UNIVERSIDADE FEDERAL DE SERGIPE

CAMPUS SÃO CRISTÓVÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Um Estudo da Representação de Documentos Jurídicos em Espaços Métricos

Dissertação de Mestrado

GUSTAVO MENEZES MACHADO

SÃO CRISTÓVÃO

2019

UNIVERSIDADE FEDERAL DE SERGIPE
CAMPUS SÃO CRISTÓVÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GUSTAVO MENEZES MACHADO

Um Estudo da Representação de Documentos Jurídicos em Espaços Métricos

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PROCC) da Universidade Federal de Sergipe (UFS) como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Dr. JUGURTA ROSA MONTALVÃO FILHO

SÃO CRISTÓVÃO
2019

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Machado, Gustavo Menezes
M149e Um estudo da representação de documentos jurídicos em espaços métricos / Gustavo Menezes Machado ; orientador Jugurta Rosa Montalvão Filho . - São Cristóvão, 2019.
74 f.

Dissertação (mestrado em Ciência da Computação) – Universidade Federal de Sergipe, 2019.

1. Computação. 2. Jurisprudência. 3. Multidimensional scaling. 4. Julgamentos. I. Machado, Gustavo Menezes orient. II. Título.

CDU 004:34

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS.
Candidato: GUSTAVO MENEZES MACHADO

Em 20 dias do mês de Agosto do ano de dois mil e dezenove, com início às 14h00min, realizou-se na Sala de Seminário do DCOMP da Universidade Federal de Sergipe, na Cidade Universitária Prof. José Aloísio de Campos, a Sessão Pública de Defesa de Dissertação de Mestrado do candidato **Gustavo Menezes Machado**, que desenvolveu o trabalho intitulado: *" Um estudo da representação de documentos jurídicos em espaços métricos"*, sob a orientação do Prof. Dr. **Jugurta Rosa Montalvão Filho**. A Sessão foi presidida pelo Prof. Dr. **Jugurta Rosa Montalvão Filho** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, Prof. Dr. **Rogério Patrício Chagas do Nascimento** (PROCC/UFS) e, em seguida, ao Prof. **Jânio Coutinho Canuto** (UFS). Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a) APROVADO *"(aprovado/reprovado)"*. Atendidas as exigências da Instrução Normativa 01/2017/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), e da Resolução nº 25/2014/CONEPE que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária *"Prof. José Aloísio de Campos"*, 20 de Agosto de 2019.

Prof. Dr. Jugurta Rosa Montalvão Filho
(PROCC/UFS)
Presidente

Prof. Dr. Rogério Patrício Chagas do Nascimento
(PROCC/UFS)
Examinador Interno

Prof. Dr. Jânio Coutinho Canuto
(UFS)
Examinador Externo

Gustavo Menezes Machado
Candidato

“À minha amada esposa Aline e
à minha filha Maria Clara"

AGRADECIMENTOS

A minha família, por sua capacidade de acreditar em mim e sempre me apoiar.

Aos meus amigos que sempre me ajudaram e me apoiaram nos momentos mais difíceis.

Ao Prof. Dr. Jugurta Rosa Montalvão Filho, por ter me orientado durante esse período, pelas conversas e ensinamentos passados.

Aos meus colegas de trabalho que me ajudaram nas dúvidas e discussões levantadas a respeito do tema pesquisado.

Ao PROCC e às pessoas com quem convivi na universidade ao longo desses anos.

A todos aqueles que de alguma forma estiveram ou estão próximos, fazendo esta vida valer cada vez mais a pena.

“O valor das coisas não está no tempo que elas duram, mas na intensidade com que elas acontecem. Por isso existem momentos inesquecíveis, coisas inexplicáveis e pessoas incomparáveis”. (Fernando Pessoa).

RESUMO

Diariamente são elaboradas dezenas de decisões a partir de interpretações das leis realizadas por tribunais de todo o país. Este conjunto de decisões similares sobre uma mesma matéria é conhecido como jurisprudência, e serve como base para julgamentos e argumentações futuras. Nos textos jurídicos escritos em português brasileiro, além das palavras serem guiadas por regras estéticas diferentes, há também o uso de referências frequentes a elementos jurídicos, o que torna a análise de textos jurídicos escritos em português brasileiro um problema estimulante. Neste trabalho, é explorado um espaço métrico associado a contextos e ao compartilhamento de símbolos entre contextos de documentos jurídicos, ou seja, trata-se da busca por um espaço adequado à representação de textos como processos judiciais, onde cada processo - ou parte dele - é representado como um ponto, e as distâncias entre esses pontos representam medidas probabilísticas. Para tal representação, foi utilizado o Multidimensional Scaling (MDS), que é uma técnica de redução de dimensionalidade onde as relações de distâncias entre os pontos no espaço projetado se aproximam das medidas de proximidade dos objetos do espaço original. A base de jurisprudência do Tribunal de Justiça do Estado de Sergipe foi utilizada, além de um conjunto controlado de palavras utilizadas na área jurídica, disponibilizado pelo Supremo Tribunal Federal. Os experimentos realizados evidenciaram que o método proposto conseguiu uma melhor classificação em 43,5% dos casos, enquanto Doc2Vec foi superior em apenas 35,7% das vezes, evidenciando a existência de um espaço métrico mais adequado à representação de textos jurídicos escritos em português brasileiro, que um espaço puramente baseado em co-ocorrência de símbolos, como o que é encontrado pelo Doc2Vec.

Palavras chave: Jurisprudência, MDS, Doc2Vec, julgamentos, textos jurídicos.

ABSTRACT

Dozens of decisions are made daily from interpretations of the laws made by courts across the country. This set of similar decisions on the same subject is known as jurisprudence and serves as the basis for future judgments and arguments. In legal texts written in Brazilian Portuguese, in addition to words being guided by different esthetic rules, there is also the use of frequent references to legal elements, which makes the analysis of legal texts written in Brazilian Portuguese a stimulating problem. This work explores a metric space associated with contexts and the sharing of symbols between contexts of legal documents, that is, the search for spaces suitable for the representation of texts as court lawsuits, where each process - or part of it - is represented as a point, and the distances between these points represents probabilistic measures. For such representation, the Multidimensional Scaling (MDS) was used, which is a technique of dimensionality reduction where the relations of distances between the points in the projected space approximate the proximity measurements of the objects of the original space. The case law of the Sergipe State Court of Justice was used, in addition to a controlled set of words used in the legal area, provided by the Federal Supreme Court. The experiments showed that the proposed method obtained a better classification in 43.5% of the cases, while Doc2Vec was superior in only 35.7% of the cases, evidencing the existence of a more adequate metric space for the representation of legal texts written in Brazilian Portuguese than a space purely based on co-occurrence of symbols, as found by Doc2Vec.

Keywords: Jurisprudence, MDS, Doc2Vec, judgments, judicial documents.

LISTA DE ILUSTRAÇÕES

Figura 1: Pontos no espaço 2D com distâncias aproximadas aos valores da matriz de distâncias……..……..…24
Figura 2: Funcionamento do CBOW……..…….……26
Figura 3: Funcionamento Skip-gram……..……..…26
Figura 4: Funcionamento do PV-DBOW……..…..……....…..…28
Figura 5: Funcionamento do PV-DM…….....……..…28
Figura 6: Exemplo de consulta ao tesauro, buscando o termo "norma penal em branco".....30
Figura 7: Tratamento dos termos do tesauro jurídico……..…..……..….31
Figura 8: Exemplo de conteúdo processual extraído de um arquivo xml pertencente á base jurisprudencial…….....…….…32
Figura 9: Pré-processamento dos dados processuais……..…....……..…35
Figura 10: Exemplo de consulta ao tesauro, procurando pelo termo "dano psicológico"…..40
Figura 11: Extração das EJs dos processos, construção da matriz de similaridade e seleção dos processos semelhantes……...……..……..……43
Figura 12: Extração dos termos dos processos, construção da matriz de similaridade e seleção dos processos semelhantes……...……..…….……...…44
Figura 13: Geração da matriz de distâncias com base nas EJs e termos…..……..……..…45
Figura 14: Redução gradativa do stress……..……..…49
Figura 15: Gráfico com vetores de processos resultantes da aplicação do MDS…..……..…49
Figura 16: Redução gradativa do stress até atingir o valor de 5% na escala de Kruskal…….51
Figura 17: Extração dos tipos dos processos, construção da matriz de similaridade e seleção dos processos semelhantes……...……..…….……...53
Figura 18: Aplicação do Doc2Vec aos processos para cálculo da semelhança entre processos……...……..……..……55
Figura 19: Comparação dos resultados da Matriz de Distâncias e do Doc2Vec aos resultados baseados nos tipos dos processos…….…..……...57

LISTA DE TABELAS

Tabela 1: Matriz simétrica com as distâncias entre as dez cidades……...…….…24
Tabela 2: Configuração do servidor utilizado para os testes dos experimentos……....…….29
Tabela 3: Exemplos de processos e seus tipos associados……..…...…..….34
Tabela 4: Exemplos de formatação dos "Elementos Jurídicos" encontrados…….…….……37
Tabela 5: Exemplo de entidades jurídicas extraídas de cinco processos exemplo…....…..38
Tabela 6: Exemplo de termos do tesauro e suas respectivas quantidades de ocorrência extraídos de dois processos…….....…….……39
Tabela 7: Exemplos de termos do tesauro extraídos de cinco processos exemplo..……..…..41
Tabela 8: Exemplo de matriz de distâncias calcula com base nas EJs e termos…...…….…45
Tabela 9: Representação vetorial dos cinco processos em 2D após aplicação do MDS….....46
Tabela 10: Classificação do *Stress* de Kruskal……...…...……..…….…..48
Tabela 11: Tipos dos 5 processos utilizados como exemplo.……...…...50
Tabela 12: Exemplo de matriz de similaridade entre processos por tipo……...…..……53
Tabela 13: Exemplo de semelhança entre processos com base nos seus tipos……..…….….54
Tabela 14: Conjuntos de EJs de cinco processos exemplo……....…….…60
Tabela 15: Conjuntos de termos dos processos com suas respectivas quantidades de ocorrência……...…….61

LISTA DE ABREVIATURAS E SIGLAS

2D	Duas dimensões
CBOW	*Continuous Bag of Words*
DOC2VEC	*Document to Vector*
EJ	Entidades Jurídicas
MDS	*Multi-dimensional scaling*
NLTK	*Natural Language Toolkit*
PLN	Processamento de Linguagem Natural
PV-DBOW	*Distributed Bag of Words version of Paragraph Vector*
PV-DM	*Distributed Memory Model of Paragraph Vectors*
STF	Supremo Tribunal Federal
STRESS	*Standardized Residual Sum of Squares*
TESAURO	Vocabulário Jurídico
TJ	Tribunal de Justiça
TJSE	Tribunal de Justiça do Estado de Sergipe
TST	Tribunal Superior do Trabalho
XML	*Extensible Markup Language*

SUMÁRIO

INTRODUÇÃO **14**
Análise do problema 16
Justificativa 17
Objetivos da pesquisa 17
Objetivos específicos 18
Organização da dissertação 18
FUNDAMENTAÇÃO TEÓRICA **19**
Jurisprudência 19
Trabalhos relacionados 21
Métodos utilizados 22
MDS (Multidimensional scaling) 22
Word2Vec 25
Continuous bag of words 25
Skip-gram 26
Doc2Vec 27
PV-DBOW 27
PV-DM 28
Ambiente 29
Métricas para avaliação dos resultados 29
METODOLOGIA **30**
Vocabulário jurídico (tesauro) 30
Base de jurisprudência 32
Tipos dos processos 32
O pré-processamento 35
Método proposto 36
Análise dos resultados **51**
Experimento 1 51
Experimento 2 52
Experimento 3 57
APLICAÇÃO **59**
Tratamento dos dados 59
Geração dos conjuntos de símbolos 59
EJs 60
Termos do tesauro 60
Comparação dos resultados 61

CONCLUSÃO **63**

TRABALHOS FUTUROS **64**

REFERÊNCIAS **65**

1. INTRODUÇÃO

Em 2001, com a lei 10.259/01, foi incentivado o uso da Tecnologia da Informação para o desenvolvimento de sistemas de comunicação de atos processuais, permitindo o ajuizamento de ações de forma eletrônica, a realização de videoconferências entre os juízes e o desenvolvimento de soluções tecnológicas para subsidiar a instrução judicial. (Hino, 2014)

De acordo com o conceito do direito comum, o direito tem uma natureza dinâmica, as leis são constantemente melhoradas para se adaptar às mudanças (necessidades e valores da sociedade, discrição legal e tecnologia) e novos conceitos legais são expostos através de julgamentos anteriores (Kumar, 2014), ou seja, são baseados fortemente no conceito de precedência (Galgani, 2012).

O total de processos que tramitaram no judiciário vem aumentando exponencialmente, o que sobrecarrega os serviços prestados à sociedade e atrasa o atendimento às necessidades dos cidadãos (Hachey, 2006). Á medida que a quantidade de casos judiciais aumenta, sobe também a complexidade desses casos e, consequentemente, dificulta-se o processo de busca e recuperação de processos (Phahlamohlaka, 2018).

Diariamente são elaboradas dezenas de decisões a partir de interpretações das leis realizadas por magistrados de todo o país. Esse conjunto de decisões é conhecido como jurisprudência, que é uma palavra composta por dois termos originários do latim: jus ("justo") + prudentia ("prudência"), nada mais é do que um conjunto de decisões dos tribunais, ou uma série de decisões similares sobre uma mesma matéria (Venosa, 2006) que serve como base para julgamentos e argumentações futuras, garantindo a uniformidade das decisões.

Essas decisões podem ser vistas como uma imensa massa de sinais simbólicos, onde cada palavra e, ainda mais relevante, cada referência a elementos jurídicos (e.g. leis, súmulas, medidas provisórias e artigos) corresponde a um símbolo, aglutinados nos textos de forma a refletir seu conteúdo semântico. A referência a elementos jurídicos carrega muita informação subjetiva relativa ao tema tratado no texto, o que torna o processamento desse tipo de texto um problema com características bem particulares.

Dado o volume crescente desses textos jurídicos, particularmente em bases jurisprudenciais usadas e armazenadas em tribunais, se faz necessária a utilização de vários tipos de processamento automático (via dispositivos eletrônicos), o que demanda, primeiramente, a sua modelagem probabilística, pois, apesar dos aspectos formais dos textos

jurídicos, os símbolos usados nesses textos geralmente não podem ser antecipados de forma determinística. Em outras palavras, assim como acontece na maioria dos casos envolvendo linguagem humana, os modelos probabilísticos se impõem em detrimento dos modelos determinísticos.

No entanto, sabe-se que a estimação de modelos probabilísticos é sensível à relação entre dados disponíveis (amostra de sinais) e à dimensão efetiva em que esses dados são representados (Menacer, 2016). No caso de dados simbólicos como textos, o problema se agrava pelo fato de não haver uma definição óbvia de métrica ou dimensão espacial.

