  e  $x_j$  fazem parte de grupos diferentes em  $\mathcal{U}$  e do mesmo grupo em  $\mathcal{V}$ .
- (d)  $x_i$  e  $x_j$  fazem parte do mesmo grupo em  $\mathcal{U}$  e de grupos diferentes em  $\mathcal{V}$ .

Assim, o índice de *Rand*  $R$  que mede a corretude do processo de agrupamento proposto por [Rand \(1971\)](#) se dá por:

$$R = \frac{(a) + (b)}{(a) + (b) + (c) + (d)}$$

O estudo publicado por [Milligan e Cooper \(1986\)](#) investiga diferentes medidas de validação dos agrupamentos encontrados, afirmando que o método índice de *Rand* ajustado como proposto em [Hubert e Arabie \(1985\)](#) é o mais recomendado na pesquisa científica. O índice ajustado deriva da tabela de contingência (Tabela 2) de  $\mathcal{U}$  e  $\mathcal{V}$  cujo valor  $n_{pq}$  de intersecção representa o número de objetos em comum entre  $\mathcal{U}$  e  $\mathcal{V}$ , ou seja,  $n_{pq} = |\mathcal{U}_p \cap \mathcal{V}_q|$ :



Tabela 2 – Tabela de contingência de  $\mathcal{U}$  e  $\mathcal{V}$ 

	$\mathcal{V}_1$	$\mathcal{V}_2$	$\dots$	$\mathcal{V}_q$	$\Sigma$
$\mathcal{U}_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1q}$	$a_1$
$\mathcal{U}_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2q}$	$a_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathcal{U}_p$	$n_{p1}$	$n_{p2}$	$\dots$	$n_{pq}$	$a_q$
$\Sigma$	$b_1$	$b_2$	$\dots$	$b_q$	

A fórmula do índice *Rand* ajustado *ARI* (do inglês: *Adjusted Rand Index*) é dada por:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}$$

O *ARI* retorna um resultado próximo de zero quando a similaridade entre os dois grupos é próxima daquela obtida caso os objetos fossem aleatoriamente dispostos nos grupos de  $\mathcal{U}$  e  $\mathcal{V}$ . Caso ambas as partições *concordem* completamente, ou seja, apenas a permutação ( $a$ ) ocorra, então o valor de *ARI* será 1.

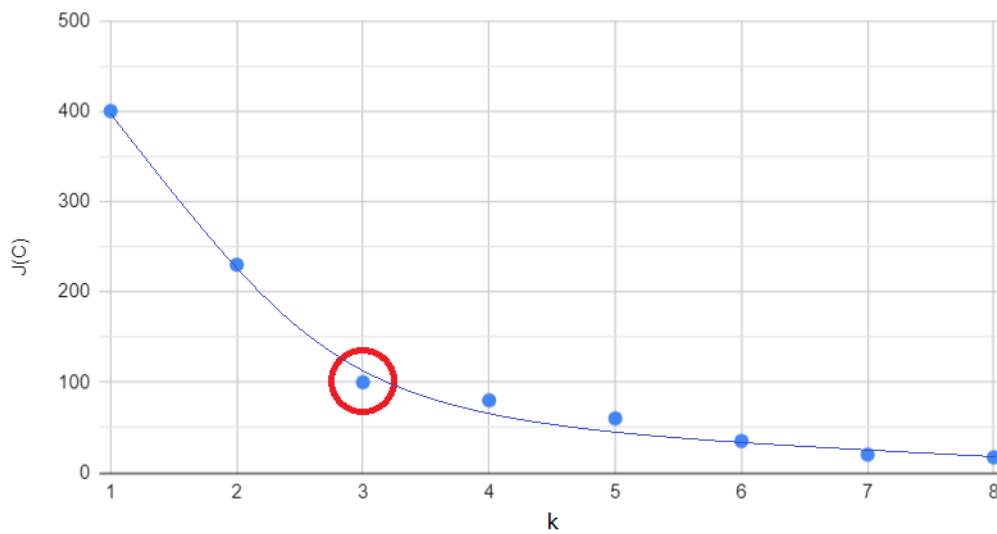
Quando não existem rótulos para os objetos de estudo, duas características dos grupos encontrados são levadas em consideração na avaliação da qualidade do agrupamento (RENDÓN et al., 2011): a compacidade e a separabilidade dos grupos. A compacidade mede quão similares os objetos de um mesmo grupo são entre si, portanto espera-se que quanto mais compacto um grupo, menor a variância das características dos objetos que o compõem. Já a separabilidade mede a diferença entre os grupos, e espera-se que em um agrupamento com alta separabilidade, a distância entre grupos seja maior do que aquela observada em agrupamentos com baixa separabilidade. Maneiras de medir distâncias entre grupos foram discutidas em na subseção 2.2.2.

Quando não se tem validação externa, um dos grandes desafios na interpretação de uma boa medida de qualidade de agrupamento está nas conclusões que se pode tirar do número obtido no processo de avaliação. Para comparar o resultado relativo de diferentes abordagens, é desejável que as métricas de avaliação possuam limites superiores e inferiores que representem sucesso e fracasso, de maneira que se possa avaliar a extensão da qualidade obtida para que a escolha do método mais eficiente seja pautada em um critério objetivo.

Tomemos como exemplo a função objetivo do método  $k$ -médias discutida na subseção 2.2.2. O objetivo do algoritmo é minimizar  $J(C)$ , que representa a média das distâncias de cada objeto de um grupo para o centroide deste grupo. Para  $k = |C| = n$ , ou seja, quando o número de grupos for igual ao número de objetos no processo de agrupamento, a distância de cada objeto para seu centroide é 0, pois cada objeto é o

centroide de seu grupo. Conforme se diminui o número de grupos, espera-se portanto que o valor de  $J(C)$  aumente.  $J(C)$  deve ter seu valor máximo quando  $k = |C| = 1$ , mas esse valor pode ser maior ou menor a depender da característica dos objetos estudados, o que faz com que seja difícil usá-lo para comparar a eficiência de diferentes algoritmos de agrupamento. Apesar disso, este método é útil na determinação do parâmetro  $k$  do algoritmo  $k$ -médias, pois quando se observa pouca variação em  $J(C)$  ao aumentar  $k$ , pode-se concluir que não há melhora significativa na compactidade dos grupos encontrados. A definição do que é uma melhora significativa, no entanto, ainda tem um aspecto de subjetividade, portanto é comum a inspeção gráfica como mostrado na Figura 4, onde os valores  $k > 3$  fornecem reduções marginais decrescentes de  $J(C)$ . Assim, este parece ser um número de grupos adequado quando o problema é analisado desta maneira. Este método de avaliação é chamado de *cotovelo* (do inglês: *elbow*), que recebe este nome pois escolhe-se o valor de  $k$  que forma um cotovelo no gráfico da variação de  $J(C)$  explicada como uma função dos valores de  $k$ .

Figura 4 – Escolha do número de grupos pela análise do cotovelo



Ao considerar as distâncias euclidianas como critério de similaridade entre objetos,  $J(C)$  avalia a compactidade dos grupos assumindo que eles possuem um formato circular ou elipsoidal, conforme discutido na subseção 2.2.1. Esta restrição pode levar a resultados pouco confiáveis na avaliação de processos de agrupamento executados por algoritmos que não se baseiam em centroides, ou em bases de dados cuja geometria dos grupos não é compatível com a geometria de  $J(C)$ . Idealmente, um método de avaliação ideal é capaz de verificar se a estrutura de agrupamento encontrada é de fato derivada da estrutura dos objetos estudados ou apenas um agrupamento artificial.

A fim de minimizar este problema, [Rousseeuw \(1987\)](#) propõem de maneira muito didática o popular método da silhueta. Intuitivamente, a ideia por trás do algoritmo é a de

apresentar, com apenas uma única estatística, o quão bem agrupado está cada objeto de estudo. O método consiste em verificar a diferença entre a distância média de um objeto de um grupo para todos os outros do mesmo grupo e a menor distância deste objeto para todos os grupos. Este valor é então dividido pela maior distância média deste objeto e todos os objetos de outros grupos. Formalmente, seja  $\mathcal{U} = \{\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_k\}$  uma partição de  $k$  grupos de  $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ . Se um processo de agrupamento atribui  $x_i$  ao grupo  $\mathcal{U}_i$ , e existe mais que um objeto em  $\mathcal{X}$ , então dizemos que  $a(x_i)$  é a distância média de  $x_i$  para todos os outros objetos do grupo  $\mathcal{U}_i$  ao qual  $x_i$  pertence. Para qualquer outro grupo  $\mathcal{U}_j \in \mathcal{U}$  para  $j \neq i$ ,  $d(x_i, \mathcal{U}_j)$  é a distância média entre o objeto  $x_i$  e todos os objetos do grupo  $\mathcal{U}_j$ . Sendo  $b(x_i) = \min\{d(x_i, \mathcal{U}_j) | j = 1..k, j \neq i\}$ , então a silhueta  $s(x_i)$  do objeto  $x_i$  é dada por:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}$$

Dado que  $-1 \leq s(x_i) \leq 1$ , o valor máximo e positivo de  $s(x_i)$  acontece quando  $b(x_i) > a(x_i) = 0$ , ou seja,  $x_i$  é idêntico aos objetos do seu grupo, e diferente dos objetos de outros grupos. Inversamente,  $s(x_i) = -1$  quando  $x_i$  é idêntico aos objetos dos outros grupos, e diferente dos objetos do grupo ao qual pertence. Assim, quanto mais próximo de 1 for o valor de  $s(x_i)$ , melhor o processo de agrupamento parece ser adequado à estrutura presente nos dados. Para que  $s(x_i)$  seja zero,  $a(x_i)$  e  $b(x_i)$  devem ter valores próximos, cenário no qual a atribuição mais correta para  $x_i$  é incerta.  $s(x_i)$  apresenta uma descontinuidade caso haja apenas um objeto em  $\mathcal{X}$ , pois neste caso tanto  $a(x_i)$  quanto  $b(x_i)$ , componentes do seu denominador, valem 0. Por isso, convencionou-se que para este caso  $s(i) = 0$ .