Nos últimos anos, a modelagem de linguagens e o Processamento de Linguagem Natural (PLN) ganharam maior visibilidade graças a utilização da representação vetorial das palavras (*word embeddings*)(Ahmad, 2016). O *embedding* de uma palavra é a sua representação em um espaço multidimensional onde palavras fracamente relacionadas, em termos probabilísticos, encontram-se distantes umas das outras. Por sua vez, palavras com relação probabilística mútua forte encontram-se próximas, pois pressupõe possuírem relação semântica (Deng e Yu, 2014).

Neste trabalho, é explorado um espaço métrico[1] associado a contextos (conjuntos de símbolos) e ao compartilhamento de símbolos entre contextos de documentos jurídicos, ou seja, trata-se da busca por espaços adequados à representação de textos como processos judiciais, onde cada processo - ou parte dele - é representado como um ponto, e as distâncias entre esses pontos representam medidas probabilísticas como, por exemplo, informação mútua entre variáveis aleatórias que representam a ocorrência de símbolos tipicamente usados na composição dos textos jurídicos.

Graças ao fato de abranger diversas áreas e sua capacidade de capturar tanto informação sintática quanto semântica, métricas implícitas usadas em *word embeddings* têm encontrado notável sucesso (Iacobacci, 2015). No entanto, para os estudos aqui realizados, com a particularidade das sequências de símbolos representarem textos jurídicos em português brasileiro, foi escolhido o uso de métricas explícitas, mas que são inspiradas nos critérios implícitos otimizados em *word embeddings*, a saber, medidas de probabilidade de co-ocorrência de símbolos em observações (Levy, 2014b).

A base de jurisprudência utilizada neste trabalho foi coletada no Tribunal de Justiça do

[1] Na matemática, um espaço métrico é um conjunto onde as distâncias entre todos os seus elementos é definida. O conjunto de todas essas distâncias é chamado de métrica no conjunto.

Estado de Sergipe, o que viabilizou a execução de experimentos no contexto específico dos textos jurídicos, destacando suas especificidades.

1.1. Análise do problema

A linguística computacional vem mostrando que é possível se obter uma boa aproximação do significado das palavras a partir do seu contexto (Baroni, 2014). Com base na hipótese de que, palavras que compartilham contextos semelhantes tendem a ter sentidos semelhantes (Ning, 2016), tornou-se possível o cálculo da diferença de significado entre as palavras e a mensuração do grau de similaridade das mesmas (Ju, 2015)(Nguyen, 2016).

Segundo (Zheng, 2017), o que aprendemos com a modelagem de linguagens é baseado no contexto das palavras e não na sua semântica. De acordo com (Guo, 2009), a qualidade dos contextos pode ser melhorada com a utilização de uma abordagem probabilística. Os vetores de contexto são definidos como as distribuições probabilísticas de seus contextos, ou seja, a distribuição de probabilidade das palavras que ocorrem no entorno de uma palavra dada (Gao, Yang & Xu, 2016).

Idealmente, a comparação entre conteúdos de, por exemplo, dois textos em linguagem natural relatando um fato, deveria ser feita baseada em semântica, pois um mesmo fato pode ser relatado usando palavras diferentes. Avanços recentes de análise automática de textos em linguagem natural. Contudo, vêm encontrando sucesso na representação de palavras em espaços vetoriais (métricos), com base apenas na análise sintática. Esse sucesso é explicado com base na repetição estatística de co-ocorrência de palavras e contextos que refletem razoavelmente os significados semânticos.

A área de predição de resultados judiciais utilizando aprendizado de máquina tem recebido crescente atenção ao longo dos anos (Liu, 2017). Segundo (Aletras, 2016), os julgamentos publicados pelos tribunais podem ser utilizados como modelos, uma vez que apresentam um número significativo de semelhanças às petições dirigidas à corte e aos resumos apresentados pelas partes em casos pendentes.

Uma vez que uma palavra pode possuir diversos significados, para a definição destes, é necessária a sua análise em ambientes e contextos específicos. Apesar de existirem métodos capazes de encontrar similaridade entre textos, os mesmos não levam em consideração a natureza específica do domínio jurídico (Moens, 2007) (Kumar, 2014).

Por outro lado, nos textos jurídicos escritos em português brasileiro, além das palavras possuírem um jargão próprio, ou seja, serem guiadas por regras estéticas diferentes daquelas usadas em textos literários e/ou coloquiais (e.g. rebuscamento, uso de termos raros, aversão à repetição e uso de termos em Latim) (Oliveira, 2018), há também o uso de referências frequentes a elementos jurídicos (leis, súmulas, etc.).

Esses dois aspectos, particularmente o segundo (pois esses elementos concentram significados relevantes à análise semântica), tornam a análise de textos jurídicos escritos em português brasileiro um problema estimulante, mas ainda pouco estudado, com especificidades que justificam inclusive o seu estudo acadêmico.

Buscando melhores resultados na construção dos vetores de contextos, decidimos trabalhar diretamente com uma base jurídica. O recorte dado ao estudo levanta como suposição guia a existência de espaços específicos em que textos jurídicos são pontos, e as distâncias refletem adequadamente dissimilaridades de conteúdos.

Com o uso consequente desses espaços métricos, espera-se que seja possível inferir (por proximidade entre processos) informações sobre processos como: Qual o entendimento de determinado tribunal em relação a determinado assunto; Qual o posicionamento de um juiz específico quando aborda um caso específico; Quais documentos foram julgados por determinado juiz, dentre muitas outras informações que podem auxiliar as partes envolvidas em um processo.

1.2. Justificativa

Se existir um espaço métrico mais adequado a textos jurídicos em português brasileiro, espera-se que ele seja utilizado como referência (sua fundamentação teórica e procedimentos utilizados para encontrá-lo) para estudos acadêmicos futuros, assim como em aplicações de interesses mais imediatos nos ambientes dos tribunais.

1.3. Objetivos da pesquisa

O objetivo deste projeto é utilizar e estudar técnicas para a representação dos espaços métricos adequados à representação de bases jurisprudenciais escritas em português brasileiro.

1.4. Objetivos específicos

Os objetivos norteadores do presente trabalho são os expostos abaixo:

- Realização de revisão da literatura relacionada à representação de palavras em espaços métricos;
- Implementação de solução capaz de realizar o agrupamento semântico dos processos utilizando as técnicas estudadas;
- Comparação dos resultados àqueles obtidos com o Doc2Vec para a mesma coleção de dados.

1.5. Organização da dissertação

Os capítulos estão organizados da seguinte forma:

- No capítulo 2 está descrita a fundamentação teórica do trabalho, o método utilizado e as métricas definidas para avaliação dos resultados;
- Capítulo 3 fala sobre a metodologia seguida, bases de dados utilizadas e descreve as atividades realizadas;
- Capítulo 4 faz uma análise dos resultados e um detalhamento dos experimentos realizados;
- Capítulo 5 demonstra uma aplicação prática para o algoritmo proposto;
- Capítulo 6 refere-se às conclusões do trabalho, seguidos pelos trabalhos futuros (Capítulo 7) e as referências bibliográficas(Capítulo 8).

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão detalhados os principais assuntos estudados durante a execução dessa pesquisa, iniciando pela explicação do termo jurisprudência, em seguida, os métodos utilizados e, finalmente, as métricas utilizadas para avaliação dos resultados obtidos.

2.1. Jurisprudência

Jurisprudência[2] é um termo jurídico que se refere a um conjunto de decisões de um tribunal interpretando leis e casos concretos (casos em que efetivamente chegaram ao Poder Judiciário por provocação de partes). Uma única decisão de um tribunal é apenas um precedente, sendo a jurisprudência formada por um conjunto de decisões judiciais julgando outras tantas decisões de juízes singulares.

Uma "decisão de um tribunal" é uma decisão colegiada (de juízes ou ministros) emanada por uma Turma ou Plenário de um tribunal. Este tribunal pode ser tanto da Justiça Comum, quanto da Justiça Especial (Trabalhista, Eleitoral ou Militar). Como exemplo: TJ's (Tribunais de Justiça), TST (Tribunal Superior do Trabalho), dentre outros.

O termo "lei" tem um sentido amplo, referindo-se a um conjunto de normas de um ordenamento jurídico como: Constituição, Leis complementares, Leis Ordinárias etc. Ademais, segundo Paulo José da Costa Jr:

> Jurisprudência e Interpretação são, como não poderia deixar de ser, conceitos estreitamente coligados. Ao Julgar, o Tribunal aplica o Direito. Para aplicá-lo, necessita antes de mais nada interpretá-lo. Estreito o relacionamento entre jurisprudência e lei, entre "interpretatio" e "jus dicere". Desse modo, a jurisprudência configura a interpretação judiciária do Direito vigente. A aplicação da lei interpretada às relações humanas, no instante em que são elas concretamente regulamentadas pelo Direito. (FRANCO, pág. 5)

[2] Série de acórdãos dos tribunais sobre interpretação do mesmo preceito jurídico e sua aplicação em face de fatos análogos. Em sentido abstrato, é a própria Ciência do direito. SIDOU, pág 470.

No tocante à competência[3], quem pode expedir jurisprudência, ou entendimento jurisprudencial, é aquele que possui jurisdição, isto é, aqueles que podem "dizer o Direito", mais precisamente, magistrados constitucionalmente competentes, sejam: juízes, desembargadores ou ministros.

A utilidade da jurisprudência é percebida em dois principais aspectos, o primeiro, orientar as decisões subsequentes à publicação da jurisprudência, no sentido de uma uniformização que visa, dentre outras coisas, garantir a Segurança Jurídica. Nesse sentido é que se assevera a jurisprudência também ser uma fonte do direito[4], pois demonstram como certo tribunal vem entendendo acerca de determinado assunto. O segundo, para suprir eventuais "lacunas da lei", ou seja, situações em que o legislador não foi muito claro durante a elaboração do dispositivo legal.

Essas decisões são divididas em 3 tipos:

- Sentença: Decisão de um juiz em primeira instância;
- Decisão monocrática: Proferida por um único juiz, em segunda instância, sobre um tema onde já existe uma interpretação uniforme.
- Acórdão: Decisão de um colegiado de juízes, ou seja, uma turma, uma corte.

Ainda sobre uniformização de decisões, vale destacar que jurisprudências reiteradas sobre determinados assuntos podem vir a virar súmulas, isto é, entendimento consolidado de determinado Tribunal.

Todavia, dizer que existe uma jurisprudência, ou uma súmula, acerca de determinada situação, não implica necessariamente que o julgador original deverá seguir o seu conteúdo. Pode acontecer de uma decisão, contrariar um entendimento anteriormente pacificado por um tribunal, o que, apesar de possível, acarreta conflitos judiciais que normalmente terminam em grau de recursos nos tribunais superiores para dirimirem estas celeumas jurídicas.

Pensando nisso, foi acrescentado à Constituição Federal de 1988 (BRASIL, 1988), através da Emenda Constitucional nº45 de 2004, o artigo 103-A que prevê a aprovação pelo

[3] A competência residual da Justiça Comum, mesmo que excluídas as causas penais, é bastante ampla, abrangendo direito privado e direito público, em suas inúmeras vertentes (p. ex., tributário, societário, previdenciário, consumerista). NEVES, pág 59.

[4] A jurisprudência adquiriu novos contornos e importância no cenário jurídico-penal, passando a ser fonte imediata reveladora de direito. É o que ocorre de forma evidente com as súmulas vinculantes, a exemplo do verbete de nº 24, que disciplina a atipicidade de crime contra a ordem tributária quando pendente o lançamento definitivo do tributo. CUNHA, pág. 61/62.

Supremo Tribunal Federal das chamadas Súmulas Vinculantes sobre reiteradas decisões de matéria constitucional. A finalidade destas Súmulas Vinculantes está expressa no §1º do mencionado artigo que menciona:

> "A súmula terá por objetivo a validade, a interpretação e a eficácia de normas determinadas, acerca das quais haja controvérsia atual entre órgãos judiciários ou entre esses e a administração pública que acarrete grave insegurança jurídica e relevante multiplicação de processos sobre questão idêntica.".

Ao contrário das súmulas e jurisprudências puras e simples, a Súmula Vinculante, como o próprio nome sugere, vincula as decisões posteriores à sua publicação, oriundas tanto do Poder Judiciário, quanto da Administração Pública e assim, uma vez contrariada, ocasiona a anulação do ato administrativo ou cassação da decisão judicial reclamada.

2.2. Trabalhos relacionados

Com o objetivo de minimizar a complexidade computacional do aprendizado através da representação vetorial contínua de palavras, (Mikolov, 2013) propôs o ***Word2Vec***, um método que recebe uma grande massa de dados de textos como entrada e retorna uma série de vetores de palavras (*word embeddings*) na saída. Para produção desses vetores, foram definidos 2 modelos complementares: um que busca predizer o contexto - ou palavras do entorno - a partir de uma palavra escolhida em um texto, e outro que é inverso, no sentido em que busca prever a palavra central a um contexto a partir deste último. A essas duas estruturas de predição complementares são frequentemente associados os conceitos de *skip-gram* e *Continuous Bag of Words* (*CBOW*).

Uma variação do *Word2Vec* é **Paragraph Vector** (Le, 2014), também conhecido como *Doc2Vec*, que é um algoritmo de aprendizado não supervisionado que aprende representações vetoriais através de pedaços de texto de tamanho variável como frases, sentenças e documentos. Concatenando o vetor de parágrafos com vetores de palavras e, considerando a semântica das palavras, mantendo a ordem das palavras, é possível predizer uma palavra a partir do seu contexto.

Uma outra extensão do *Word2Vec* é o **FastText** (Bojanowski et al., 2016), que é baseado no modelo *Skip-gram* e leva em consideração a morfologia das palavras, ou seja, a

sua estrutura interna. Cada palavra é representada como um conjunto de n-grams de caracteres, formados a partir da combinação das palavras e sub-palavras, representando palavras como sendo a soma dos seus *n-grams*.

Com a criação de um modelo específico ponderado dos mínimos quadrados, treinado sobre a contagem global de co-ocorrência das palavras, foi desenvolvido o **GLOVE** (Pennington, 2014), treinando apenas sobre os elementos não nulos da matriz de co-ocorrência de palavras, não sendo necessário percorrer toda a matriz ou procurar por uma janela de contexto específica em todo corpus.

Uma outra técnica é o **Item2Vec** (Barkan,2016) onde, através do treinamento de um modelo de word embedding, utilizando o modelo *Skip-gram* do *word2vec*, a partir da análise das relações de co-ocorrência entre palavras (itens) em diferentes contextos, são construídos vetores de itens. Cada vetor de itens representa as informações de um contexto específico.

Além das técnicas mencionadas acima existem muitas outras baseadas no *Word2Vec*, a exemplo do *node2vec* (Grover, 2016), *graph2vec* (Prieto, 2017), *topic2vec*(Niu, 2016), *med2vec* (Sun, 2016), *sense2vec* (Trask, 2015), etc.

2.3. Métodos utilizados

2.3.1. MDS (*Multidimensional scaling*)

Técnica estatística originária da psicometria, que é uma área da psicologia ligada à matemática aplicada. É uma das técnicas de análise dimensional mais utilizada (Gruenhage, 2016).

Possibilita o mapeamento de um conjunto de pontos de um espaço multidimensional de alta dimensão em um espaço de baixa dimensão, de forma que as relações de distâncias entre os pontos no espaço projetado igualem a similaridade (ou dissimilaridade) entre esses pontos, lembrando que, nem sempre tais relações são perfeitamente representáveis em espaços reduzidos, muitas vezes são obtidas apenas aproximações (no caso de espaços com dimensão menor que a necessária).

No MDS, uma matriz de distâncias entre pares de objetos quaisquer é imposta inicialmente, sendo que a escolha da medida de distância é arbitrária (e.g. em ciências

humanas, não é incomum que essas distâncias sejam estabelecidas subjetivamente, através de formulários de consulta a voluntários).

O MDS é utilizado para reconstruir pontos a partir de pares de distâncias. Assim, dada a matriz de distâncias entre N objetos, são criados N vetores numéricos com dimensão D, e uma nova matriz de distâncias é computada entre os pares de vetores. Vale notar que as distâncias entre vetores são igualmente arbitrárias, embora a distância euclidiana seja usualmente escolhida.

Dessa maneira, a posição dos pontos no espaço de dimensão D é ajustada gradativamente até que a discrepância entre as duas matrizes de distância seja mínima, de acordo com o critério de comparação entre essas matrizes, usualmente chamado de '*stress*' (*Standardized Residual Sum of Squares*).

Por consequência, se as duas matrizes forem construídas com a mesma medida de distância, então, para D maior ou igual à dimensão subjacente à representação dos objetos originais, o 'stress' esperado é nulo. Por outro lado, é possível se induzir representações em dimensão D cada vez menores, ao preço de medidas de '*stress'* cada vez maiores.

Um exemplo prático de aplicação do MDS ocorre quando, a partir das distâncias entre cidades, se pretende gerar pontos e plotá-los num gráfico para facilitar a visualização da localização dessas cidades. Considerando 10 cidades: São Paulo, Rio de Janeiro, Manaus, Natal, Porto Alegre, Brasília, Salvador, Rio Branco, Cuiabá e Belo Horizonte. A tabela 1 representa a matriz simétrica que foi construída para representar as distâncias aproximadas entre elas em quilômetros.

	SP	RJ	MA	NA	PA	BR	SA	RB	CB	BH
SP	0	348	2689	2322	850	875	1455	2704	1326	490
RJ	348	0	2849	2086	1122	934	1210	2982	1576	340
MA	2689	2849	0	2765	3131	1932	2606	1149	1452	2556
NA	2322	2086	2765	0	3172	1775	876	3617	2524	1832
PA	850	1122	3131	3172	0	1619	2302	2814	1679	1340
BR	875	934	1932	1775	1619	0	1060	2247	874	625

SA	1455	1210	2606	876	2302	1060	0	3206	1915	965
RB	2704	2982	1149	3617	2814	2247	3206	0	1414	2786
CB	1326	1576	1452	2524	1679	874	1915	1414	0	1372
BH	490	340	2556	1832	1340	625	965	2786	1372	0

Tabela 1: Matriz simétrica com as distâncias entre as dez cidades

Fonte: Elaborada pelo autor.

A título de ilustração, aplicou-se o MDS a essa matriz de distâncias para que se encontrassem os pontos num espaço de duas dimensões (2D) com distâncias aproximadas às informadas na matriz de distâncias e o resultado foi apresentado na imagem abaixo.

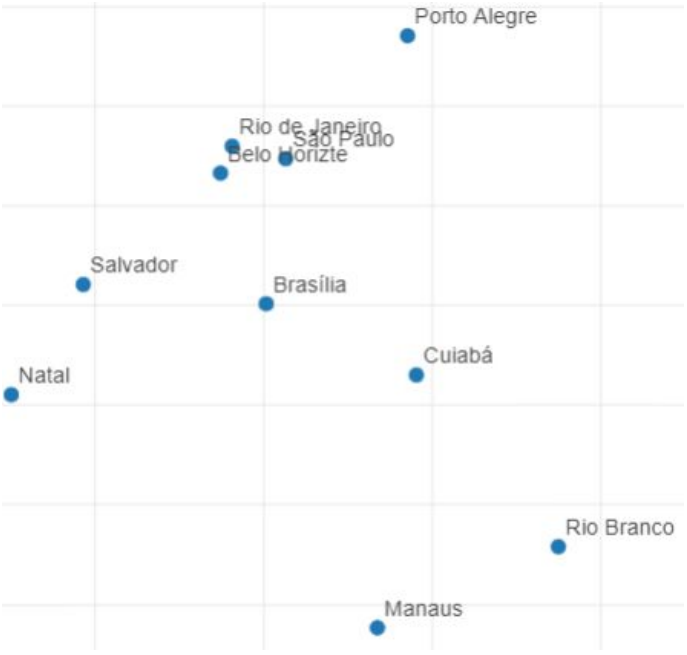

Figura 1: Pontos no espaço 2D com distâncias aproximadas aos valores da matriz de distâncias

Fonte: Elaborada pelo autor.

Apesar de os pontos terem sido gerados no sentido contrário, isto é, de cabeça para baixo, em relação ao que estamos acostumados a visualizar o mapa do Brasil, percebe-se que as posições se aproximam bastante das posições reais representadas no mapa. Além disso, a disposição aparentemente invertida dos pontos ilustra uma característica inerente ao MDS,

que é a arbitrariedade de orientação dos pontos ajustados, devido à invariância da matriz de distâncias à rotação.

2.3.2. Word2Vec

Proposto por (Mikolov, 2013), é utilizado para calcular a representação vetorial de palavras, de forma que, medidas de similaridade entre os vetores que representam as palavras sejam numericamente semelhantes a medidas probabilísticas associadas aos contextos dessas palavras.

O método Word2Vec reduziu a complexidade computacional do aprendizado das representação distribuídas de palavras, sendo capaz de capturar aspectos semânticos e sintáticos com uma granularidade fina, por meio de operações aritméticas vetoriais simples.

Um conjunto de textos é apresentado como entrada ao word2vec e, como saída, é criada uma representação distribuída para esses dados (vetores com características das palavras do corpus apresentado), preservando-se suas relações semânticas com as palavras ao seu redor (contexto).

Existem duas arquiteturas de implementação para essa técnica, o CBOW e o *Skip-gram*.

2.3.2.1. Continuous bag of words

Dada uma palavra em uma sentença (palavra central ou palavra alvo), o CBOW utiliza o contexto, ou palavras próximas como "entrada" para tentar prever a probabilidade dessa palavra alvo ocorrer. A ordem das palavras não influencia na predição final.

Por exemplo: Sendo palavra(pos) a palavra central e, considerando uma janela de contexto de tamanho 5, a entrada será composta pelas palavras da posição palavra(pos-2), palavra(pos-1), palavra(pos+1) e palavra(pos+2), conforme demonstrado na figura 2:

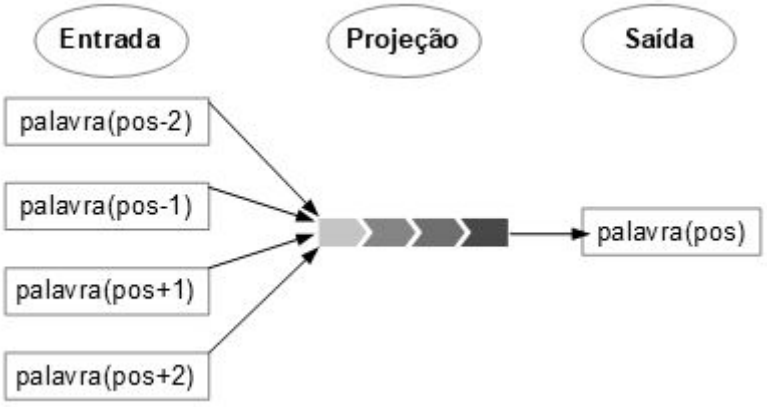

Figura 2: Funcionamento do CBOW

Fonte: Elaborada pelo autor.

2.3.2.2. Skip-gram

Conforme demonstrado na figura 3, esse método recebe uma palavra alvo como "entrada" e tenta prever o seu contexto. A partir da palavra palavra(pos), apresentada como "entrada do método", procura maximizar a predição de quais palavras (palavra(pos-2), palavra(pos-1), palavra(pos+1) e palavra(pos+2)) apareceram próximas à palavra de "entrada".

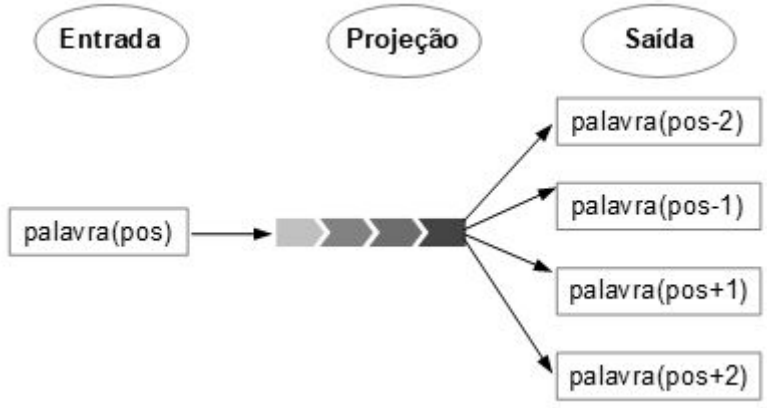

Figura 3: Funcionamento *Skip-gram*

Fonte: Elaborada pelo autor.

2.3.3. Doc2Vec

Proposto por (Le, 2014), é um método não supervisionado que "aprende" as representações distribuídas a partir de textos não estruturados de qualquer tamanho (frases, parágrafos ou documentos).

É uma extensão do Word2Vec onde o "aprendizado" vem tanto das palavras que compõem cada documento (relacionamento entre as palavras) quanto a partir dos próprios documentos (contexto onde as palavras estão inseridas). Os vetores de palavras são comuns a todos os parágrafos e cada parágrafo do documento é mapeada em um vetor de parágrafos.

Existem duas abordagens para se construir os vetores de parágrafos, o Modelo Distribuído de *Bag-Of-Words* para Vetores de Parágrafos (PV-DBOW) e o Modelo de Memória Distribuída de Vetores de Parágrafos (PV-DM).

2.3.3.1. PV-DBOW

É semelhante ao modelo *Skip-gram* e não leva em consideração a ordem das palavras. Utiliza os vetores de parágrafos concatenados aos vetores de palavras para prever as palavras seguintes, como exibido na figura seguinte. Outra forma de funcionamento é, ignorando as palavras do contexto de entrada, forçar o modelo a prever palavras aleatoriamente com base no parágrafo de saída

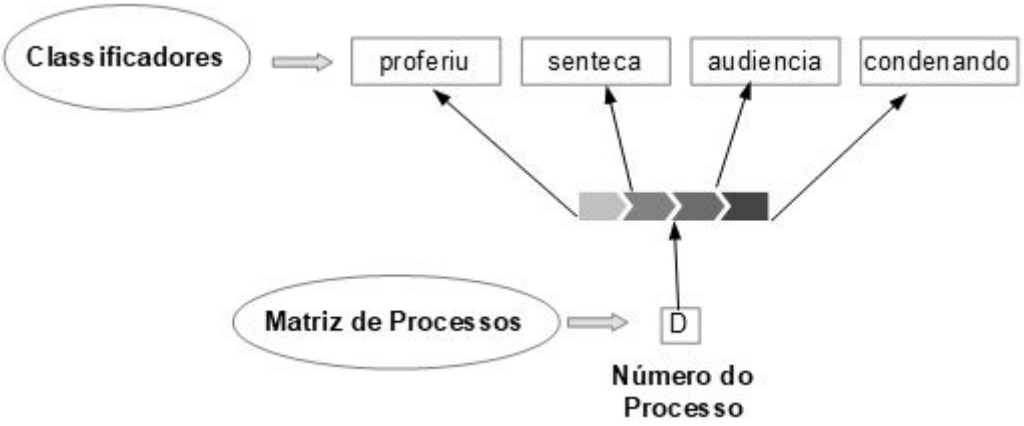

Figura 4: Funcionamento do PV-DBOW

Fonte: Elaborada pelo autor.

2.3.3.2. PV-DM

É um método semelhante ao modelo CBOW e leva em consideração a ordem das palavras. O contexto semântico de onde a palavra está inserida é considerado, o que ajuda a manter o nível de proximidade entre as palavras.

A "matriz de processos (documentos)" atua como uma memória, que captura o que está faltando no contexto atual e é compartilhado entre todos os contextos gerados apenas pelo mesmo parágrafo, não sendo compartilhamento entre parágrafos. Como exemplificado na figura a seguir.

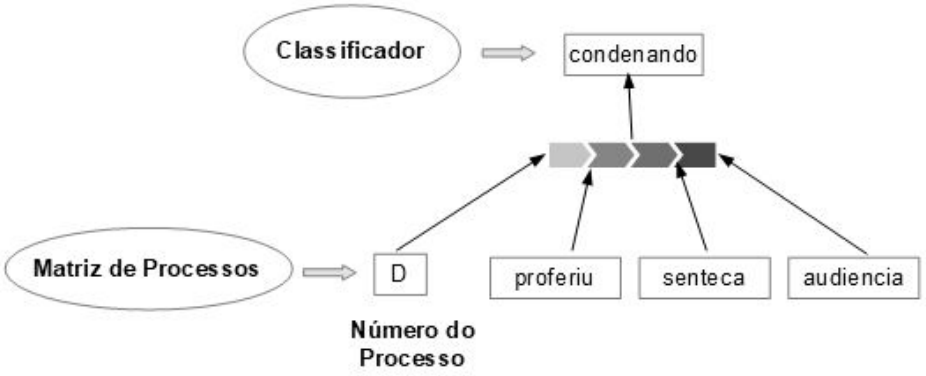

Figura 5: Funcionamento do PV-DM

Fonte: Elaborada pelo autor.

2.4. Ambiente

Os treinamentos e testes foram executados em um computador da marca “Lenovo” que possui as seguintes características:

Item	Descrição
Processador	Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHz (20 núcleos)
Memória	256GB de RAM
Sistema Operacional	Linux CentOS 7

Tabela 2: Configuração do servidor utilizado para os testes dos experimentos

Fonte: Elaborada pelo autor.

Para implementação e execução dos experimentos foi utilizada a linguagem de programação Python 3.7 com as bibliotecas Numpy 1.15, NLTK 3.3, Gensim 3.7 e Bokeh 0.13.

2.5. Métricas para avaliação dos resultados

Foram definidas duas métricas

- Analisar os resultados obtidos e avaliar se as semelhanças encontradas entre os processos equivalem aos tipos dos processos. Esses tipos são rotulados manualmente por especialistas, no momento em que os processos são cadastrados;
- Comparar os resultados obtidos com o método proposto, para o mesmo conjunto de documentos, aos resultados encontrados com o Doc2Vec, em termos de agrupamento de documentos.