Diferentes estudos concluem que a silhueta é uma medida de validação interna adequada para avaliar os processos de agrupamento, como pode ser visto em [Xiong e Li \(2013\)](#), [Moulavi et al. \(2014\)](#), [Rendón et al. \(2011\)](#) e [Tomasini et al. \(2016\)](#), pois além de apresentar resultados melhores, ela é capaz de fornecer uma maneira de avaliar tanto individualmente os objetos quanto grandes grupos de objetos através de estatísticas como média, mediana e moda de seus valores. Apesar disso, o método requer que a distância entre todos os objetos seja calculada, o que faz com que seu tempo de execução seja da ordem de  $O(n^2)$ . Isso se dá porque a medida de distância entre o objeto e um grupo é equivalente à *ligação média*, um dos casos do método de mínima variância de Ward ([JR, 1963](#)), como apresentado na subseção 2.2.2. [Hruschka, Castro e Campello \(2004\)](#) propõem a silhueta simplificada como uma das medidas de validação internas usada em seu trabalho, medida essa que altera  $d(x_i, \mathcal{U}_j)$  de maneira que seja calculada como a distância de  $x_i$  até o centroide de  $\mathcal{U}_j$ , e [Wang et al. \(2017\)](#) concluem que a versão simplificada apresenta resultados competitivos com performance computacional superior.

Segundo [Rendón et al. \(2011\)](#), as medidas de validação interna produzem os melhores resultados em avaliar se o agrupamento foi bem sucedido em capturar a estrutura

dos dados. No entanto, quando se quer descobrir se um processo de agrupamento pode criar grupos que estejam em conformidade com um determinado conjunto de rótulos, a melhor medida interna pode não representar a melhor solução para o problema, o que faz necessária a análise tanto de medidas internas quanto externas, ou avaliações subjetivas por meio de um especialista de domínio.

## 2.3 Direito e as ciências exatas

A definição de *jurimetria* que, no melhor do nosso conhecimento, é a pioneira, define o termo como uma área do conhecimento que estuda a aplicação de teorias e ferramentas de Estatística e Computação em prol da previsibilidade do Direito. Esta definição do termo remonta ao final dos anos 40 e início dos anos 50, quando [Loevinger \(1948\)](#) publicou seu longo artigo no qual discorre sobre seu assombro pela dificuldade de acesso ao Direito e pelo debate acadêmico sobre a definição do que é a lei.

Em sua leitura sobre a cronologia deste debate, [Loevinger \(1948\)](#) relembra os primórdios da justiça instanciados no código de Hamurabi, passa pela lógica aristotélica e pelos trabalhos filosóficos sobre a origem da lei desenvolvidos ao longo da história, principalmente na Europa ocidental e na América do Norte. Loevinger conclui que apesar da sua extensão, pouca foi a contribuição deste debate para o transporte de informação pertinente. Afirma ainda que as ciências sociais tendiam a usar de fatos para justificar um ponto de vista previamente estabelecido, o que segundo ele, iria de encontro àquela prática da verdadeira ciência, na qual se usa das evidências, de dados e fatos para só então depois se formar uma teoria sobre o assunto estudado. E é com base neste suposto *modus operandi* das ciências sociais, e em especial do Direito, que Loevinger faz a conexão desta disciplina com os métodos quantitativos supracitados como meios para mudar o estudo da disciplina.

O termo *jurisprudência* refere-se ao conjunto de decisões jurídicas tomadas sobre um determinado assunto ([FRANÇA, 1971](#)). Loevinger, no entanto, define jurisprudência como um exercício de mera especulação sobre o funcionamento das leis, e argumenta que para que o ser humano dê o próximo passo no sentido do progresso, a prática da lei precisa deixar de ser feita através da jurisprudência e passar a ser feita através da jurimetria. Propõe, então, que os métodos estatísticos e matemáticos sejam empregados em questões que vão do comportamento das testemunhas, juízes e legisladores ao estudo da linguagem e da comunicação jurídicas. O termo ganhou relevância acadêmica e, em 1959, o periódico *Modern Uses of Logic in Law* passou a ser publicado pela *American Bar Association*. Em 1966, mudou de nome para *Jurimetrics Journal* e, em 1978, passou a se chamar *Jurimetrics: The Journal of Law, Science, & Technology*, nome da publicação até hoje.

Desde a publicação de [Loevinger \(1948\)](#), diversos trabalhos científicos publicados estudam questões relacionadas ao Direito empregando métodos quantitativos sem necessariamente citar o termo jurimetria ([ASH; CHEN, 2018](#); [MANDAL et al., 2017](#); [SUGATHADASA et al., 2018](#)). Mesmo a Conferência Internacional de Inteligência Artificial e Direito (*ICAAIL, do inglês: International Conference on Artificial Intelligence and Law*) não cita o termo no seu texto de apresentação. Assim, apesar da definição proposta por Loevinger ser adequada ao se referir a estes trabalhos, é possível que a comunidade científica das ciências exatas não tenha convergido para o uso do termo, o que pode dificultar a recuperação de pesquisas dessa natureza.

*Legal informatics* é outro termo relevante para o estudo da aplicação de métodos quantitativos ao Direito. [Biasiotti et al. \(2008\)](#) definem o termo como a disciplina que lida com o uso das tecnologias da informação e comunicação (TIC) para processar informações e suportar atividades no domínio jurídico, como a criação, cognição e aplicação da lei. Esta definição mais ampla abrange qualquer tipo de tecnologia da informação, o que coloca os domínios científicos aos quais se associa de maneira mais fácil de se perceber, pois no contexto de *legal informatics*, qualquer sistema computadorizado que lida com dados jurídicos atende à definição. É de particular interesse deste trabalho, no entanto, os sistemas que fazem uso de técnicas de aprendizado de máquina para a organização de grandes volumes de documentos jurídicos.

Em um estudo que visa identificar atividades jurídicas que poderiam se beneficiar de sistemas inteligentes, e considerando o estado da arte da pesquisa de inteligência artificial, [Surden \(2014\)](#) argumenta que dentre as atividades legais, aquelas que envolvem recuperação da informação, predição de provável resultado de um processo legal, busca por informação implícita potencialmente útil em aberturas ou defesas de ações judiciais e classificação e agrupamento de documentos são bons candidatos para a aplicação de sistemas inteligentes. [Surden \(2014\)](#) conclui ainda que apesar de muitas das atividades desempenhadas por um profissional de direito requererem capacidades cognitivas que os sistemas inteligentes atuais ainda não conseguem reproduzir, as técnicas de aprendizado de máquina podem já ser capazes de produzir resultados úteis no domínio.

[Surden \(2014\)](#) elucida para a relevância, dentre outros tópicos, da organização automática de documentos jurídicos, que é precisamente o objetivo deste trabalho. Como pode ser visto no capítulo 3, diversos trabalhos se propõem a endereçar problemas que encontram equivalência nas oportunidades apontadas por Surden, e dentre estes trabalhos encontram-se também trabalhos brasileiros que lidam com as questões particulares ao sistema judiciário brasileiro.

Segundo [Zabala e Silveira \(2014\)](#), poucos são os desdobramentos científicos relevantes de publicações tupiniquins sobre o tema aqui discutido, ainda que o interesse no assunto seja crescente. Usando o termo *jurimetria*, afirmam:

Uma das mais destacadas atuações da jurimetria é a análise de informações organizadas em bancos de dados públicos, fundamentais para o entendimento da situação socioeconômica vigente. A organização e análise de dados proporcionam um ambiente favorável para a produção de leis coerentes, criando um alicerce comum para discussões políticas.