3. METODOLOGIA

3.1. Vocabulário jurídico (tesauro)

Além do conteúdos dos processos, utilizamos também o vocabulário jurídico (tesauro, 2019), acessíveis através do site do Supremo Tribunal Federal. Esse tesauro nada mais é do que um conjunto controlado de palavras utilizadas na área jurídica, visando a padronização da informação.

Atualmente existe uma área específica no STF formada por aproximadamente 70 pessoas responsáveis por avaliar documentos jurídicos e extrair informações relevantes que possam contribuir com o aprimoramento do tesauro, ajudando cada vez mais os profissionais que a utilizam. Logo a seguir exibimos um exemplo de consulta ao tesauro.

Figura 6: Exemplo de consulta ao tesauro, buscando o termo “norma penal em branco”

Fonte: Elaborada pelo autor.

O tesauro possui a seguinte estrutura:

- **Descritor:** Termo escolhido para representar um conceito no tesauro (indexador)
- **Não-descritor**: Descreve o mesmo conceito do descritor mas não é utilizado na indexação
- **Nota:** Definição ou orientação de como utilizar o termo
- **Termo genérico (TG):** Termo mais abrangente respeitando uma hierarquia entre os termos
- **Termo específico (TE):** Termos subordinados ao genérico ou cadeia hierárquica
- **Termo relacionado (TR):** Indica uma relação entre os termos que não formam uma hierarquia, mas que são associados mentalmente
- **Categoria (CAT):** Dividida em 3 ramos:
    - **Ramos do direito:** Constitucional, civil, etc.
    - **Especificadores:** Agrupam termos que restringem o conceito de um descritor.
    - **Identificadores:** Agrupam nomes de pessoas, instituições, etc.

Após a avaliação das informações do tesauro, foi possível realizar o levantamento de todos os termos principais e seus respectivos termos associados, conforme ilustração a seguir:

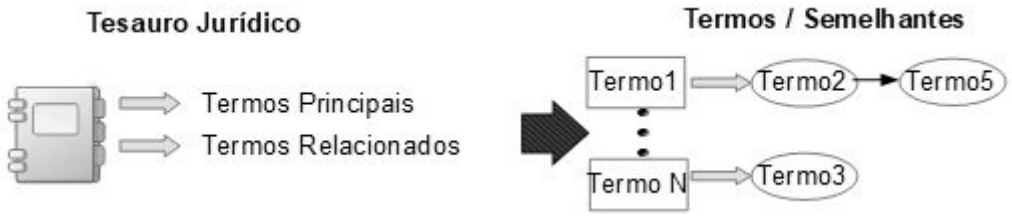

Figura 7: Tratamento dos termos do tesauro jurídico

Fonte: Elaborada pelo autor.

3.2. Base de jurisprudência

Os dados utilizados neste trabalho são dados de domínio público, acessíveis através do site do Tribunal de Justiça do Estado de Sergipe (TJSE), referentes aos julgamentos realizados por este tribunal.

A base utilizada foi disponibilizados por (Oliveira, 2018), que publicou 119 arquivos XML (*eXtensible Markup Language*), de aproximadamente 30MB cada, extraídos da base de dados jurisprudencial do Tribunal de Justiça do Estado de Sergipe, em Setembro de 2016.

Cada um desses arquivos armazena informações referentes a quatro tipos de processos:

- Acórdãos do Segundo Grau (181.994 processos);
- Decisões monocráticas do Segundo Grau (37.044 processos);
- Acórdãos da Turma Recursal (37.161 processos);
- Decisões monocráticas da Turma Recursal (23.149 processos).

A figura 8 exemplifica os dados de um processo extraído de um dos arquivos xml da base de jurisprudência utilizada.

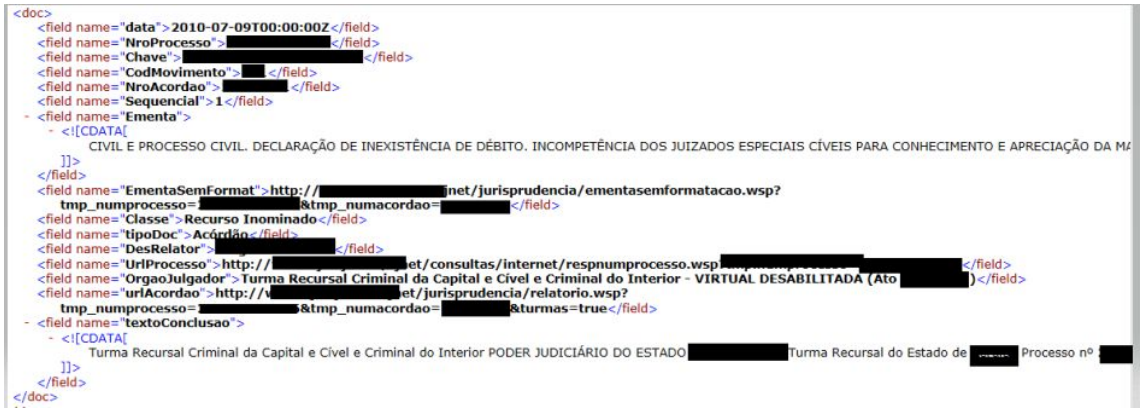

Figura 8: Exemplo de conteúdo processual extraído de um arquivo xml pertencente á base jurisprudencial

Fonte: Elaborada pelo autor.

3.3. Tipos dos processos

Assim que um novo processo é criado, o mesmo é categorizado em alguns tipos pré-definidos de processos. Durante todo o trâmite processual, o processo é analisado por analistas e técnicos judiciários e administrativos, advogados, juízes, promotores, etc. Podendo

essa categorização ser reavaliada e alterada caso algum(uns) dos interessados tenha(m) uma opinião diferente da classificação original.

Logo abaixo foram exibidos alguns exemplos de processos, seguidos por seus respectivos tipos. Atentar para o fato de que um processo pode possuir mais de um tipo.

Nr. Processo	**Tipo**	
201XXXXXXXXX	ASSISTÊNCIA JUDICIÁRIA GRATUITA	INDENIZAÇÃO POR DANO MORAL
201XXXXXXXXX	ASSISTÊNCIA JUDICIÁRIA GRATUITA	SEGURO
201XXXXXXXXX	ASSISTÊNCIA JUDICIÁRIA GRATUITA	ANTECIPAÇÃO DE TUTELA / TUTELA ESPECÍFICA
201XXXXXXXXX	JUROS DE MORA - LEGAIS / CONTRATUAIS	ASSISTÊNCIA JUDICIÁRIA GRATUITA
201XXXXXXXXX	INDENIZAÇÃO POR DANO MORAL	LEI DE IMPRENSA
201XXXXXXXXX	INCLUSÃO INDEVIDA EM CADASTRO DE INADIMPLENTES	ANTECIPAÇÃO DE TUTELA / TUTELA ESPECÍFICA
201XXXXXXXXX	TELEFONIA	ABATIMENTO PROPORCIONAL DO PREÇO
201XXXXXXXXX	ASSISTÊNCIA JUDICIÁRIA GRATUITA	INDENIZAÇÃO POR DANO MORAL

201XXXXX XXXX	INDENIZAÇÃO POR DANO MORAL	
201XXXXX XXXX	ASSISTÊNCIA JUDICIÁRIA GRATUITA	SEGURO
201XXXXX XXXX	PAGAMENTO INDEVIDO	INDENIZAÇÃO POR DANO MORAL
201XXXXX XXXX	ASSISTÊNCIA JUDICIÁRIA GRATUITA	SEGURO
201XXXXX XXXX	NULIDADE	HONORÁRIOS ADVOCATÍCIOS
201XXXXX XXXX	PERDAS E DANOS	SEGURO
201XXXXX XXXX	NULIDADE	PROVAS
201XXXXX XXXX	INDENIZAÇÃO POR DANO MORAL	INDENIZAÇÃO POR DANO MATERIAL
201XXXXX XXXX	INTERPRETAÇÃO / REVISÃO DE CONTRATO	CARTÃO DE CRÉDITO
201XXXXX XXXX	ASSISTÊNCIA JUDICIÁRIA GRATUITA	HONORÁRIOS ADVOCATÍCIOS

Tabela 3: Exemplos de processos e seus tipos associados

Fonte: Elaborada pelo autor.

3.4. O pré-processamento

Na fase de pré-processamento dos documentos, todos os processos foram analisados e foram extraídos apenas os campos: chave, tipo do processo e os campos textuais. Após a seleção dessas informações, todo texto foi convertido para minúsculo e os símbolos de acentuação e caracteres especiais foram removidos.

Partindo do princípio de que palavras muito frequentes como "a" ou "um" não agregam muito valor como características de contextos (Baroni et al., 2014) e que, é possível se obter bons resultados com a eliminação de *stop-words*, uma vez que sua remoção não afeta a ordem ou o significado das demais palavras (Lilleberg, 2015) (Rahmawati, 2016), optamos pela remoção das *stop-words* nessa fase de pré-processamento. A lista de *stopwords* da biblioteca NLTK[5] do python foi utilizada, sendo composta por 204 *stopwords* em português.

A figura 9 resume todas as etapas realizadas durante essa fase de pré-processamento, bem como os processos e seus respectivos conteúdos textuais tratados, obtidos após a sua execução.

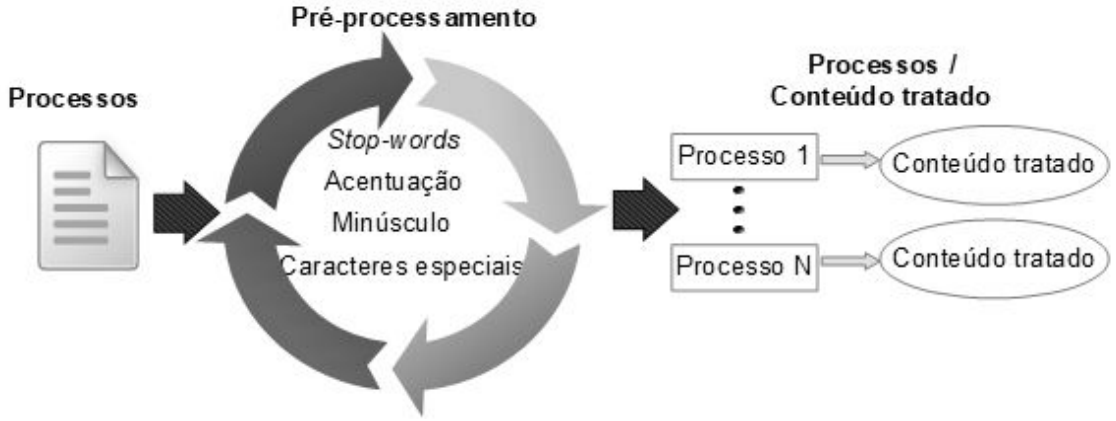

Figura 9: Pré-processamento dos dados processuais

Fonte: Elaborada pelo autor.

[5] É um conjunto de ferramentas Python, utilizado no processamento de linguagem natural, desenvolvido pelo curso de linguística computacional do Departamento de Ciência da Computação e Informática da Universidade da Pensilvânia. A relação de *stopwords* (em português) utilizada pertence ao pacote NLTK.corpus e está disponível através do endereço: http://www.nltk.org/nltk_data/

3.5. Método proposto

Superada a fase de tratamento dos dados, foi realizada uma busca (expressões regulares[6]) pela sua fundamentação legal, ou seja, referência a artigos, leis, súmulas, etc, que embasaram as argumentações dos processos analisados. A essa fundamentação demos o nome de "Entidades Jurídicas" (EJs).

Com a elaboração dessas expressões regulares foi possível extrair características de:

- **Artigos** - como número do artigo, leis referenciadas e indicação de onde encontrar mais informações sobre as mesmas: Código de Defesa do Consumidor (CDC), Constituição Federal (CF), Código Civil (CC), Código Penal Brasileiro (CPB), Código Processual Penal Brasileiro (CPPB) etc.
- **Súmulas** - Extração do número identificador da súmula
- **Apelações Cíveis** - Identificador do número da apelação Cível de referência
- **Agravos Regimentais** - Identificação do Agravo Regimental
- **Medidas Provisórias** - Identificador
- **Apelações em Geral** - Código identificador

Todas as ocorrências dessas entidades jurídicas foram formatadas para um padrão específico e adicionadas aos contextos dos seus respectivos processos. As referências a "artigos" foram substituídas pela letra A concatenada com o número do artigo e ao seu identificador de origem (ex: CPC, Código de Processo Civil); as referências a súmulas pela letra S concatenada ao número da súmula; e às leis pela letra L concatenada ao número da lei e ao artigo de referência, conforme demonstrado na tabela 4.

Padrão encontrado	Padrão formatado
art. 48, parágrafo único, lei 9.099	A48L9099
ARTIGO 333, II DO CPC	A333CPC

[6] Expressão regular é um tipo de texto-padrão utilizado para verificar se determinado texto se encaixa em um padrão específico. Quando utilizadas da forma correta, podem simplificar muitas tarefas de programação e processamento de texto (Goyvaerts, 2011)

súmula nº 296	S296
AgRg no REsp 994.910/MG	AGRG994910MG
MP nº. 2.172	MP2172
Ac. nº 1099	AC1099

Tabela 4: Exemplos de formatação dos "Elementos Jurídicos" encontrados

Fonte: Elaborada pelo autor.

Com o levantamento dessas entidades de cada um dos processos, foi iniciada a fase de implementação dos métodos estudados através da codificação vetorial das "Entidades Jurídicas" contidas em cada texto e utilização desses vetores para geração da representação semântica resultante de cada texto jurídico. A tabela seguinte exibe exemplos de conteúdo textual e suas respectivas entidades jurídicas extraídas de cinco processos da base.

Processo	**Conteúdo textual**	**Entidades Jurídicas**	**Tipo**
20090080 1883	...ART. 24 DA LEI Nº 12.016... ...art. 8º da Lei 9.099... ...artigo 47, parágrafo único do CPC.... ...e Súmula 631 ...	A24L12016, A8L9099, A47CPC, S631	REPRESENTAÇÃO EM JUÍZO
20100090 1107	...no art. 3º, alínea b, da Lei nº 6.194... ...O ART. 515, § 3º DO CPC.... ...do artigo 206 do Código Civil.... ...À MEDIDA	A3L6194, A515CPC, A206CC, MP340, S14	SEGURO

	PROVISÓRIA Nº 340… ...balizada da Súmula nº. 14...		
2010009 01119	...no art. 55, 2ª parte, da Lei nº 9.099... ...da Medida Provisória nº. 451....	A55L9099, MP451	SEGURO
2010009 01261	...DO ART. 405 DO CC... ...no art. 97 da Constituição Federal… ...do artigo 8º da Lei 11.482...	A405CC, A97CF, A8L11482	CORREÇÃO MONETÁRIA, JUROS DE MORA - LEGAIS / CONTRATUAIS, ASSISTÊNCIA JUDICIÁRIA GRATUITA, SEGURO
2010009 01183	...DO ART. 405 DO CC… ...art. 97 da CF... ...o artigo 3º da Lei 6.194....	A405CC, A97CF, A3L6194	CORREÇÃO MONETÁRIA, JUROS DE MORA - LEGAIS / CONTRATUAIS, ASSISTÊNCIA JUDICIÁRIA GRATUITA, SEGURO

Tabela 5: Exemplo de entidades jurídicas extraídas de cinco processos exemplo

Fonte: Elaborada pelo autor.