Para exemplificar a afirmação citada, [Zabala e Silveira \(2014\)](#) apresentam casos de trabalhos de análise de dados jurídicos brasileiros que tiveram impactos legais, mas apontam também que o debate político antecedeu a análise das informações sobre os temas discutidos, fato este que pode ser um empecilho para a efetiva tomada de decisão com base em dados pertinentes. Sugerem ainda que a *jurimetria* pode ser olhada de três prismas: elaboração legislativa e gestão pública, decisão judicial e instrução probatória. Esta sugestão dialoga com as oportunidades de aplicação de métodos quantitativos no direito propostas por [Surden \(2014\)](#), o que leva a crer que esta divisão pode ser também expandida para o domínio mais abrangente de *legal informatics*, previamente discutido nesta subseção.

A organização de documentos se enquadraria então no prisma de elaboração legislativa e gestão pública e na atividade jurídica descrita por Surden como recuperação da informação, mas entendemos que os resultados da pesquisa que visa encontrar representações e métricas de similaridades adequadas ao domínio jurídico podem também contribuir com as atividades envolvidas na predição de resultado de um processo legal, pois a escolha do método de representação também influencia no desempenho de tarefas de classificação automática de documentos.

Trabalhos como [Oliveira \(2016\)](#), [Castro \(2017\)](#), [Ravagnani \(2017\)](#) e [Nunes \(2020\)](#) são bons exemplos de como o tema vem ganhando relevância no Brasil no domínio do Direito, disciplina essa que pode se beneficiar de avanços computacionais na representação e organização de um dos seus materiais de trabalho: os documentos que contém as informações relevantes. Eles mostram como problemas importantes do Direito podem tirar proveito de ferramentas para que os praticantes da disciplina lidem cada vez mais com os problemas pertinentes ao sistema judiciário brasileiro, abstraindo os desafios relacionados às TICs de maneira a contribuir com a execução das suas atividades. Do ponto de vista das ciências exatas aplicadas, os processos legais são uma excelente fonte de objetos de estudo, permitindo a criação de bases de dados volumosas e com complexidade variável. Parte significativa da informação jurídica disponível já possui assuntos, matérias, comarcas, magistrados e as partes envolvidas em um processo, todas disponibilizadas como dados estruturados. Além disso, os teores dos documentos possuem informação textual com redação de alto rigor qualitativo e frequentemente dotada de pouco ou nenhum erro ortográfico. Estes fatores nos levam a concluir que a exploração destes dados e suas aplicações têm um potencial acadêmico ainda pouco explorado, porém muito promissor.

### 3 Trabalhos Relacionados

Este capítulo discute tanto trabalhos que estudam aplicações de aprendizado de máquina no domínio jurídico quanto trabalhos que possuem documentos longos escritos em língua portuguesa como fonte de dados. Apresenta ainda trabalhos que discutem técnicas de agrupamento como maneira de se estabelecer classes para os documentos agrupados.

Furquim e Lima (2012) analisaram o agrupamento de documentos jurídicos por meio de uma versão modificada do algoritmo de agrupamento proposto em Aggarwal, Gates e Yu (2004). As avaliações feitas por Furquim e Lima (2012) consistem de um conjunto de 1.192 documentos datados entre 2006 e 2009, representados através de *bag of terms and law references*, uma extensão da representação *bag of words* na qual o vocabulário é composto por 13.354 expressões extraídas de dois tesouros jurídicos: o Vocabulário Controlado Básico<sup>1</sup> e o Tesouro da Justiça Federal<sup>2</sup>. Além das expressões, as referências às leis também são consideradas parte do vocabulário. A contagem das expressões dos tesouros e as referências às leis são ponderadas com base na hierarquia em que aparecem no tesouro e com base no tipo da lei referenciada. Em seguida, os documentos foram agrupados e os centroides dos grupos foram usados para categorizar 105 documentos. Os autores concluem que as referências às leis melhoram o desempenho da atividade de agrupamento.

O trabalho realizado por Castro et al. (2019) aborda o tema de extração de entidades nomeadas no domínio jurídico usando técnicas de aprendizado profundo cuja representação interna das palavras é feita por meio de *word embeddings*. Para tanto, foram treinados modelos baseados na arquitetura *ELMo* (PETERS et al., 2018), tanto no domínio jurídico quanto no domínio geral da língua portuguesa. O corpus do domínio jurídico foi composto por 1.305 documentos obtidos no site do Processo Judicial Eletrônico e anotados por um estudante de Direito, enquanto que o corpus geral consiste em extrações da Wikipedia e do *brwac* (BOOS et al., 2014), um compilado de extrações textuais de páginas *web* com domínio .br. Castro et al. (2019) comparam o desempenho de extração de entidades em duas representações: tradicional e *embeddings*, concluindo que o uso em conjunto não trouxe melhora quando comparado ao uso apenas da técnica tradicional no domínio geral da língua portuguesa.

Já no contexto de inteligência artificial aplicada ao setor jurídico em idiomas estrangeiros, trabalhos como Branting et al. (2019) e Castano et al. (2019) mostram como algoritmos de aprendizado de máquina contribuem para automação da explicação de decisões jurídicas. Enquanto que o primeiro apresenta o uso de redes neurais com mecanis-

<sup>1</sup> [http://biblioteca2.senado.gov.br:8991/F/?func=find-b-0local\\_base=sen10](http://biblioteca2.senado.gov.br:8991/F/?func=find-b-0local_base=sen10)

<sup>2</sup> <https://www.cjf.jus.br/cjf/biblioteca/tesouro-juridico>



mos de atenção capazes de encontrar termos diretamente relacionados à jurisprudência, o segundo mostra uma abordagem baseada em ontologias que é capaz de relacionar o aparecimento de termos em decisões processuais com termos similares documentados em tesouros desenvolvidos por especialistas no domínio, possibilitando assim a representação automática do conhecimento jurídico contido nestes documentos.

Em [Hartmann et al. \(2017\)](#), diversos algoritmos de representação numérica de palavras são treinados e comparados frente à sua utilidade em tarefas como *Part of Speech Tagging* e analogias sintáticas e semânticas. Os algoritmos foram treinados em um corpus composto por mais de um bilhão de palavras da língua portuguesa, coletado de fontes diversas como Wikipédia, Revista Mundo Estranho, textos científicos divulgados pela FAPESP, entre outros. O estudo sugere nas conclusões que não é apropriado usar analogias (semânticas ou sintáticas) para avaliar a qualidade das representações de palavras, mas que as representações geradas podem ser úteis em diversas atividades no domínio de processamento de linguagem natural.

Em [Magalhães e Souza \(2019\)](#), 50 notícias pertencentes a quatro categorias foram coletadas na internet, e os algoritmos de agrupamento hierárquico, *k-means* e *affinity propagation* foram usados para agrupá-las. A representação escolhida para os documentos foi a mesma para todos os algoritmos: *bag of words* criado depois da remoção de *stopwords* e de *stemming* nas palavras restantes. A métrica de avaliação de desempenho foi o número de vezes que cada algoritmo agrupou notícias da mesma classe no mesmo grupo, critério no qual o algoritmo *k-means* apresentou melhor desempenho.

[Faraco et al. \(2018\)](#) usa 1.849 teses e dissertações da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior como fonte de dados textuais para a tarefa de agrupamento. Os documentos são representados como vetores *tf-idf* e agrupados com a técnica *k-means*. A qualidade dos agrupamentos foi avaliada qualitativamente e por amostragem, e então os nomes dos grupos foram concebidos com base no teor dos documentos que continham. O trabalho conclui que a aplicação de técnicas de agrupamento em documentos de texto apresenta a qualidade de encontrar agrupamentos não óbvios.

[Afonso \(2016\)](#) apresenta um sistema de agrupamento de textos em língua portuguesa composto de duas fases: indexação e agrupamento. Na fase de indexação aplica-se regras que visam extrair sintagmas nominais com o objetivo de representar o conteúdo dos documentos, como feito em [Maia \(2009\)](#). Para os termos extraídos nesta representação são atribuídos pesos *tf-idf*, que não são usados no contexto de vetores *bag of words* pois estes possuem alta dimensionalidade. Em vez disso, os autores propõem representar cada documento apenas com os termos de indexação extraídos na etapa anterior, e a justificativa para fazê-lo é o grande esforço computacional necessário em caso se opte por usar todas as ocorrências de palavras do corpus na representação individual dos documentos. Na fase de agrupamento, os vetores individuais são tratados como entrada de um algoritmo



evolutivo que seleciona automaticamente o número de grupos ideal para os documentos do corpus, comparando a similaridade dos vetores que representam cada documento. Os dados usados no experimento são quatro conjuntos de artigos científicos que variam entre 60 e 120 documentos pertencentes a seis categorias no corpus de 120 elementos, cinco categorias nos corpus de 100 elementos e a 3 categorias no corpus de 60 elementos. A avaliação foi feita comparando-se o número de documentos corretamente classificados em cada grupo, sendo que a predominância de documentos em um dado grupo define o rótulo considerado correto neste grupo. O algoritmo proposto por [Afonso \(2016\)](#) obteve desempenho superior a no mínimo 71,7% (no maior corpus) e no máximo 97,6% (no menor corpus).