Com o conteúdo dos processos e a relação de semelhança entre os termos do tesauro Jurídico, foi possível identificar uma maior relação entre as palavras e expressões, uma vez que, mesmo uma palavra não tendo sido encontrada no texto, a informação do tesauro possibilitou o levantamento dos seus "Termos Relacionados", auxiliando na busca de semelhança entre os processos.

Para explicar melhor o papel dos termos do tesauro, vamos considerar o seguinte exemplo: Dois processos A e B possuem, dentre outros termos, os seguintes termos e suas respectivas quantidade de ocorrências desses termos em cada processo, conforme tabela abaixo:

Processos	**Termos do tesauro / Quantidade de ocorrência**
A	sofrimento psíquico: 6
B	dano psicológico: 10, violência psicológica: 4

Tabela 6: Exemplo de termos do tesauro e suas respectivas quantidades de ocorrência extraídos de dois processos

Fonte: Elaborada pelo autor.

Conforme exemplo de consulta ao tesauro exibido na figura 10, o termo “dano psicológico” está relacionado aos termos “integridade psíquica”, “sofrimento psíquico” e “violência psicológica”.

Figura 10: Exemplo de consulta ao tesauro, procurando pelo termo "dano psicológico"

Fonte: Elaborada pelo autor.

Caso a comparação fosse realizada apenas pelos termos, não seria encontrada nenhuma semelhança entre os processos A e B. Com a utilização do tesauro, foi possível detectar o relacionamento entre os termos e identificar que todos os três termos listados se referem a assuntos semelhantes, identificando novas características para comparação entre os processos. A tabela 7 demonstra exemplos de termos do tesauro extraídos de cinco processos exemplo.

Processos	**Termos do tesauro / Quantidade de ocorrência**
200900801883	impetrante: 6, litisconsorte: 7, mandado seguranca: 10, seguranca: 10
201000901107	acidente: 9, dpvat: 15, honorarios: 5, indenizacao: 38, invalidez: 33, pericia: 5, seguradora: 7, seguro: 19, seguro obrigatorio: 9
201000901119	acidente: 7,dpvat: 10, indenizacao: 36, invalidez: 32, invalidez permanente: 23, prova pericial: 6, seguradora: 6

201000901261	correcao: 10, dpvat: 10, indenizacao: 27, juros mora: 5, pagamento: 12, quitacao: 5, recorrente: 5, segurado: 6, seguro obrigatorio: 5, sinistro: 5
201000901183	correcao: 10, correcao monetaria: 8, dpvat: 10, indenizacao: 25, juros: 8, juros mora: 5, morte: 6, pagamento: 12, segurado: 6, seguro: 26

Tabela 7: Exemplos de termos do tesauro extraídos de cinco processos exemplo

Fonte: Elaborada pelo autor.

Para calcular a semelhança entre os processos, foi criada uma matriz simétrica M para representar o grau de similaridade entre todos os processos ("distância" entre os processos). Essa similaridade foi calculada a partir da comparação entre os contextos dos processos, partindo do princípio de que quanto maior a quantidade de Entidades Jurídicas e termos do tesauro comuns aos dois contextos, menor será a distância entre esses processos e, consequentemente, maior será a similaridade entre os contextos comparados.

Cada linha e coluna dessa matriz de distância, M, contidas na base estudada, onde Mij representa a força do relacionamento entre o contexto do processo "i" (cp_i) e o contexto do processo "j" (cp_j). Essa métrica de associação foi elaborada baseada na associação entre palavras e contextos, que é muito comum na literatura do PLN e similaridade entre palavras (Levy, 2014b).

Com base no coeficiente de Jaccard[7], que compara o número de itens comuns com o número total de itens excluindo-se os itens comuns, (Levy, 2014b) conseguiu descrever uma métrica semelhante para o cenário "palavra-contexto". Apesar de essa métrica não ser definida explicitamente, pelo fato de ter sido descrita em detalhes, foi possível entendê-la e adaptá-la ao cenário "processo-processo" utilizado neste trabalho, chegando à seguinte fórmula:

$$M_{(i,j)} = 1 - \frac{|(cp_i \cap cp_j)|}{|(cp_i \cup cp_j)|} \qquad (1)$$

[7] Coeficiente de Jaccard = a / (a+b+c), onde "a" é o número de itens encontrados em A e B; "b" é o número de itens do local B, que não pertencem a A; "c" é o número de itens do local A, que não pertencem a B; (Jaccard, 1908).

A célula $M_{(i,j)}$ armazena o valor que representa a quantidade de elementos (cardinalidade, representada pelo símbolo "| |") comuns aos dois contextos, dividido pela soma da quantidade de elementos que pertencem a qualquer um dos contextos. Sendo essa mesma métrica utilizada tanto para comparação baseada em Ejs quanto para baseada nos termos do tesauro.

Para comparação entre os processos, levando em consideração as suas Entidades Jurídicas, foi utilizada a seguinte métrica:

$$semelhancaEJ = 1 - \frac{qtdItensIntersec}{qtdEjProcA + qtdEjProcB - qtdItensIntersec} \quad (2)$$

onde qtdEjProcA e qtdEjProcB são as quantidade totais de EJ de dois processos quaisquer A e B respectivamente e qtdItensIntersec é a quantidade de EJ pertencentes aos dois processos, levando em consideração a quantidade de vezes que essas EJ apareceram nos dois processos comparados.

Com o cálculo da semelhança baseado em EJs, quanto mais próximo de zero for o resultado. maior será a semelhança entre os processos e, quanto mais próximo de 1, mais diferentes eles serão. Ao final dessa etapa, após a análise dos dados e avaliação dos resultados, verificou-se que os 5% (por linha) menores valores de semelhança baseado em EJs representavam os processos mais semelhantes, ou seja, dentre os valores da linha da matriz de similaridade por EJ, foram selecionados os menores valores (referentes a 5% do total de processo da amostra).

A ilustração abaixo resume o papel das entidades jurídicas nesse processo, que vai desde a sua extração e a construção da matriz de similaridade entre os processos com base nas suas EJs, utilizando a equação (2), exibida anteriormente, até a seleção dos processos mais semelhantes com base nas EJs que possuem.

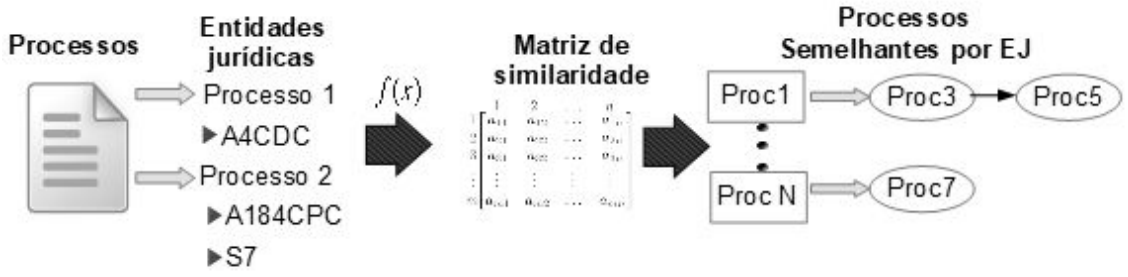

Figura 11: Extração das EJs dos processos, construção da matriz de similaridade e seleção dos processos semelhantes

Fonte: Elaborada pelo autor.

Conforme mencionado anteriormente, essa mesma métrica foi utilizada na comparação entre processos levando em consideração os termos do tesauro:

$$semelhancaTermos = 1 - \frac{qtdItensIntersec}{qtdTermosProcA + qtdTermosProcB - qtdItensIntersec} \qquad (3)$$

onde qtdTermosProcA e qtdTermosProcB são as quantidade totais de palavras e expressões que aparecem em dois processos quaisquer A e B respectivamente e qtdItensIntersec é a quantidade de termos comuns aos dois processos levando em consideração a quantidade de ocorrência desses termos nos dois processos comparados.

Essa fórmula mantém a mesma relação da anterior, quanto mais próximo de zero for o resultado, mais semelhantes os processos serão e, quanto mais próximo de 1, menos relacionado estará o conteúdo deles. Sendo considerados também, dentre os valores da linha da matriz de similaridade por termos, apenas os menores valores (referentes a 5% do total de processo da amostra).

A figura seguinte resume o papel dos termos do tesauro nesse processo, que vai desde a sua extração e construção da matriz de similaridade entre os processos, com base nos seus termos, utilizando a equação (3), exibida anteriormente, até a seleção dos processos mais semelhantes com base nos termos que possuem.

Figura 12: Extração dos termos dos processos, construção da matriz de similaridade e seleção dos processos semelhantes

Fonte: Elaborada pelo autor.

Após o cálculo das duas comparações acima, uma média aritmética é aplicada entre os dois valores, sendo o resultado final um valor que representa a semelhança entre os processos comparados:

$$semelhancaProcessos = \frac{semelhancaEJ + semelhancaTermos}{2} \qquad (4)$$

Realizando esse cálculo com alguns processos de exemplo, obtivemos a seguinte matriz de distâncias:

Processos	20090080188 3	20100090110 7	20100090111 9	20100090126 1	20100090118 3
20090080188 3	0.	0.97	0.96	0.96	0.96
20100090110 7	0.97	0.	0.2	0.68	0.72
20100090111 9	0.96	0.2	0.	0.67	0.7
20100090126 1	0.96	0.68	0.67	0.	0.13
20100090118 3	0.96	0.72	0.7	0.13	0.

Tabela 8: Exemplo de matriz de distâncias calcula com base nas EJs e termos

Fonte: Elaborada pelo autor.

A imagem a seguir demonstra a construção da matriz de distâncias, levando em consideração os processos semelhantes com base nas suas EJs e os termos do tesauro, utilizando a equação (4), exibida anteriormente, para construção da matriz de distâncias.

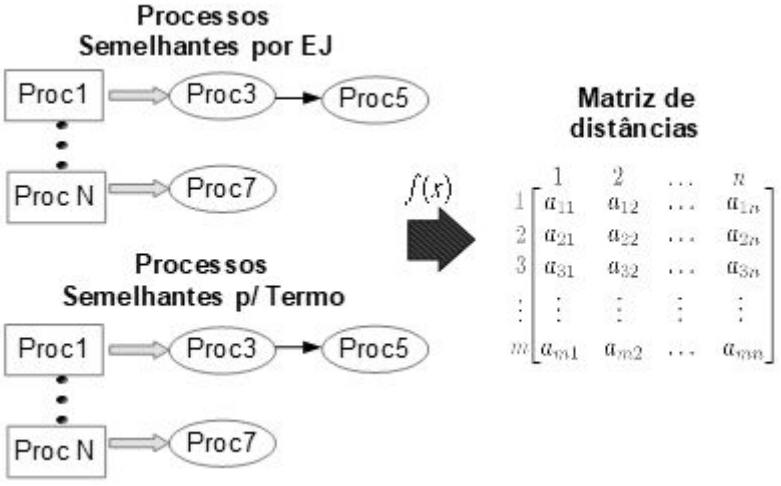

Figura 13: Geração da matriz de distâncias com base nas EJs e termos

Fonte: Elaborada pelo autor.

Com a matriz de distâncias calculada, o MDS foi utilizado para impor uma representação vetorial 2D em que cada ponto corresponde a um dos processos de exemplo utilizados nesta ilustração. Assim, esperando que as distâncias no plano refletissem o relacionamento semântico dos processos. A tabela 9 exibe a representação vetorial 2D dos cinco processos após a aplicação do MDS.

Processos	**00900801883**	**00100901107**	**00100901119**	**00100901261**	**00100901183**
X	1.25503807	0.33417767	0.33967187	0.3682725	0.39225114
Y	0.50431231	0.19774689	0.21825977	0.879214	0.91875521

Tabela 9: Representação vetorial dos cinco processos em 2D após aplicação do MDS

Fonte: Elaborada pelo autor.

De acordo com os artigos estudados, apesar de existirem trabalhos que utilizaram mais de 300 dimensões para representar os vetores de palavras (Ma, 2015) (Levy, 2014a) (Mikolov, 2013) (Baroni,2014), e trabalhos que consideraram menos de 200 dimensões (Prout, 2016) (Huang, 2012) (Huang, 2016) (Barkan, 2016), a maioria dos trabalhos analisados utilizou de 200 (Nii, 2016) (Fuchida, 2016) (Víta, 2016) (Oomoto, 2017) (Zhu, 2017) a 300 dimensões (Pennington, 2014) (Mikolov,2013) (Li et al., 2017) (Murgia et al., 2016) (Rahmawati, 2016) (Loukachevitch, 2016). Entretanto, para fins de ilustração e apresentação visual da ideia ao leitor deste trabalho, limitamos a projeção da representação dos documentos jurídicos a duas dimensões.

Com a construção da matriz de distâncias foi possível constatar que processos com maior grau de correlação, contextos semelhantes, apresentam representações próximas umas das outras, enquanto que processos rotulados com assuntos diferentes são representados por pontos distantes uns dos outros.

Assume-se que o conjunto de k processos, $\{ x_1, x_2, ..., x_k \}$ corresponde a uma coleção de pontos distribuídos num espaço cuja dimensão, m, é desconhecida. Dessa forma, apenas com base na matriz de distâncias M, busca-se um conjunto de pontos equivalentes, num espaço de dimensão n conhecida, cujas coordenadas formam as linhas de uma matriz V (dimensão $k \times n$). Como restrição para que essa equivalência seja obtida, deseja-se que as

relações de proximidade (representadas em M) sejam aproximadamente preservadas. Para tanto, construímos uma outra matriz de distâncias MC, representando a distância calculada entre os pontos de V.

Apesar de várias métricas serem utilizadas para o cálculo da similaridade semântica entre as palavras, a exemplo da distância de Manhattan, distância de Jaccard e do coeficiente de correlação de Pearson, a distância euclidiana e a similaridade entre cossenos são medidas comumente utilizadas para esse tipo de tarefa (Zhang, 2014) (Ning, 2016) (Wang, 2016). Além disso, levando em consideração que (Lapesa, 2014) obteve resultados totalmente equivalentes utilizando essas duas últimas medidas mencionadas, optamos por utilizamos a distância euclidiana como medida de distância para esse espaço de representação vetorial, ou seja, para fazer o mapeamento do espaço original de dimensão m para o espaço projetado de dimensão n.

A distância euclidiana entre os pontos representados pelas linhas da matriz V foi calculada através da seguinte fórmula:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (V_{ik} - V_{jk})^2} \qquad (5)$$

onde d_{ij} é a distância euclidiana entre os pontos i e j no espaço n-dimensional.