Em [Aggarwal, Gates e Yu \(2004\)](#) argumenta-se que agrupamentos de texto podem ser usados para simplificar a classificação de grandes conjuntos de documentos. Propõe-se o agrupamento como maneira de criação de rótulos para algoritmos de classificação, o que configura a técnica de aprendizado semi-supervisionado uma vez que a qualidade do agrupamento é medida com documentos previamente rotulados e que são incorporados às próximas etapas de treinamento, de maneira iterativa.

Neste trabalho, propomos o uso de técnicas de representação de documentos como ferramentas para estudar o problema de agrupamento de decisões judiciais frente a seu assunto. Além das representações, serão avaliados também diferentes algoritmos de agrupamento de dados, sendo que o material do estudo é um corpus com cerca de 3 milhões de decisões jurídicas.



## 4 Metodologia

Este capítulo detalha o corpus usado neste projeto e apresenta em detalhes o plano de trabalho.

### 4.1 Corpus

O Corpus é composto por 3.167.830 decisões de primeira instância extraídas do sistema *e-SAJ*<sup>1</sup>. Para realizar a extração do corpus, um coletor dos dados foi desenvolvido. A partir dele, foram estruturadas as decisões judiciais de primeira instância disponíveis e seus respectivos rótulos. Cada documento é composto no corpus seguindo o seguinte esquema:

- *id*: o número único que identifica o processo
- *Classe*: classe processual definida pelo Conselho Nacional de Justiça. A Tabela 3 mostra as dez primeiras classes ordenadas pela frequência em que aparecem no corpus.
- *Assunto*: o assunto judicial relaciona o processo quanto à matéria, e as classificações dos assuntos possíveis são feitas pelo Conselho Nacional de Justiça e podem verificadas no site do Conselho<sup>2</sup>. A distribuição dos dez primeiros assuntos pode ser vista na Tabela 4.
- *Magistrado*: o juiz responsável pela sentença. O corpus possui 1801 magistrados, sendo que aqueles que mais apareceram nos documentos podem ser vistos na Tabela 5.
- *Comarca*: a unidade legislativa onde foi tramitado o processo. Pode corresponder a uma região administrativa composta de mais de um município onde trabalha, juízes de primeiro grau. 318 comarcas figuram no período de extração do corpus, e as dez mais frequentes podem ser vistas na Tabela 6.
- *Foro*: usado para determinar qual lugar tem o poder de julgar. Os mais frequentes podem ser vistos na 7.
- *Vara*: a vara onde o processo tramitou. Corresponde a um tribunal ou ao local onde trabalha o juiz. Ao todo, o corpus possui 347 varas, e a Tabela 8 apresenta as varas com mais processos.

---

<sup>1</sup> <https://bit.ly/36SPXEw>

<sup>2</sup> <http://www.cnj.jus.br/sgt>

- *Data\_disp*: a data na qual o documento foi disponibilizado no e-Saj. As datas de disponibilização extraídas vão de 14/10/2019 a 27/10/2019. A Tabela 9 apresenta a quantidade de documentos em cada uma delas. Existe ainda uma data de disponibilização inválida, 01/10/3199, na qual 4.503 documentos foram cadastrados. Este volume representa 0,1421% do total.
- *Teor*: conteúdo da sentença

Tabela 3 – As 10 classes mais frequentes no corpus

Classe	Quantidade	%
execução fiscal	740.880	23,39
procedimento comum cível	640.329	20,21
procedimento do juizado especial cível	472.625	14,92
cumprimento de sentença	289.291	9,13
execução de título extrajudicial	136.844	4,32
ação penal - procedimento ordinário	118.904	3,75
requisição de pequeno valor	63.012	1,99
reclamação pré-processual	57.231	1,81
cumprimento de sentença contra a fazenda pública	52.721	1,66
monitória	50.532	1,60

Tabela 4 – Os 10 assuntos mais frequentes no corpus

Assunto	Quantidade	%
dívida ativa	245.109	7,74
iptu/ imposto predial e territorial urbano	238.140	7,52
assunto não informado.	143.131	4,52
indenização por dano moral	127.705	4,03
indenização por dano material	96.678	3,05
obrigação de fazer / não fazer	75.831	2,39
prestação de serviços	72.340	2,28
contratos bancários	51.250	1,62
acidente de trânsito	49.110	1,55
rescisão do contrato e devolução do dinheiro	48.006	1,52

Tabela 5 – Os 10 magistrados mais frequentes no corpus (continua)

Magistrado	Quantidade	%
fabio francisco taborda	79.574	2,51

Tabela 5 – *Os 10 magistrados mais frequentes no corpus (conclusão)*

<b>Magistrado</b>	<b>Quantidade</b>	<b>%</b>
anderson josé borges da mota	34.630	1,09
patricia soares de albuquerque	27.598	0,87
fernanda silva goncalves	26.977	0,85
josé vitor teixeira de freitas	22.369	0,71
josé daniel dinis gonçalves	20.393	0,64
ana cecilia marques faria	17.328	0,55
enoque cartaxo de souza	16.796	0,53
joão baptista galhardo júnior	15.357	0,48
ariana consani brejão degregório gerônimo	14.586	0,46

Tabela 6 – *As 10 comarcas mais frequentes no corpus*

<b>Comarca</b>	<b>Quantidade</b>	<b>%</b>
são paulo	575.794	18,18
campinas	89.025	2,81
são vicente	88.013	2,78
santos	79.404	2,51
guarulhos	54.509	1,72
araçatuba	52.205	1,65
ribeirão preto	50.163	1,58
presidente prudente	41.437	1,31
são josé do rio preto	40.986	1,29
santo andré	39.802	1,26

Tabela 7 – *Os 10 foros mais frequentes no corpus*

<b>Foro</b>	<b>Quantidade</b>	<b>%</b>
foro central cível	151.868	4,79
foro de são vicente	88.013	2,78
foro de santos	79.387	2,51
foro de campinas	75.938	2,40
foro regional ii - santo amaro	56.902	1,80
foro de guarulhos	54.509	1,72
foro central - fazenda pública/acidentes	54.311	1,71
foro de araçatuba	50.185	1,58
foro de ribeirão preto	49.789	1,57
foro de são josé do rio preto	40.986	1,29

Tabela 8 – *As 10 varas mais frequentes no corpus*

Vara	Quantidade	%
vara da fazenda pública	222.502	7,02
juizado especial cível e criminal	167.072	5,27
saf - serviço de anexo fiscal	165.261	5,22
2ª vara cível	162.402	5,13
vara do juizado especial cível e criminal	158.847	5,01
1ª vara cível	153.761	4,85
1ª vara	152.702	4,82
3ª vara cível	148.503	4,69
2ª vara	142.464	4,50
vara única	133.824	4,22

Tabela 9 – *Quantidade de documentos de acordo com as datas de disponibilização.*

Data	Quantidade	%	Data	Quantidade	%
14/10/2019	30	0,0009	21/10/2019	12.750	0,4025
15/10/2019	110.270	3,4809	22/10/2019	10.674	0,3369
16/10/2019	1.163.250	36,7207	23/10/2019	54.470	1,7195
17/10/2019	973.451	30,7293	24/10/2019	203.514	6,4244
18/10/2019	469.048	14,8066	25/10/2019	98.919	3,1226
19/10/2019	3.226	0,1018	26/10/2019	4.714	0,1488
20/10/2019	511	0,0161	27/10/2019	58.500	1,8467

As classes e assuntos compõem, ainda, uma hierarquia na qual um assunto pode aparecer em mais que uma classe. A Tabela 10 mostra em quantas classes diferentes aparecem cada um dos 20 assuntos mais comuns no corpus:

Tabela 10 – *Quantidade de classes por assunto (continua)*

Assunto	Quantidade de classes
assunto não informado.	25
obrigações	18
obrigação de fazer / não fazer	16
locação de imóvel	16
liquidação / cumprimento / execução	14
prestação de serviços	14
inadimplemento	13
tratamento médico-hospitalar	12
inventário e partilha	12

Tabela 10 – Quantidade de classes por assunto (conclusão)

Assunto	Quantidade de classes
contratos bancários	12
indenização por dano moral	12
liminar	11
espécies de contratos	11
valor da execução / cálculo / atualização	11
antecipação de tutela / tutela específica	11
compra e venda	11
fornecimento de medicamentos	11
indenização por dano material	10
pagamento	10
tráfico de drogas e condutas afins	10

## 4.2 Tesouro Jurídico do Supremo Tribunal Federal

O Tesouro do Supremo Tribunal Federal (TSTF) é uma ferramenta de controle para 15.795 terminologias usadas no setor jurídico. Segundo o site do Supremo Tribunal Federal<sup>3</sup>, os elementos contidos no TSTF são:

**Descritor:** Termo escolhido para representar um conceito no Tesouro e que será utilizado na indexação e na recuperação de determinado assunto. Quando houver outros termos que representem o mesmo conceito, antes do termo descritor, constará a sigla **USE**.

**Não-descritor:** Termo que, embora descreva o mesmo conceito que o descritor, não é autorizado na indexação, para evitar a proliferação de sinônimos. Antes de cada não-descritor, constará a sigla **UP**.

**Nota explicativa (NE):** Fornece uma definição do termo ou uma orientação sobre como utilizá-lo em uma indexação.

**Termo genérico (TG):** Indica que há relação hierárquica entre termos com relação gênero-espécie e que este descritor representa o termo com o conceito mais abrangente.

**Termo específico (TE):** Indica os termos subordinados ao termo genérico na cadeia hierárquica.

**Termo relacionado (TR):** Indica relação entre termos que não formam uma hierarquia (gênero-espécie), mas que são associados mentalmente, de forma automática. Servem para orientar o indexador quanto às possibilidades de encadeamento de descritores e para sugerir ao usuário formas de limitar ou expandir uma pesquisa.

**Categoria (CAT):** O TSTF é organizado em três grandes grupos de categorias: **Ramos do Direito** (direito constitucional, direito civil, etc.), **Especificadores** (agrupam termos que restringem o conceito de um

<sup>3</sup> <<https://bit.ly/2XefZbI>>

descriptor, revelando a situação concreta em que o descriptor foi empregado) e **Identificadores** (agrupam nomes de pessoas, instituições, países, estados-membros, programas, etc.)

Para facilitar a manipulação do TSTF, foi desenvolvido um extrator e os dados que ele contém foram armazenados em formato de tabela. A Tabela 11 apresenta como a informação do Tesouro está representada.

### 4.3 Plano de trabalho

O teor do documento e o assunto são os principais objetos de interesse deste trabalho. A hipótese de pesquisa é investigar se no teor existem características inerentes à redação dos documentos que permitem identificar seu assunto através de aprendizado não supervisionado, e se uma representação criada com base no domínio jurídico da língua portuguesa é melhor que uma representação gerada no domínio geral. Pretende-se ainda comparar a eficiência entre a representação por agregação de vetores de palavras com a representação de vetores de documentos, ambos gerados no domínio jurídico, na tarefa de agrupamento. As atividades de pré-processamento do teor dos documentos estão descritas no Apêndice A - Atividades de pré-processamento.

Após o pré-processamento dos documentos, os teores serão usados para treinar um modelo Word2Vec (MIKOLOV et al., 2013) com 100 dimensões cuja implementação está disponível na biblioteca de código aberto Gensim (ŘEHŮŘEK; SOJKA, 2010). Os termos do TSTF que são compostos por apenas uma palavra e que também aparecem no corpus serão analisados no contexto do modelo criado. Esta análise consiste em observar, segundo a distância de cosseno (discutida na subseção 2.2.1 do capítulo 2), qual o *ranking* entre o termo analisado e seus respectivos termos relacionados, termos específicos, termos genéricos, termos descritores e termos não-descritores, segundo o TSTF. Nesta etapa do estudo, objetiva-se verificar se o conhecimento contido no Tesouro está também representado no modelo. Quanto mais próximos os termos analisados estiverem de seus termos associados, melhor o modelo representa o conhecimento que está contido no TSTF. Em outras palavras, quanto mais próximo de 1 o *ranking* dos termos associados em relação ao termo analisado, melhor o modelo representa o conhecimento do TSTF.

A análise descrita acima será repetida também para o modelo GloVe (PENNINGTON; SOCHER; MANNING, 2014), cuja implementação é disponibilizada pelo grupo de pesquisa em processamento de linguagem natural da Universidade de Stanford<sup>4</sup>. Então o experimento será repetido para o modelo FastText (BOJANOWSKI et al., 2016), e os três modelos serão comparados frente a sua capacidade de representar o conhecimento contido no TSTF. A distribuição dos *rankings* produzidos pelos modelos será estudada, e

<sup>4</sup> <<https://stanford.io/3kwj6LV>>



Tabela 11 – Exemplo de termos do TSTF em formato tabular

termo	use	up	te	tr	tg	cat	nota
A MAIORI AD MINUS	/QUEM PODE O MAIS PODE O MENOS/					/DPC DIREITO PROCESSUAL CIVIL/	
À PRIMEIRA VISTA	/ICTU OCULI/					/ASP ASPEC- TOS/	
A VOZ DO BRASIL				/EMISSORA DE RÁDIO/		/INS INSTITUI- ÇÕES/	
ABALROAÇÃO	/ABALROAMENTO/					/DIC DIREITO CIVIL/	
ABANDONO			/ABANDONO DA CAUSA/ABANDONO DE CARGO/ABANDONO DO POSTO/			/NOC NOMES COMUNS/	
ABDUÇÃO				/ESTRANGEIRO/		/DIN DIREITO INTERNACIO- NAL/	
ABERRAÇÃO NO ATAQUE	/ERRO DE EXECU- ÇÃO/					/DPE DIREITO PENAL/	
ABERRATIO ICTUS	/ERRO DE EXECU- ÇÃO/					/DPE DIREITO PENAL/	
ABERTURA DE CRÉDITO				/CRÉDITO/IOF/		/DIC DIREITO CIVIL/	
ABIN	/AGÊNCIA BRASI- LEIRA DE INTELI- GÊNCIA (ABIN)/					/INS INSTITUI- ÇÕES/	

aquele modelo com menor *ranking* médio será considerado o modelo que melhor captura o conhecimento contido no TSTF.

Concluídas as análises dos modelos, será feita uma análise descritiva que visa identificar características no corpus que possam contribuir na tarefa de agrupar os documentos de acordo com seus assuntos. Essa atividade é de muita importância para o trabalho, uma vez que possibilita identificar palavras e expressões comuns a alguns assuntos e raras em outros que podem ser incluídas nas representações dos documentos para potencialmente melhorar a qualidade do agrupamento.

Para as tarefas a seguir, os assuntos que ocorrem menos de 10.000 vezes no corpus serão descartados, e cada assunto será representado por no máximo 40.000 documentos que serão selecionados aleatoriamente caso o assunto tenha mais do que 40.000 documentos. Assim, os assuntos terão entre 10.000 e 40.000 documentos. O vocabulário que será usado na representação dos documentos será criado de acordo com as seguintes diretrizes:

1. Os termos compostos por uma palavra que aparecem tanto no TSTF quanto no corpus farão parte do vocabulário.
2. Os termos que são exclusivos a um assunto e que aparecem ao menos 100 vezes no corpus farão parte do vocabulário.
3. Os termos que acontecem com mais frequência em alguns poucos assuntos serão identificados de acordo com o inverso da sua frequência em todos os assuntos do corpus, e também serão adicionados no vocabulário.

A terceira diretriz parte da premissa de que os termos comuns a todos os assuntos contribuem menos para a tarefa de agrupamento pretendida do que aqueles termos que aparecem apenas em poucos assuntos. Uma vez que objetiva-se agrupar os documentos de acordo com os seus respectivos assuntos, é desejável que o vocabulário usado para representar estes documentos traga informação relevante para diferenciar um assunto do outro. Para a execução da terceira diretriz, será computado para cada termo o seu *ICA*, ou Índice de Concentração nos Assuntos, que é dado por  $ICA(t) = \log_{10} \left( \frac{A}{n_t} \right)$ , onde  $A$  é a quantidade de assuntos considerada na análise e  $n_t$  é a quantidade de assuntos nos quais o termo  $t$  aparece. Desta maneira, os termos que aparecem em muitos assuntos terão um *ICA* menor. Finalmente, apenas aqueles termos cujo *ICA* é maior que a média do *ICA* acrescida de 1,5 desvios padrões, e que apareçam no mínimo 100 vezes no corpus, serão adicionados ao vocabulário.

A representação dos documentos através do vocabulário será feita através da soma dos vetores das palavras do vocabulário que ocorrem nos documentos. Os vetores das palavras serão ponderados pelo *tf-idf* da palavra que representam, e cada documento terá

uma representação para cada um dos modelos gerados no corpus: Word2Vec, FastText e GloVe de 100 dimensões. A distribuição da distância de cosseno entre os documentos do mesmo assunto e entre documentos de assuntos diferentes será analisada. Ela objetiva contribuir com a análise qualitativa da qualidade das representações geradas por cada modelo, bem como do vocabulário. Espera-se que documentos no mesmo assunto possuam uma distância entre si menor que aquela observada entre documentos de assuntos diferentes.

Em seguida, para cada uma das representações, os documentos serão agrupados com uso dos algoritmos *k*-médias esférico e *expectation-maximization* como implementados na biblioteca *scikit-learn* (PEDREGOSA et al., 2011). Conforme a documentação da biblioteca<sup>5</sup> e as discussões apresentadas na subseção 2.2.2, os vetores dos documentos precisam possuir a mesma norma para que a implementação usada neste trabalho produza resultados equivalentes ao *k*-médias esférico. Uma vez que é de interesse deste trabalho agrupar os documentos de acordo com seus assuntos, a métrica *Adjusted Rand Index* (HUBERT; ARABIE, 1985) será empregada na avaliação objetiva dos agrupamentos.