Após a construção da matriz de distâncias calculadas (MC), a mesma foi comparada à matriz de distâncias original (M) a partir da avaliação da função de ajuste definida. Quanto menor o valor do ajuste, "*stress*" (*standardized residual sum of squares*), maior será a similaridade entre essas duas matrizes. A função de *stress* é representada pela seguinte fórmula:

$$stress = \sqrt{\frac{\sum_{k=1}^{n} (M_{ik} - MC_{jk})^2}{\sum_{k=1}^{n} (M_{ik})^2}} \qquad (6)$$

A matriz V é inicializada com valores aleatórios e ajustada gradativamente pela medida de adequação de ajuste acima (Kruskal, 1964), buscando obter o menor valor de *stress*

(configuração de melhor ajuste) entre as duas matrizes, ou seja, uma maior correlação entre elas.

Para validação da qualidade do ajuste foi utilizada a tabela abaixo como referência (Kruskal, 1964):

Stress	**Qualidade do ajuste**
20%	Pobre
10%	Razoável
5%	Bom
2,5%	Excelente
0%	Perfeito

Tabela 10: Classificação do *Stress* de Kruskal

Fonte: Elaborada pelo autor.

Ao atingir um valor de *stress* aceitável, as distâncias armazenadas na matriz de distâncias calculadas MC se aproximam das distâncias da matriz de distâncias reais M e o ajuste é interrompido. Na imagem abaixo foi alcançado um valor de stress de 8% na execução de um experimento onde foram realizadas 100 iterações e uma taxa inicial de aprendizado de 0.1, utilizando-se 2 dimensões para representar 5 documentos processados.

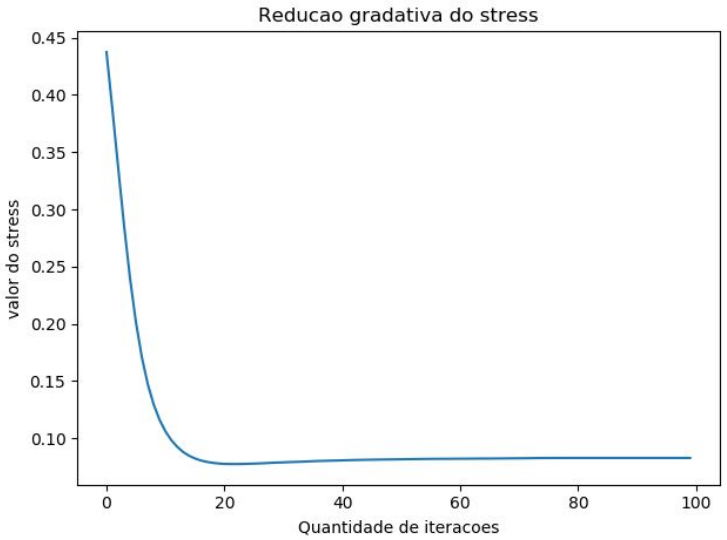

Figura 14: Redução gradativa do *stress*

Fonte: Elaborada pelo autor.

Com a adequação do ajuste foi possível representar a proximidade entre os processos semelhantes e distanciamento entre os processos não similares, conforme demonstrado na imagem abaixo.

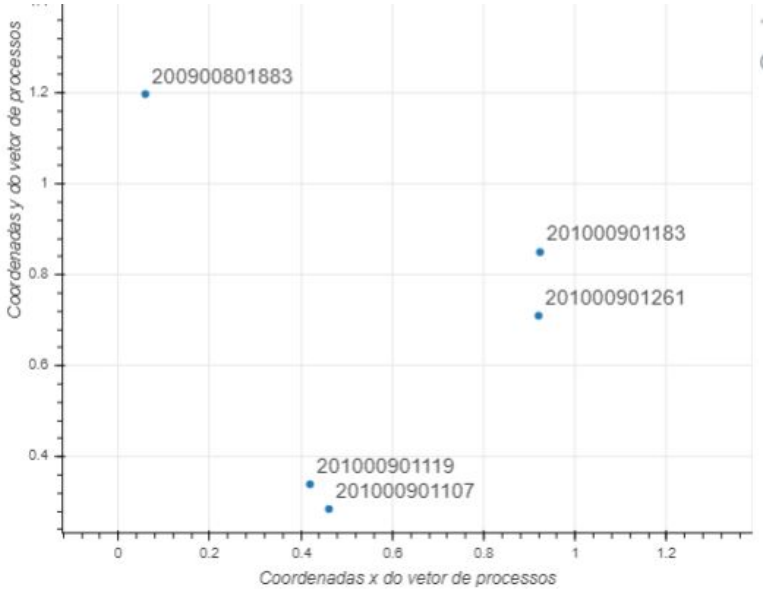

Figura 15: Gráfico com vetores de processos resultantes da aplicação do MDS

Fonte: Elaborada pelo autor.

Com a representação acima foi possível confirmar o relacionamento semântico entre os pares de processos: 201000901119 e 201000901107; 201000901261 e 201000901183 e o distanciamento do processo 200900801883 sendo que não possui semelhanças com os demais. Conforme exposto na tabela 11, esta representação se encontra de acordo com a classificação dos processos com base nos seus tipos.

Processo	**Tipo**
200900801883	Representação em juízo
201000901107	Seguro
201000901119	Seguro
201000901261	Correção monetária, Juros de mora - legais / contratuais, Assistência judiciária gratuita, Seguro
201000901183	Correção monetária, Juros de mora - legais / contratuais, Assistência judiciária gratuita, Seguro

Tabela 11: Tipos dos 5 processos utilizados como exemplo.

Fonte: Elaborada pelo autor.

4. Análise dos resultados

Foram realizados três experimentos para comparar os métodos e assim, validar o novo método proposto:

1. Aplicação do **MDS** em uma amostra dos dados, regulando o fator de ajuste, até chegar a um valor considerado bom pela escala de Kruskal;
2. Apresentação dos mesmos documentos ao **Doc2Vec**, comparando seus resultados ao **MDS** para verificar a qualidade dos resultados encontrados;
3. Construção de matriz de distâncias sem levar em consideração as EJs e os termos do tesauro para confirmar se realmente ambos influenciaram os resultados encontrados.

4.1. Experimento 1

Uma amostra contendo 448 processos foi selecionada aleatoriamente e, após o tratamento dos dados, extração das EJs, termos do tesauro e construção da matriz de distâncias, foi aplicado o MDS, sendo utilizadas vinte dimensões para representação desses processos e uma taxa inicial de aprendizado de 0.0035. Após 1192 iterações, alcançamos um valor de stress de 5% (0.05), o que é considerado "bom", segundo a escala de Kruskal (Tabela 10), conforme ilustrado na figura 16.

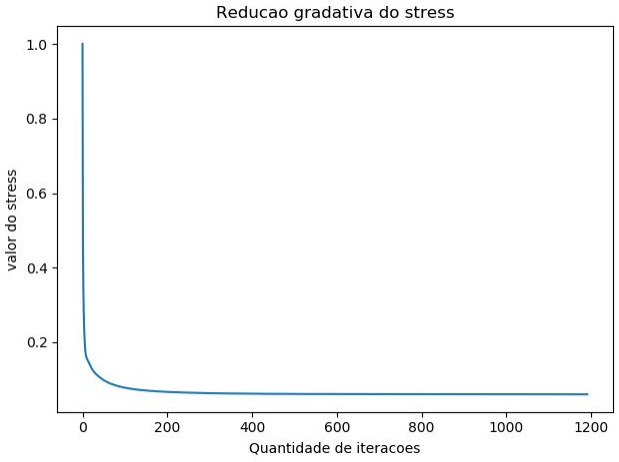

Figura 16: Redução gradativa do stress até atingir o valor de 5% na escala de Kruskal

Fonte: Elaborada pelo autor.

4.2. Experimento 2

A comparação entre os processos a partir do seus tipos foi realizada utilizando a mesma métrica utilizada para o cálculo da matriz de distâncias:

$$semelhancaProcTP = 1 - \frac{qtdTpsItsc}{qtdTpsProcA + qtdTpsProcB - qtdTpsItsc} \qquad (7)$$

onde qtdTpsProcA e qtdTpsProcB são as quantidade totais de tipos associados aos processos A e B respectivamente e qtdTpsItsc é a quantidade de tipos comuns aos dois processos.

Essas distâncias foram utilizadas como métricas desejadas (ou aproximadamente ideais) por serem fruto da análise humana de cada documento por especialistas, uma vez que os processos são analisados, ao longo do seu ciclo de vida (por juízes, promotores, advogados, analistas e técnicos judiciários e administrativos), servindo de modelo para análise dos resultados obtidos nas fases posteriores.

A tabela 12 é um exemplo de matriz de distância resultante da comparação de oito processos com base nos tipos dos processos.

	Proc1	**Proc2**	**Proc3**	**Proc4**	**Proc5**	**Proc6**	**Proc7**	**Proc8**
Proc1	0.	0.67	0.33	0.67	0.67	1.	0.6	0.
Proc2	0.67	0.	0.75	0.67	1.	1.	0.83	0.67
Proc3	0.33	0.75	0.	0.75	0.75	0.75	0.4	0.33
Proc4	0.67	0.67	0.75	0.	1.	1.	0.83	0.67
Proc5	0.67	1.	0.75	1.	0.	1.	0.83	0.67
Proc6	1.	1.	0.75	1.	1.	0.	0.83	1.
Proc7	0.6	0.83	0.4	0.83	0.83	0.83	0.	0.6
Proc8	0.	0.67	0.33	0.67	0.67	1.	0.6	0.

Tabela 12: Exemplo de matriz de similaridade entre processos por tipo

Fonte: Elaborada pelo autor.

Após o cálculo da semelhancaProcTP para todos os processos, calculamos a média de todos os valores de semelhança (linha a linha da matriz acima) para auxiliar no encontro do valor ideal que identifique um maior grau de semelhança entre os processos.

Analisando os dados e avaliando os resultados, verificou-se que os 5% menores valores de semelhança, baseados nos tipos dos processos, representavam os processos mais semelhantes, ou seja, com a seleção dos menores valores (referentes a 5% do total de processos da amostra), foi possível obter uma relação de semelhança entre todos os processos analisados com base apenas nos tipos dos processos.

A tabela abaixo demonstra um exemplo de semelhança encontrada entre os processos, onde a primeira coluna representa o processo principal e a coluna seguinte os seus processos semelhantes.

Processos	Processos semelhantes			
Proc1	Proc3	Proc4	Proc7	...
Proc2	Proc3	Proc5		...
Proc3	Proc6	Proc2	Proc1	...
Proc4	Proc8	Proc6	Proc1	...
Proc5	Proc5			...
Proc6	Proc7	Proc4	Proc8	...
Proc7	Proc5	Proc6		...
Proc8	Proc4	Proc7	Proc3	...
...	...	...	...	...

Tabela 13: Exemplo de semelhança entre processos com base nos seus tipos

Fonte: Elaborada pelo autor.

A figura 17 resume o papel dos "tipos dos processos" nesse processo, que vai desde a sua extração e construção da matriz de similaridade entre os processos com base nos seus tipos, utilizando a equação (7), exibida anteriormente, até a seleção dos processos mais semelhantes também baseado nos seus tipos.

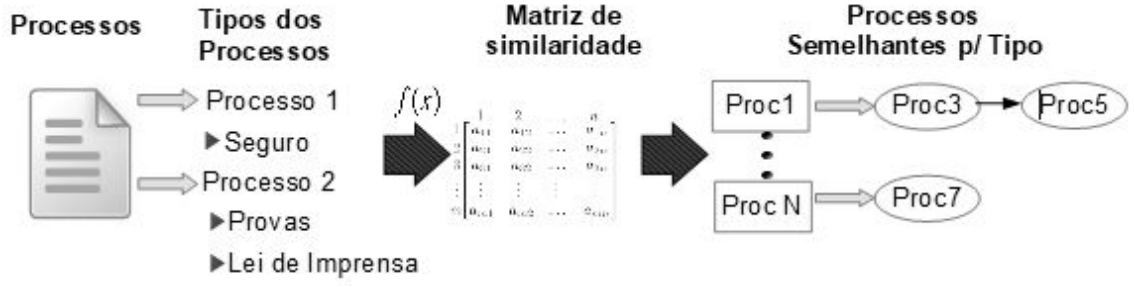

Figura 17: Extração dos tipos dos processos, construção da matriz de similaridade e seleção dos processos semelhantes

Fonte: Elaborada pelo autor.

Para comparação dos resultados encontrados, foi utilizada a implementação do Doc2Vec disponível na biblioteca de modelagem semântica de textos Gensim (Rehurek, 2010). O Gensim é uma biblioteca gratuita do Python que foi escrita por Radim Rehurek em 2009, utilizada para modelagem de espaços vetoriais e modelagem de tópicos. É considerada a melhor opção para processamento de grandes coleções de texto (Zahidi, 2019).

Os dados foram preparados seguindo os mesmos critérios dos dados apresentados ao MDS, conversão do texto para minúsculo, remoção de *stop-words*, acentos e caracteres especiais. O Doc2Vec foi executado com a mesma parametrização do MDS, 20 dimensões, a taxa inicial de aprendizado (alpha) foi de 0.0035 e foram realizadas 3000 iterações. Os demais parâmetros não foram alterados, sendo utilizados com os valores padrão da ferramenta.

A figura 18 resume o processo de execução do Doc2Vec, onde os dados tratados são apresentados como entrada e o como saída tem-se uma lista de processos com os seus respectivos processos semelhantes.

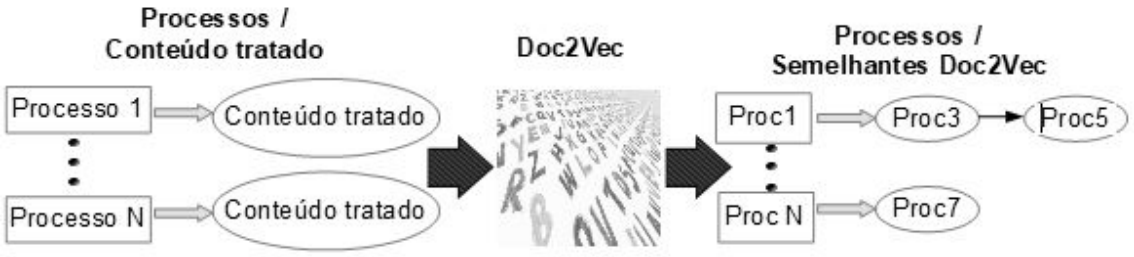

Figura 18: Aplicação do Doc2Vec aos processos para cálculo da semelhança entre processos
Fonte: Elaborada pelo autor.

Após a execução do modelo Doc2Vec, pudemos listar, para cada um dos processos da base, quais os seus processos semelhantes e comparar os resultados à semelhança por tipos de processos conforme demonstrado na equação:

$$semelhancaD2V = 1 - \frac{qtdProcItsc}{qtdProcTP + qtdProcD2V - qtdProcItsc} \qquad (8)$$

onde qtdProcTP e qtdProcD2V são as quantidade totais de processos semelhantes ao processo corrente, resultantes da construção da matriz de distâncias baseada nos tipos dos processos e

da execução do Doc2Vec respectivamente. Enquanto que qtdProcItsc é a quantidade de processos semelhantes resultantes dos dois métodos.

Realizamos também o cálculo dos processos semelhantes baseado nos valores da matriz de distâncias (M), comparando o resultado desse cálculo ao resultado da semelhança baseada nos tipos dos processos conforme descrito abaixo:

$$semelhancaMD = 1 - \frac{qtdProcItsc}{qtdProcTP + qtdProcMD - qtdProcItsc} \qquad (9)$$

onde qtdProcTP e qtdProcMD são as quantidade totais de processos semelhantes ao processo corrente, resultantes da construção da matriz de distâncias baseada nos tipos dos processos e da matriz de distâncias entre processos (M) respectivamente. Enquanto que qtdProcItsc é a quantidade de processos semelhantes resultantes dos dois métodos.

Com o cálculo das equações (8) e (9), demonstradas acima, foi possível comparar as semelhanças obtidas pelo Doc2Vec e pela Matriz M, aos valores da matriz de similaridade entre os processos com base nos tipos dos processos e, verificar qual dos dois métodos conseguiu classificar mais processos semelhantes, conforme demonstra figura abaixo.

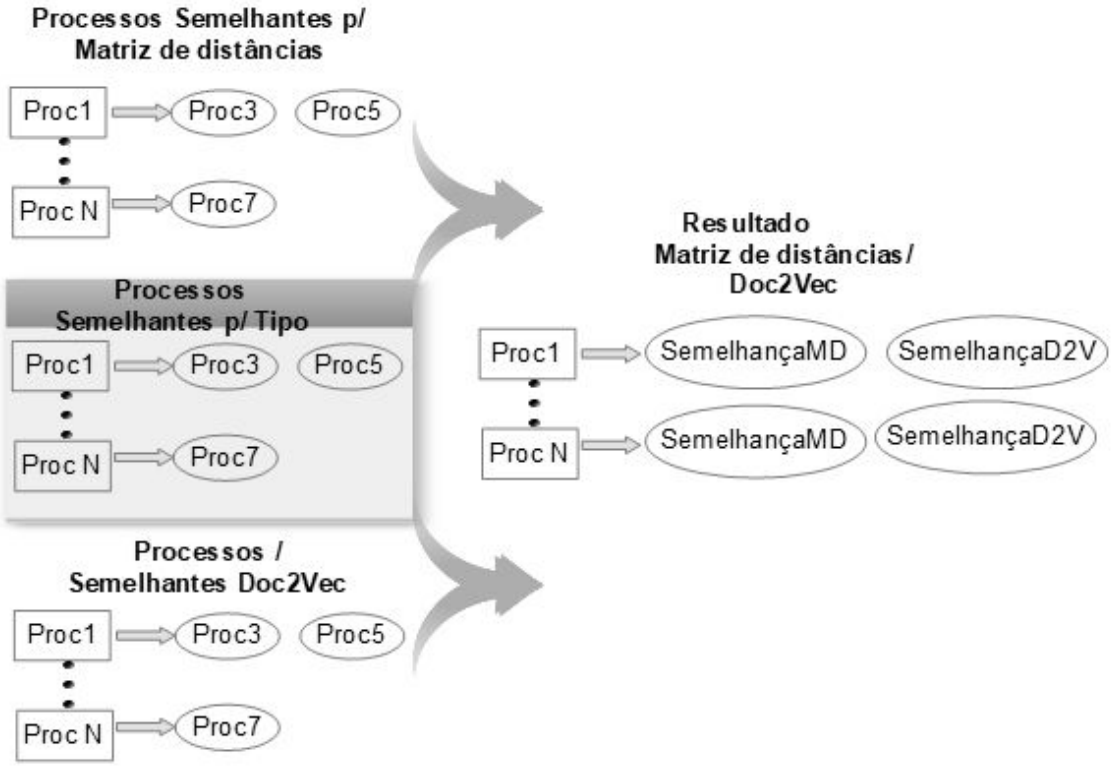

Figura 19: Comparação dos resultados da Matriz de Distâncias e do Doc2Vec aos resultados baseados nos tipos dos processos

Fonte: Elaborada pelo autor.

Após a execução dos dois métodos (construção de M e Doc2Vec), realizamos um teste com uma base contendo 448 processos. Comparando seus resultados, verificamos que, com a utilização da matriz de distâncias, o método proposto conseguiu classificar melhor em 43,5% (195 processos) das vezes, enquanto que o Doc2Vec foi superior em 35,7% (160 processos) dos casos. Sendo que, em 8,9% (40 processos) dos processos analisados, os dois métodos retornaram o mesmo resultado e, em 11,9% (53 processos) dos casos, os resultados de ambos os métodos foram diferentes das semelhanças calculadas a partir dos tipos dos processos.

4.3. Experimento 3

Outro teste que foi realizado foi a construção da matriz de distâncias M considerando apenas as palavras que fazem parte do conteúdo dos processos e a frequência em que aparecem nos documentos, sem levar em consideração os termos mais significativos da área jurídica (EJ) ou se os termos estavam catalogados no Vocabulário Jurídico do STF.

Apresentando apenas os documentos processados (sem *stop-words*, acentos, caracteres especiais e com o conteúdo convertido para minúsculo), da mesma maneira que foram apresentados ao Doc2Vec.

Após executar novamente os dois métodos, construção de M (sem EJ e termos) e o Doc2Vec, realizamos testes na mesma amostra de 448 processos. Comparando seus resultados, não conseguimos obter bons resultados utilizando a matriz de distâncias, onde o MDS conseguiu uma melhor classificação em apenas 12,9% (58 processos) das vezes enquanto Doc2Vec foi superior em 65,8% (295 processos) dos casos. Em 2,4% (11 processos) dos processos analisados, os dois métodos retornaram o mesmo resultado e, em 18,9% (84 processos) dos casos, os resultados de ambos os métodos foram diferentes das semelhanças calculadas à partir dos tipos dos processos.

5. APLICAÇÃO

Dentre as várias possibilidades de aplicação do método estudado, optamos pelo agrupamento de novos documentos, de tal forma que esses novos documentos são mapeados no espaço de documentos (se tornando um ponto nesse espaço) e, por proximidade a documentos (através dos tipos explicados na seção 3.3), documentos semelhantes ao documento apresentado são retornados sem a intervenção humana.

Para realização dessa aplicação, dividimos o processo em três etapas:

1. Tratamento dos textos dos novos processos;
2. Geração dos conjuntos de símbolos referentes aos novos processos tanto para levantamento das EJs quanto para os termos do tesauro;
3. Comparação dos conjuntos de símbolos dos novos processos aos conjuntos da base de testes para construção dos vetores de distâncias dos novos processos.

A métrica definida na equação (1) foi utilizada durante todo o processo. Tanto na comparação dos tipos dos novos processos, quanto na comparação das EJs quanto dos termos do tesauro.

5.1. Tratamento dos dados

Os dados foram tratados da mesma forma que nos demais experimentos. Foram removidas as *stop-words*, acentos, caracteres especiais e todo o conteúdo dos textos foi convertido para minúsculo.

5.2. Geração dos conjuntos de símbolos

Os novos processos foram comparados aos processos da base de treinamento levando-se em consideração:

5.2.1. EJs

Os conjuntos das Entidades Jurídicas dos novos processos foram comparados aos conjuntos das EJs de todos os processos da base de treinamento.

Entidades jurídicas dos processos / Quantidade de Ocorrência
A42CDC, A6L8987, A184CPC, S83, A177L8078, A4CDC, A55L9099, S7, AGRG1122762SP, AGRG1050470SP, AGRG854002RS, AGRG820665RS
A5CF, MP168, A177CC, A2028CC, A178CC, A11CC, S284, S282, S279, A55L9099, A46L9099, MP172
A3L6194, S14, A11L1060, A8L11482, MP451, A55L9099, A97CF, A5CF, A3CF, A515CPC, A333CPC, A5L6194, A102CF, A127CF, MP340
A3L6194, S14, A8L11482, A11L1060, A54L9099, A97CF, A5CF, A3CF, MP340, MP451, A269CPC, A55L9099, A10L9099, A515CPC, A333CPC, A5L6194, A102CF, A127CF, S426
A3L6194, S14, A8L11482, MP340, A5L6194, A97CF, A5CF, A3CF, MP451, A55L9099, A515CPC, A333CPC, A102CF, A127CF, A11L1060, A46L9099

Tabela 14: Conjuntos de EJs de cinco processos exemplo

Fonte: Elaborada pelo autor.

Como resultado dessa etapa, foi gerado um vetor da similaridade dos novos processos com os processos da base de treinamento.

5.2.2. Termos do tesauro

Foi realizada a comparação entre os conjuntos de termos do tesauro dos processos da base de treinamentos aos conjuntos de termos dos novos processos.

Termos(tesauro) dos processos
consumidor: 25, recorrente: 25, dano: 21, moral: 16, responsabilidade: 16, dano moral: 15, consumo: 12, indenizacao: 12, autos: 11, cdc: 10, fornecedor: 10
indenizacao: 27, seguro: 26, complementacao: 12, dignidade: 12, correcao: 10, dpvat: 10, inconstitucionalidade: 10
dano: 25, moral: 22, dano moral: 19, consumidor: 17, assistencia: 15, assistencia tecnica: 14 responsabilidade: 13, cdc: 11, bem: 10, indenizacao: 10
indenizacao: 34, invalidez: 33, pagamento: 33, salario: 18, salario minimo: 18, dpvat: 17, seguro: 17, invalidez permanente: 15, acidente: 12, medida: 12, complementacao: 10, prova: 10
pagamento: 41, invalidez: 35, indenizacao: 27, seguradora: 18, dpvat: 14, invalidez permanente: 14, seguro: 14, complementacao: 12, honorarios: 12, recorrente: 11, civil: 10, grau invalidez: 10, quitacao: 10

Tabela 15: Conjuntos de termos dos processos com suas respectivas quantidades de ocorrência

Fonte: Elaborada pelo autor.

Mais uma vez, é gerado apenas um novo vetor de similaridade entre o processo novo e os processos da base.

5.3. Comparação dos resultados

Com a similaridade entre os processos por EJs e por termos, foi aplicada a equação (4) para se obter a semelhança entre os processos e, a partir daí, foram selecionados os 5% dos processos da base com valores que representam um maior grau de semelhança, ou seja, valores mais próximos de zero.

Os novos processos também foram apresentados ao modelo treinado no Doc2Vec de forma que um conjunto de processos semelhantes foi inferido e comparado aos processos selecionados no parágrafo anterior

Foram apresentados 2193 novos documentos e mais uma vez o método proposto obteve resultados superiores aos Doc2Vec, conseguindo classificar melhor em 53,5% (1175 processos) das vezes enquanto Doc2Vec foi superior em 42,5% (932 processos) dos casos.

Em 4% (86 processos) dos processos analisados, os dois métodos retornaram o mesmo resultado.

6. CONCLUSÃO

Neste trabalho foram estudadas técnicas para a representação de espaços métricos adequados à representação de bases jurisprudenciais escritas em português brasileiro. Diante dos resultados expostos nos experimentos ficou comprovado que existe realmente um espaço métrico adequado para a representação de textos jurídicos escritos em português brasileiro.

Os experimentos foram realizados com os dados da base jurisprudencial do Tribunal de Justiça de Sergipe. Essa base é composta por aproximadamente 280 mil processos, distribuídos em quatro tipos de decisões: decisões monocráticas do Segundo Grau, acórdãos do Segundo Grau, decisões monocráticas da Turma Recursal e acórdãos da Turma Recursal.

Além da base de jurisprudência, também foi utilizado, buscando a padronização das informações analisadas, o vocabulário jurídico (tesauro) do Supremo Tribunal Federal (STF), que é um vocabulário controlado utilizado por pessoas que compartilham uma mesma linguagem.

Por se tratar de um método elaborado especificamente para documentos jurídicos, criado a partir da análise e avaliação das informações contidas nos processos jurídicos, foi possível se obter melhores resultados do que com a aplicação de métodos genéricos como o Doc2Vec.

Conforme evidenciado nos resultados apresentados, o método proposto conseguiu uma melhor classificação em 43,5% dos casos enquanto que o Doc2Vec foi superior em 35,7%. Para confirmar a eficiência de todo o processo realizado, executamos o MDS sem a utilização das EJs e dos termos do tesauro. Sem a utilização desses dois recursos, o MDS obteve uma melhor classificação em apenas 12,9% das vezes enquanto que o Doc2Vec foi superior em 65,8% dos casos, comprovando a importância das EJs e dos termos em todas as etapas que foram realizados neste trabalho.

7. TRABALHOS FUTUROS

Com o objetivo de aprimorar a comparação entre os contextos dos processos, pretendemos melhorar a detecção de novas entidades jurídicas, aumentando assim a qualidade dos contextos analisados e o número de características a serem avaliadas. Exemplo de entidades jurídicas identificadas: medida provisória, apelação cível, agravo regimental, agravo de instrumento, recurso especial, resolução, etc.

Buscando aumentar a abrangência e eficiência do algoritmo, pretendemos executar experimentos com um maior número de processos, dos mais variados tipos, bem como conseguir acesso a uma versão atualizada da base de dados de jurisprudência do TJSE.

8. REFERÊNCIAS

AHMAD A AND AMIN M. R., **Bengali word embeddings and it's application in solving document classification problem**. 2016. 19th International Conference on Computer and Information Technology (ICCIT). Anais...IEEE, fev. 2017. Disponível em: <https://ieeexplore.ieee.org/document/7860236>.

ALETRAS, N., TSARAPATSANIS, D., PREOȚIUC-PIETRO, D., & LAMPOS, V. **Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective**. PeerJ Computer Science, 2(Ml), e93. 2016. Disponível em: <https://doi.org/10.7717/peerj-cs.93/>.

BARKAN, O., & KOENIGSTEIN, N. **ITEM2VEC: Neural item embedding for collaborative filtering.** IEEE International Workshop on Machine Learning for Signal Processing, MLSP, 2016. Disponível em: <https://arxiv.org/vc/arxiv/papers/1603/1603.04259v2.pdf>.

BARONI, M., DINU, G., & KRUSZEWSKI, G. **Don't count , predict ! A systematic comparison of context-counting vs . context-predicting semantic vectors.** Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics., 238–247. 2014. Disponível em: <https://www.aclweb.org/anthology/P14-1023>.

BOJANOWSKI, P., GRAVE, E., JOULIN, A., & MIKOLOV, T. **Enriching Word Vectors with Subword Information**, 5, 135–146. 2016. Disponível em: <https://aclweb.org/anthology/Q17-1010>.

BRASIL. **Constituição (1988). Constituição da República Federativa do Brasil**. Disponível em: <http://www.planalto.gov.br/ccivil_03/Constituicao/Constituiçao.htm>. Acesso em: Julho 2019.

CUNHA, ROGÉRIO SANCHES. **Manual de direito penal: parte geral**. 7.ed. rev., ampl. e atual. - Salvador: JusPODIVM. 2019.