Uma nova representação dos documentos será feita, porém desta vez serão usados modelos treinados no domínio geral da língua portuguesa. Serão usados os modelos Word2Vec, FastText e GloVe de 100 dimensões pré-treinados e disponibilizados por Hartmann et al. (2017). Uma vez que tratam-se de modelos que representam palavras, as etapas acima descritas para a representação dos documentos serão repetidas, bem como o processo de agrupamento e avaliação dos grupos gerados. Finalmente, as abordagens serão comparadas frente à qualidade dos grupos encontrados.

Finalmente, uma nova representação do corpus será desenvolvida mediante aplicação do algoritmo Doc2Vec proposto em Le e Mikolov (2014b) e também disponível na biblioteca Gensim, e as análises acima descritas serão repetidas para os resultados obtidos neste novo experimento.

As atividades previstas são as seguintes:

- A Preparação e apresentação do exame de qualificação;
- B Revisão bibliográfica sobre *word embeddings* e mineração de textos de documentos jurídicos;
- C Treinamento dos modelos;
- D Criação das representações com os modelos treinados;
- E Realização dos experimentos de agrupamento;
- F Redação da dissertação;

---

<sup>5</sup> <<https://bit.ly/2ZU4j5S>>

G Elaboração e submissão de artigos científicos.

## 4.4 Cronograma

O cronograma das atividades previamente listadas encontra-se na Tabela 12.

Tabela 12 – *Cronograma estimado de desenvolvimento do trabalho.*

Atividades	out/2020 até nov/2020	dez/2020 até jan/2021	fev/2021 até mar/2021	abr/2021 até mai/2021
A	X			
B	X	X	X	X
C	X	X		
D	X	X	X	
E		X	X	
F		X	X	X
G			X	X

# Referências

- AFONSO, A. R. Brazilian portuguese text clustering based on evolutionary computing. *IEEE Latin America Transactions*, IEEE, v. 14, n. 7, p. 3370–3377, 2016. Citado 2 vezes nas páginas 30 e 31.
- Aggarwal, C. C.; Gates, S. C.; Yu, P. S. On using partial supervision for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, v. 16, n. 2, p. 245–255, Feb 2004. ISSN 1041-4347. Citado 2 vezes nas páginas 29 e 31.
- AGGARWAL, C. C.; ZHAI, C. A survey of text clustering algorithms. In: *Mining text data*. [S.l.]: Springer, 2012. p. 77–128. Citado na página 17.
- ALPAYDIN, E. *Introduction to Machine Learning*. 2nd. ed. [S.l.]: The MIT Press, 2010. ISBN 026201243X, 9780262012430. Citado na página 2.
- ARBELAITZ, O. et al. An extensive comparative study of cluster validity indices. *Pattern Recognition*, Elsevier, v. 46, n. 1, p. 243–256, 2013. Citado na página 22.
- ARLIA, D.; COPPOLA, M. Experiments in parallel clustering with dbscan. In: SPRINGER. *European Conference on Parallel Processing*. [S.l.], 2001. p. 326–331. Citado na página 21.
- ASH, E.; CHEN, D. L. Case vectors: Spatial representations of the law using document embeddings. *Available at SSRN 3204926*, 2018. Citado na página 27.
- BIASIOTTI, M. et al. Legal informatics and management of legislative documents. *Global Center for ICT in Parliament Working Paper*, v. 2, 2008. Citado na página 27.
- BIGI, B. Using kullback-leibler distance for text categorization. In: SPRINGER. *European Conference on Information Retrieval*. [S.l.], 2003. p. 305–319. Citado na página 14.
- BILGIN, M.; ŞENTÜRK, İ. F. Sentiment analysis on twitter data with semi-supervised doc2vec. In: IEEE. *2017 international conference on computer science and engineering (UBMK)*. [S.l.], 2017. p. 661–666. Citado na página 10.
- BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738. Citado na página 2.
- BLUM, A.; HOPCROFT, J.; KANNAN, R. Foundations of data science. *Vorabversion eines Lehrbuchs*, v. 5, 2016. Citado na página 8.
- BOJANOWSKI, P. et al. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. Citado na página 38.
- BOJANOWSKI, P. et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 5, p. 135–146, 2017. Citado na página 10.

- BOOS, R. et al. brwac: A wacky corpus for brazilian portuguese. In: BAPTISTA, J. et al. (Ed.). *Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2014. p. 201–206. ISBN 978-3-319-09761-9. Citado na página 29.
- BORA, M. et al. Effect of different distance measures on the performance of k-means algorithm: an experimental study in matlab. *arXiv preprint arXiv:1405.7471*, 2014. Citado na página 18.
- BRANTING, K. et al. Semi-supervised methods for explainable legal prediction. In: ACM. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. [S.l.], 2019. p. 22–31. Citado na página 29.
- BUCHTA, C. et al. Spherical k-means clustering. *Journal of Statistical Software*, American Statistical Association, v. 50, n. 10, p. 1–22, 2012. Citado na página 19.
- CASTANO, S. et al. Crime knowledge extraction: an ontology-driven approach for detecting abstract terms in case law decisions. In: ACM. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. [S.l.], 2019. p. 179–183. Citado na página 29.
- CASTRO, P. V. Q. d. et al. Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico. Universidade Federal de Goiás, 2019. Citado na página 29.
- CASTRO, R. M. d. Direito, econometria e estatística. 2017. Citado na página 28.
- DAI, A. M.; OLAH, C.; LE, Q. V. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*, 2015. Citado 2 vezes nas páginas 10 e 16.
- DEERWESTER, S. et al. Indexing by latent semantic analysis. *Journal of the American society for information science*, Wiley Online Library, v. 41, n. 6, p. 391–407, 1990. Citado na página 8.
- DEY, S. et al. A comparative study of support vector machine and naive bayes classifier for sentiment analysis on amazon product reviews. In: IEEE. *2020 International Conference on Contemporary Computing and Applications (IC3A)*. [S.l.], 2020. p. 217–220. Citado na página 8.
- DHILLON, I. S.; MODHA, D. S. Concept decompositions for large sparse text data using clustering. *Machine learning*, Springer, v. 42, n. 1-2, p. 143–175, 2001. Citado na página 19.
- ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231. Citado 2 vezes nas páginas 20 e 21.
- FARACO, F. M. et al. Análise de agrupamentos sobre textos: Um estudo dos resumos do banco de teses e dissertações da capes. In: *Congresso Internacional de Conhecimento e Inovação-Ciki*. [S.l.: s.n.], 2018. v. 1, n. 1. Citado na página 30.
- FERRERO, J. et al. Usingword embedding for cross-language plagiarism detection. *arXiv preprint arXiv:1702.03082*, 2017. Citado na página 10.

- FRANÇA, R. L. Da jurisprudência como direito positivo. *Revista da Faculdade de Direito, Universidade de São Paulo*, v. 66, p. 201–222, 1971. Citado na página 26.
- FURQUIM, L. O. de C.; LIMA, V. L. S. D. Clustering and categorization of brazilian portuguese legal documents. In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2012. p. 272–283. Citado na página 29.
- FURQUIM LUIS OTÁVIO DE COLLA, L. V. L. S. d. *Agrupamento e categorização de documentos jurídicos*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2011. Citado 2 vezes nas páginas 1 e 2.
- GAN, J.; TAO, Y. Dbscan revisited: Mis-claim, un-fixability, and approximation. In: *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. [S.l.: s.n.], 2015. p. 519–530. Citado na página 21.
- GAONKAR, M. N.; SAWANT, K. Autoepsdbscan: Dbscan with eps automatic for large dataset. *International Journal on Advanced Computer Theory and Engineering*, v. 2, n. 2, p. 11–16, 2013. Citado na página 21.
- GHOSH, J.; STREHL, A. Similarity-based text clustering: A comparative study. In: *Grouping Multidimensional Data*. [S.l.]: Springer, 2006. p. 73–97. Citado na página 13.
- GONZALEZ, M.; LIMA, V. L. Recuperação de informação e processamento da linguagem natural. In: *XXIII Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2003. v. 3, p. 347–395. Citado na página 1.
- HARRIS, Z. S. Distributional structure. *Word*, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954. Citado na página 2.
- HARTMANN, N. et al. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *CoRR*, abs/1708.06025, 2017. Disponível em: <<http://arxiv.org/abs/1708.06025>>. Citado 3 vezes nas páginas 16, 30 e 41.
- HONG, T.-P. et al. Using tf-idf to hide sensitive itemsets. *Applied Intelligence*, Springer, v. 38, n. 4, p. 502–510, 2013. Citado na página 7.
- HRUSCHKA, E. R.; CASTRO, L. N. de; CAMPELLO, R. J. Evolutionary algorithms for clustering gene-expression data. In: IEEE. *Fourth IEEE International Conference on Data Mining (ICDM'04)*. [S.l.], 2004. p. 403–406. Citado na página 25.
- HUANG, A. Similarity measures for text document clustering. In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. [S.l.: s.n.], 2008. v. 4, p. 9–56. Citado 3 vezes nas páginas 13, 14 e 16.
- HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, n. 1, p. 193–218, 1985. Citado 2 vezes nas páginas 22 e 41.
- JACCARD, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, v. 37, p. 547–579, 1901. Citado na página 14.
- JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. [S.l.]: Prentice-Hall, Inc., 1988. Citado 4 vezes nas páginas 11, 12, 13 e 16.

- JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, MCB UP Ltd, 1972. Citado na página 7.
- JR, J. H. W. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, Taylor & Francis Group, v. 58, n. 301, p. 236–244, 1963. Citado 2 vezes nas páginas 17 e 25.
- JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 1st. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000. ISBN 0130950696. Citado na página 1.
- KARAMI, A.; JOHANSSON, R. Choosing dbscan parameters automatically using differential evolution. *International Journal of Computer Applications*, Foundation of Computer Science (FCS), v. 91, n. 7, p. 1–11, 2014. Citado na página 21.
- KARYPIS, M. S. G.; KUMAR, V.; STEINBACH, M. A comparison of document clustering techniques. In: *TextMining Workshop at KDD2000 (May 2000)*. [S.l.: s.n.], 2000. Citado 3 vezes nas páginas 16, 17 e 18.
- KIM, D. et al. Multi-co-training for document classification using various document representations: Tf-idf, lda, and doc2vec. *Information Sciences*, Elsevier, v. 477, p. 15–29, 2019. Citado na página 10.
- KIM, H.; KIM, H. K.; CHO, S. Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling. *Expert Systems with Applications*, Elsevier, v. 150, p. 113288, 2020. Citado na página 19.
- KRIEGEL, H.-P. et al. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 1, n. 3, p. 231–240, 2011. Citado na página 21.
- LAI, W. et al. A new dbscan parameters determination method based on improved mvo. *IEEE Access*, IEEE, v. 7, p. 104085–104095, 2019. Citado na página 21.
- LAKSHMI, A. R.; BALAKRISHNA, V. Efficient clustering of text document using spherical k-means algorithm. *International Journal of Computer Science and Information Technologies*, v. 7, n. 5, p. 2187–2190, 2016. Citado na página 19.
- LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. *Discourse Processes*, Routledge, v. 25, n. 2-3, p. 259–284, 1998. Disponível em: <<https://doi.org/10.1080/01638539809545028>>. Citado na página 2.
- LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: *International conference on machine learning*. [S.l.: s.n.], 2014. p. 1188–1196. Citado na página 10.
- LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: *International conference on machine learning*. [S.l.: s.n.], 2014. p. 1188–1196. Citado na página 41.
- LEE, H.; YOON, Y. Engineering doc2vec for automatic classification of product descriptions on o2o applications. *Electronic Commerce Research*, Springer, v. 18, n. 3, p. 433–456, 2018. Citado na página 10.



- LEE, S.; JIN, X.; KIM, W. Sentiment classification for unlabeled dataset using doc2vec with jst. In: *Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World*. [S.l.: s.n.], 2016. p. 1–5. Citado na página 10.
- LEVY, O.; GOLDBERG, Y. Dependency-based word embeddings. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. [S.l.: s.n.], 2014. p. 302–308. Citado na página 9.
- LI, M. et al. The seeding algorithms for spherical k-means clustering. *Journal of Global Optimization*, Springer, p. 1–14, 2019. Citado na página 19.
- LIKAS, A.; VLASSIS, N.; VERBEEK, J. J. The global k-means clustering algorithm. *Pattern recognition*, Elsevier, v. 36, n. 2, p. 451–461, 2003. Citado na página 18.
- LING, W. et al. Two/too simple adaptations of word2vec for syntax problems. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2015. p. 1299–1304. Citado na página 10.
- LOEVINGER, L. Jurimetrics—the next step forward. *Minn. L. Rev.*, HeinOnline, v. 33, p. 455, 1948. Citado 2 vezes nas páginas 26 e 27.
- LOOHACH, R.; GARG, K. Effect of distance functions on k-means clustering algorithm. *International Journal of Computer Applications*, Citeseer, v. 49, n. 6, p. 7–9, 2012. Citado na página 18.
- LOVINS, J. B. Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, v. 11, n. 1-2, p. 22–31, 1968. Citado na página 10.
- MAGALHÃES, L. H. de; SOUZA, R. R. Agrupamento automático de notícias de jornais on-line usando técnicas de machine learning para clustering de textos no idioma português. *Múltiplos Olhares em Ciência da Informação*, v. 9, n. 2, 2019. Citado na página 30.
- MAIA, L. C. G. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. *Perspectivas em Ciência da Informação*, SciELO Brasil, v. 14, n. 3, p. 237–237, 2009. Citado na página 30.
- MANDAL, A. et al. Measuring similarity among legal court case documents. In: *Proceedings of the 10th Annual ACM India Compute Conference*. [S.l.: s.n.], 2017. p. 1–9. Citado na página 27.
- MANNING, C. D.; MANNING, C. D.; SCHÜTZE, H. *Foundations of statistical natural language processing*. [S.l.]: MIT press, 1999. Citado na página 5.
- METZLER, D.; DUMAIS, S.; MEEK, C. Similarity measures for short segments of text. In: SPRINGER. *European conference on information retrieval*. [S.l.], 2007. p. 16–27. Citado na página 13.
- MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. Disponível em: <<http://arxiv.org/abs/1301.3781>>. Citado 2 vezes nas páginas 1 e 2.

- MIKOLOV, T. et al. *Computing numeric representations of words in a high-dimensional space*. [S.l.]: Google Patents, 2019. US Patent 10,241,997. Citado na página 16.
- MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 3111–3119. Citado 2 vezes nas páginas 10 e 38.
- MILLIGAN, G. W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *psychometrika*, Springer, v. 45, n. 3, p. 325–342, 1980. Citado na página 18.
- MILLIGAN, G. W.; COOPER, M. C. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate behavioral research*, Taylor & Francis, v. 21, n. 4, p. 441–458, 1986. Citado na página 22.
- MOULAVI, D. et al. Density-based clustering validation. In: SIAM. *Proceedings of the 2014 SIAM international conference on data mining*. [S.l.], 2014. p. 839–847. Citado na página 25.
- NASSIF, L. F. da C.; HRUSCHKA, E. R. Document clustering for forensic computing: An approach for improving computer inspection. In: IEEE. *2011 10th International Conference on Machine Learning and Applications and Workshops*. [S.l.], 2011. v. 1, p. 265–268. Citado na página 17.
- NUNES, D. Jurimetria e tecnologia: Diálogos essenciais com o direito processual. *Revista de Processo/ vol*, v. 299, n. 2020, p. 407–450, 2020. Citado na página 28.
- OLIVEIRA, A. d. Comportamento de gestores de recursos públicos: identificação de contingências previstas e vigentes relativas à prestação de contas. 2016. Citado na página 28.
- PAIK, J. H. A novel tf-idf weighting scheme for effective ranking. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2013. (SIGIR '13), p. 343–352. ISBN 978-1-4503-2034-4. Disponível em: <<http://doi.acm.org/10.1145/2484028.2484070>>. Citado na página 1.
- PAPAKYRIAKOPOULOS, O. et al. Bias in word embeddings. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. [S.l.: s.n.], 2020. p. 446–457. Citado na página 9.
- PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 41.
- PENA, J. M.; LOZANO, J. A.; LARRANAGA, P. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters*, Elsevier, v. 20, n. 10, p. 1027–1040, 1999. Citado na página 18.
- PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543. Citado 2 vezes nas páginas 10 e 38.

- PETERS, M. E. et al. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018. Citado na página 29.
- PORTER, M. F. et al. An algorithm for suffix stripping. *Program*, Citeseer, v. 14, n. 3, p. 130–137, 1980. Citado na página 6.
- QAISER, S.; ALI, R. Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, v. 181, n. 1, p. 25–29, 2018. Citado na página 7.
- RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: PISCATAWAY, NJ. *Proceedings of the first instructional conference on machine learning*. [S.l.], 2003. v. 242, p. 133–142. Citado na página 7.
- RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, Taylor & Francis Group, v. 66, n. 336, p. 846–850, 1971. Citado na página 22.
- RAVAGNANI, G. dos S. Automação da advocacia, gestão de contencioso de massa e a atuação estratégica do grande litigante. *Revista de Processo/ vol*, v. 265, n. 2017, p. 219–256, 2017. Citado na página 28.
- ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010. p. 45–50. <<http://is.muni.cz/publication/884893/en>>. Citado na página 38.
- RENDÓN, E. et al. Internal versus external cluster validation indexes. *International Journal of computers and communications*, v. 5, n. 1, p. 27–34, 2011. Citado 2 vezes nas páginas 23 e 25.
- RESHMA, P.; RAJAGOPAL, S.; LAJISH, V. A novel document and query similarity indexing using vsm for unstructured documents. In: IEEE. *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. [S.l.], 2020. p. 676–681. Citado na página 8.
- ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, North-Holland, v. 20, p. 53–65, 1987. Citado na página 24.
- ROUSSEEUW, P. J.; KAUFMAN, L. Finding groups in data. *Hoboken: Wiley Online Library*, 1990. Citado na página 18.
- SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information processing & management*, Elsevier, v. 24, n. 5, p. 513–523, 1988. Citado 2 vezes nas páginas 7 e 8.
- SALTON, G.; WONG, A.; YANG, C.-S. A vector space model for automatic indexing. *Communications of the ACM*, ACM New York, NY, USA, v. 18, n. 11, p. 613–620, 1975. Citado na página 5.
- SCHUBERT, E. et al. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, ACM New York, NY, USA, v. 42, n. 3, p. 1–21, 2017. Citado na página 21.