DENG, L. . **Deep Learning: Methods and Applications**. Foundations and Trends® in Signal Processing, 7(3–4), 197–387. 2014. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.691.3679&rep=rep1&type=pdf>.

FRANCO, ALBERTO SILVA. **Código Penal e sua Interpretação Jurisprudencial**. 4 ed. rev. e ampl. - São Paulo: Editora Revista dos Tribunais. 1993.

FUCHIDA, Y., TANIGUCHI, T., TAKANO, T., MORI, T., TAKENAKA, K., & BANDO, T. **Driving word2vec: Distributed semantic vector representation for symbolized naturalistic driving data.** IEEE Intelligent Vehicles Symposium, Proceedings, 2016, 1313–1320. . Anais...IEEE, ago. 2016. Disponível em: <https://ieeexplore.ieee.org/document/7535560>.

GALGANI, F., COMPTON, P., & HOFFMANN, A. **Combining Different Summarization Techniques for Legal Text**. Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, 115–123. 2012. Disponível em: <https://dl.acm.org/citation.cfm?id=2388647>.

GAO, Y., YANG, X., XU, J., & XU, B. **Valence-arousal ratings prediction with co-occurrence word-embedding**. 2016 International Conference on Asian Language Processing (IALP), 293–296. . Anais...IEEE, mar. 2017. Disponível em: <https://ieeexplore.ieee.org/document/7875989>.

GOYVAERTS, JAN; LEVITHAN, STEVEN. **Expressões Regulares Cookbook**. 2. ed. Novatec Editora, São Paulo. 2011.

GROVER, A., & LESKOVEC, J. **node2vec: Scalable Feature Learning for Networks**. KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 855-864 2016. Disponível em: <https://dl.acm.org/citation.cfm?doid=2939672.2939754>.

GRUENHAGE, G., OPPER, M., & BARTHELME, S. **Visualizing the effects of a changing distance on data using continuous embeddings**. Computational Statistics and Data Analysis, 104, 51–65. 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167947316301438?via%3Dihub>.

GUO, J., XU, G., CHENG, X., & LI, H. **Named entity recognition in query**. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '09, 267. 2019. Disponível em: <https://doi.org/10.1145/1571941.1571989>.

HACHEY, B., & GROVER, C. **Extractive summarisation of legal texts**. Artificial Intelligence and Law, 14(4), 305–345. 2006. Disponível em: <https://doi.org/10.1007/s10506-007-9039-z>.

HINO, M., & CUNHA, M. **Technology in Practice in Brazilian Judiciary: The Process of Computerization**, 1–11. 2014. Disponível em: http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1152&context=amcis2014>.

HUANG, E. H., SOCHER, R., MANNING, C. D., & NG, A. YA. **Improving Word Representations via Global Context and MultipleWord Prototypes**. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, (July), 873–882. 2012. Disponível em: <https://dl.acm.org/citation.cfm?id=2390645>.

HUANG, J., XU, K., & VYDISWARAN, V. G. V. **Analyzing Multiple Medical Corpora Using Word Embedding**. Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016, (1), 527–533. 2016. Disponível em: https://doi.org/10.1109/ICHI.2016.94>.

IACOBACCI, I., PILEHVAR, M. T., & NAVIGLI, R. **SENSEMBED: Learning Sense Embeddings for Word and Relational Similarity**. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference

on Natural Language Processing (Volume 1: Long Papers).Acl, (1), 95–105. 2015. Disponível em: <https://doi.org/10.3115/v1/P15-1010>.

JACCARD, P. **Nouvelles recherches sur la distribution florale**. Bulletin de La Société Vaudoise Des Sciences Naturelles, 44, 223–270. 1908. Disponível em: <https://doi.org/10.5169/seals-268384>.

JU, R., ZHOU, P., LI, C. H., & LIU, L. **An efficient method for Document categorization based on Word2vec and latent semantic analysis**. Proceedings - 15th IEEE International Conference on Computer and Information Technology, CIT 2015, 14th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2015, 13th IEEE International Conference on Dependable, Autonomic and Se, 2276–2283. 2015. Disponível em: <https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.336>.

KRUSKAL, J. B. **Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis**. Psychometrika, 29(1), 1–27. 1964. Disponível em: <https://doi.org/10.1007/BF02289565>.

KUMAR, S. **Similarity Analysis of Legal Judgments and applying 'Paragraph-link' to Find Similar Legal Judgments**. International Institute of Information Technology Hyderabad, India, (April). 2014. Disponível em: <https://www.semanticscholar.org/paper/Similarity-Analysis-of-Legal-Judgments-and-applying-Kumar/edbcbc4f128e16b719c2d7fafbf92a800ec5d1aa>.

LAPESA, G., & EVERT, S. **A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection**. Transactions of the Association for Computational Linguistics, 2, 531–545. 2014. Disponível em: <https://www.aclweb.org/anthology/papers/Q/Q14/Q14-1041/>.

LE, Q. V., & MIKOLOV, T. **Distributed Representations of Sentences and Documents**. International Conference on Machine Learning - ICML 2014, 32, 1188–1196. 2014. Disponível em: <https://cs.stanford.edu/~quocle/paragraph_vector.pdf>.

LEVY, O., & GOLDBERG, Y. **Linguistic Regularities in Sparse and Explicit Word Representations**. Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 171–180. 2014a. Disponível em: <https://doi.org/10.3115/v1/W14-1618>.

LEVY, O., & GOLDBERG, Y. **Neural Word Embedding as Implicit Matrix Factorization**. Advances in Neural Information Processing Systems (NIPS), 2177–2185. 2014b. Disponível em: <https://dl.acm.org/citation.cfm?id=2969070>.

LI, Q., SHAH, S., LIU, X., NOURBAKHSH, A., & FANG, R. **Tweet Topic Classification Using Distributed Language Representations**. Proceedings - 2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016, 81–88. 2017. Disponível em: <https://doi.org/10.1109/WI.2016.0022>.

LILLEBERG, J. **Support Vector Machines and Word2vec for Text Classification with Semantic Features**. 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC). Anais...IEEE, SET. 2015. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7259377>.

LIU, Z., & CHEN, H. **A Predictive Performance Comparison of Machine Learning Models for Judicial Cases**. 2017 IEEE Symposium Series on Computational Intelligence (SSCI). Anais...IEEE, fev. 2018. Disponível em: <https://ieeexplore.ieee.org/document/8285436>.

LOUKACHEVITCH, N., & ALEKSEEV, A. **Use of neighbor sentence co-occurrence to improve word semantic similarity detection**. Proceedings of the International FRUCT Conference on Intelligence, Social Media and Web, ISMW FRUCT 2016, 9. 2016. Disponível em: <https://doi.org/10.1109/FRUCT.2016.7584768>.

MA, L., & ZHANG, Y. **Using Word2Vec to process big text data**. Proceedings - 2015 IEEE

International Conference on Big Data, IEEE Big Data 2015, 2895–2897. 2015. Disponível em: <https://doi.org/10.1109/BigData.2015.7364114>.

MENACER M.A., BOUMERDAS A., ZAKARIA C., SMAILI K. **A New Language Model Based on Possibility Theory**. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2016. Lecture Notes in Computer Science, vol 9623. Springer, Cham. 2016. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-75477-2_8>.

MIKOLOV, T., CORRADO, G., CHEN, K., & DEAN, J. **Efficient Estimation of Word Representations in Vector Space**. Proceedings of the International Conference on Learning Representations (ICLR 2013), 1–12. 2013. Disponível em: <https://arxiv.org/abs/1301.3781>.

MOENS, M. F. **Summarizing court decisions**. Information Processing and Management, 43(6), 1748–1764. 2007. Disponível em: <https://doi.org/10.1016/j.ipm.2007.01.005>.

MURGIA, A., GHIDINI, G., EMMONS, S. P., & BELLAVISTA, P. **Lightweight Internet Traffic Classification: A Subject-Based Solution with Word Embeddings**. 2016 IEEE International Conference on Smart Computing, SMARTCOMP 2016. 2016. Disponível em: <https://doi.org/10.1109/SMARTCOMP.2016.7501703>.

NEVES, DANIEL AMORIM ASSUMPÇÃO. **Novo Código de Processo Civil Comentado**. Salvador: Ed. JusPodivm. 2016.

NII, M., TUCHIDA, Y., IWAMOTO, T., UCHINUNO, A., & SAKASHITA, R. **Nursing-care text evaluation using word vector representations realized by word2vec**. 2016 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2016, 2165–2169. 2016. Disponível em: <https://doi.org/10.1109/FUZZ-IEEE.2016.7737960>.

NING, W., & YU, M. **Exploiting distributional semantics to benefit machine learning in automated classification of Chinese clinical text**. Proceedings - 2016 IEEE International

Conference on Bioinformatics and Biomedicine, BIBM 2016, 1096–1102. 2016. Disponível em: <https://doi.org/10.1109/BIBM.2016.7822674>.

NIU, L., DAI, X., ZHANG, J., & CHEN, J. **Topic2Vec: Learning distributed representations of topics**. Proceedings of 2015 International Conference on Asian Language Processing, IALP 2015, 193–196. 2016. Disponível em: <https://doi.org/10.1109/IALP.2015.7451564>.

OLIVEIRA, R. A. N., & JUNIOR, M. C. **Experimental analysis of stemming on jurisprudential documents retrieval**. Information (Switzerland), 9(2). 2018. Disponível em: <https://doi.org/10.3390/info9020028>.

OLIVEIRA, ROBERT A. N. **Análise e Avaliação Experimentais de Técnicas para Recuperação de Documentos Jurisprudenciais**. Disponível em: <https://github.com/ranophoenix/radicalizacaojurisprudencial>. Acesso em: Julho 2019.

OOMOTO, K., OIKAWA, H., YAMAMOTO, E., YOSHIDA, M., OKABE, M., & UMEMURA, K. **Polysemy detection in distributed representation of word sense**. 2017 9th International Conference on Knowledge and Smart Technology: Crunching Information of Everything, KST 2017, 28–33. 2017. Disponível em: <https://doi.org/10.1109/KST.2017.7886073>.

PENNINGTON, J., SOCHER, R., & MANNING, C. D. **GloVe: Global Vectors for Word Representation**. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543. 2014. Disponível em: <https://doi.org/10.3115/v1/D14-1162>.

PHAHLAMOHLAKA, M. C., COETZEE, M., & LAW, A. C. **CaseRank : Ranking Case Law Using Precedent and Principal Component Analysis**. 2018 Conference on Information Communications Technology and Society (ICTAS). Anais...IEEE, mai. 2018. Disponível em: <https://ieeexplore.ieee.org/document/8368765>.

PRIETO, L. P., RODRÍGUEZ-TRIANA, M. J., KUSMIN, M., & LAANPERE, M. **graph2vec: Learning Distributed Representations of Graphs.** CEUR Workshop Proceedings, 1828, 53–59. 2017. Disponível em: <https://arxiv.org/abs/1707.05005>.

PROUT, X. D. B. **Unlock big data emotions - Weighted word embeddings for sentiment classification**. 2016 IEEE International Conference on Big Data (Big Data), 5, 3833–3838. Anais...IEEE, fev. 2017. Disponível em: <https://ieeexplore.ieee.org/document/7841056>.

RAHMAWATI, D., & KHODRA, M. L. **Word2vec semantic representation in multilabel classification for Indonesian news article.** 4th IGNITE Conference and 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, ICAICTA 2016, 0–5. 2016. Disponível em: <https://doi.org/10.1109/ICAICTA.2016.7803115>.

REHUREK, R., & SOJKA, P. **Software framework for topic modelling with large corpora.** LREC Workshop on New Challenges for NLP Frameworks, 45–50. 2010. Disponível em: <http://www.muni.cz/research/publications/884893>.

RUMELHART, D. E., Hinton, G. E., & Williams, R. J. **Learning representations by back-propagating errors.** Nature, 323(6088), 533–536. 1986. Disponível em: <https://doi.org/10.1038/323533a0>.

SIDOU, J.M. OTHLON. **Dicionário Jurídico: Academia Brasileira de Letras Jurídicas**. 5 ed. Rio de Janeiro: Forense Universitária. 1999.

SUN, F., GUO, J., LAN, Y., XU, J., & XUEQI, C. **Sparse word embeddings using l1 regularized online learning.** IJCAI International Joint Conference on Artificial Intelligence, 2016-Janua, 2915–2921. 2016. Disponível em: <https://dl.acm.org/citation.cfm?id=3060832.3061029>.

TRASK, A., MICHALAK, P., & LIU, J. sense2vec - **A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings.** IJCAI'16 Proceedings of the

Twenty-Fifth International Joint Conference on Artificial Intelligence Pages 2915-2921. 2015. Disponível em: <http://arxiv.org/abs/1511.06388>.

VAN, D. M., L. J. P., & HINTON, G. E. **Visualizing high-dimensional data using t-sne.** Journal of Machine Learning Research 9 (2008) 2579-2605. 2014. Disponível em: <http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>.

VENOSA, SÍLVIO DE SALVO. **Direito civil: parte geral, 6ª edição**. São Paulo: Atlas, 2006.

VITA, M., & KRÍŽ, V. **Word2vec based system for recognizing partial textual entailment.** 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), 8, 513–516. 2016. Disponível em: <https://doi.org/10.15439/2016F419>.

TESAURO, Supremo Tribunal Federal. **Vocabulário Jurídico (tesauro)**. Disponível em: <http://www.stf.jus.br/portal/jurisprudencia/pesquisarVocabularioJuridico.asp>. Acesso em: Julho 2019.

WANG, Z., MA, L., & ZHANG, Y. **A novel method for document summarization using Word2Vec.** In 2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC) (Vol. 129, pp. 523–529). 2016. Disponível em: <https://doi.org/10.1109/ICCI-CC.2016.7862087>.

ZAHIDI, Y., EL YOUNOUSSI, Y., & AZROUMAHLI, C. **Comparative Study of the Most Useful Arabic-supporting Natural Language Processing and Deep Learning Libraries**. 2019 5th International Conference on Optimization and Applications (ICOA), 1–10. 2019. Disponível em: <https://doi.org/10.1109/ICOA.2019.8727617>.

ZHANG, C., ZHANG, L., WANG, C. J., & XIE, J. Y. **Text Summarization Based on Sentence Selection with Semantic Representation.** Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 2014-Decem, 584–590. 2014. Disponível em: <https://doi.org/10.1109/ICTAI.2014.93>.

ZHENG, Y., SHI, Y., GUO, K., LI, W., & ZHU, L. **Enhanced word embedding with multiple prototypes.** 4th International Conference on Industrial Economics System and Industrial Security Engineering, IEIS 2017, (91546201). 2017. Disponível em: <https://doi.org/10.1109/IEIS.2017.8078651>.

ZHU, L., WANG, G., & ZOU, X. **A Study of Chinese Document Representation and Classification with Word2vec.** Proceedings - 2016 9th International Symposium on Computational Intelligence and Design, ISCID 2016, 1, 298–302. 2017. Disponível em: <https://doi.org/10.1109/ISCID.2016.1075>.