- SCHÜTZE, H.; MANNING, C. D.; RAGHAVAN, P. *Introduction to information retrieval*. [S.l.]: Cambridge University Press Cambridge, 2008. v. 39. Citado na página 5.
- SEDGWICK, P. Pearson's correlation coefficient. *Bmj*, British Medical Journal Publishing Group, v. 345, p. e4483, 2012. Citado na página 14.
- SINGH, V. K.; TIWARI, N.; GARG, S. Document clustering using k-means, heuristic k-means and fuzzy c-means. In: IEEE. *2011 International Conference on Computational Intelligence and Communication Networks*. [S.l.], 2011. p. 297–301. Citado na página 7.
- STREHL, A.; GHOSH, J.; MOONEY, R. Impact of similarity measures on web-page clustering. In: *Workshop on artificial intelligence for web search (AAAI 2000)*. [S.l.: s.n.], 2000. v. 58, p. 64. Citado 2 vezes nas páginas 13 e 16.
- SUGATHADASA, K. et al. Legal document retrieval using document vector embeddings and deep learning. In: SPRINGER. *Science and Information Conference*. [S.l.], 2018. p. 160–175. Citado na página 27.
- SURDEN, H. Machine learning and law. *Wash. L. Rev.*, HeinOnline, v. 89, p. 87, 2014. Citado 2 vezes nas páginas 27 e 28.
- SZEKELY, G. J.; RIZZO, M. L. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, v. 22, n. 2, 2005. Citado na página 17.
- TANG, B. et al. Comparing dimension reduction techniques for document clustering. In: SPRINGER. *Conference of the Canadian Society for Computational Studies of Intelligence*. [S.l.], 2005. p. 292–296. Citado na página 9.
- TOMASINI, C. et al. A methodology for selecting the most suitable cluster validation internal indices. In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. [S.l.: s.n.], 2016. p. 901–903. Citado na página 25.
- TRIEU, L. Q.; TRAN, H. Q.; TRAN, M.-T. News classification from social media using twitter-based doc2vec model and automatic query expansion. In: *Proceedings of the Eighth International Symposium on Information and Communication Technology*. [S.l.: s.n.], 2017. p. 460–467. Citado na página 10.
- TRSTENJAK, B.; MIKAC, S.; DONKO, D. Knn with tf-idf based framework for text categorization. *Procedia Engineering*, Elsevier, v. 69, p. 1356–1364, 2014. Citado na página 7.
- TUMMERS, J. et al. Coronaviruses and people with intellectual disability: An exploratory data analysis. *Journal of Intellectual Disability Research*, Wiley Online Library, 2020. Citado na página 8.
- TUNALI, V.; BILGIN, T.; CAMURCU, A. An improved clustering algorithm for text mining: Multi-cluster spherical k-means. *International Arab Journal of Information Technology (IAJIT)*, v. 13, n. 1, 2016. Citado na página 19.
- WANG, F. et al. An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In: SPRINGER. *International Conference on Machine Learning and Data Mining in Pattern Recognition*. [S.l.], 2017. p. 291–305. Citado 2 vezes nas páginas 16 e 25.

- WOLKIND, S.; EVERITT, B. A cluster analysis of the behavioural items in the pre-school child. *Psychological medicine*, Cambridge University Press, v. 4, n. 4, p. 422–427, 1974. Citado na página 11.
- XIONG, H.; LI, Z. *Clustering Validation Measures*. [S.l.]: Citeseer, 2013. Citado na página 25.
- YIH, W.-T.; MEEK, C. Improving similarity measures for short segments of text. In: *AAAI*. [S.l.: s.n.], 2007. v. 7, n. 7, p. 1489–1494. Citado na página 13.
- ZABALA, F. J.; SILVEIRA, F. F. Jurimetria: estatística aplicada ao direito. *Revista Direito e Liberdade*, v. 16, n. 1, p. 87–103, 2014. Citado 2 vezes nas páginas 27 e 28.
- ZHANG, W.; YOSHIDA, T.; TANG, X. A comparative study of tf\* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, Elsevier, v. 38, n. 3, p. 2758–2765, 2011. Citado na página 9.
- ZHONG, S.; GHOSH, J. A comparative study of generative models for document clustering. In: CITESEER. *Proceedings of the workshop on clustering high dimensional data and its applications in SIAM data mining conference*. [S.l.], 2003. Citado na página 19.
- ZHOU, H.; WANG, P.; LI, H. Research on adaptive parameters determination in dbscan algorithm. *Journal of Xi'an University of Technology*, v. 28, n. 3, p. 289–292, 2012. Citado na página 21.
- ZHU, Z. et al. Hot topic detection based on a refined tf-idf algorithm. *IEEE Access*, IEEE, v. 7, p. 26996–27007, 2019. Citado na página 7.



## Apêndices





# APÊNDICE A – Atividades de preprocessamento

O conteúdo dos documentos será submetido às seguintes atividades de preprocessamento:

- Substituição dos caracteres `\n` e `\r` por espaços
- Remoção de *stopwords*
- Substituição de todas as letras maiúsculas por minúsculas
- Substituição de todas as letras acentuadas por sua contraparte sem acentuação
- Substituição de todos os caracteres numéricos por 0.

A lista de *stopwords* é composta pelas seguintes palavras: “de”, “a”, “o”, “que”, “e”, “é”, “do”, “da”, “em”, “um”, “para”, “com”, “não”, “uma”, “os”, “no”, “se”, “na”, “por”, “mais”, “as”, “dos”, “como”, “mas”, “ao”, “ele”, “das”, “à”, “seu”, “sua”, “ou”, “quando”, “muito”, “nos”, “já”, “eu”, “também”, “só”, “pelo”, “pela”, “até”, “isso”, “ela”, “entre”, “depois”, “sem”, “mesmo”, “aos”, “seus”, “quem”, “nas”, “me”, “esse”, “eles”, “você”, “essa”, “num”, “nem”, “suas”, “meu”, “às”, “minha”, “numa”, “pelos”, “elas”, “qual”, “nós”, “lhe”, “deles”, “essas”, “esses”, “pelas”, “este”, “dele”, “tu”, “te”, “vocês”, “vos”, “lhes”, “meus”, “minhas”, “teu”, “tua”, “teus”, “tuas”, “nosso”, “nossa”, “nossos”, “nossas”, “dela”, “delas”, “esta”, “estes”, “estas”, “aquele”, “aquela”, “aqueles”, “aquelas”, “isto”, “aquilo”, “estou”, “está”, “estamos”, “estão”, “estive”, “estive”, “estivemos”, “estiveram”, “estava”, “estávamos”, “estavam”, “estivera”, “estivéramos”, “esteja”, “estejamos”, “estejam”, “estivesse”, “estivéssemos”, “estivessem”, “estiver”, “estivermos”, “estiverem”, “hei”, “há”, “havemos”, “hã”, “houve”, “houvermos”, “houveram”, “houvera”, “houveramos”, “haja”, “hajamos”, “hajam”, “houvesse”, “houvéssemos”, “houvessem”, “houver”, “houvermos”, “houverem”, “houverei”, “houverá”, “houveremos”, “houverão”, “houveria”, “houveríamos”, “houveriam”, “sou”, “somos”, “são”, “era”, “éramos”, “eram”, “fui”, “foi”, “fomos”, “foram”, “fora”, “fôramos”, “seja”, “sejamos”, “sejam”, “fosse”, “fôssemos”, “fossem”, “for”, “formos”, “forem”, “serei”, “será”, “seremos”, “serão”, “seria”, “seríamos”, “seriam”, “tenho”, “tem”, “temos”, “tém”, “tinha”, “tínhamos”, “tinham”, “tive”, “teve”, “tivemos”, “tiveram”, “tivera”, “tivéramos”, “tenha”, “tenhamos”, “tenham”, “tivesse”, “tivéssemos”, “tivessem”, “tiver”, “tivermos”, “tiverem”, “terei”, “terá”, “teremos”, “terão”, “teria”, “teríamos”, “teriam”.



## Anexos

