Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Identificação de Precedentes Judiciais por Agrupamento Utilizando Processamento de Linguagem Natural

Igor Tadeu Silva Viana Stemler

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Marcelo Ladeira

Coorientador
Prof. Dr. Thiago de Paulo Faleiros

Brasília
2019

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

SIG24i

Stemler, Igor Tadeu Silva Viana
Identificação de Precedentes Judiciais por Agrupamento Utilizando Processamento de Linguagem Natural / Igor Tadeu Silva Viana Stemler; orientador Marcelo Ladeira; co orientador Thiago de Paulo Faleiros. -- Brasília, 2019.
62 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2019.

1. Poder Judiciário. 2. Precedentes Judiciais. 3. Demandas Repetitivas. 4. Mineração de Texto. 5. Processamento de linguagem natural. I. Ladeira, Marcelo, orient. II. Faleiros, Thiago de Paulo, co-orient. III. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Identificação de Precedentes Judiciais por Agrupamento Utilizando Processamento de Linguagem Natural

Igor Tadeu Silva Viana Stemler

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Marcelo Ladeira (Orientador)
CIC/UnB

Prof. Dr. Díbio Leandro Borges
CIC/UnB

Dr. João Alberto de Oliveira Lima
Senado Federal

Prof. Dr. Alan Ricardo da Silva
Est/UnB

Prof.a Dr.a Aletéia Patrícia Favacho de Araújo
Coordenadora do Programa de Pós-graduação em Computação Aplicada

Brasília, 05 de julho de 2019

Dedicatória

Dedico este trabalho ao meu filho Miguel Stemler, meus pais, irmãos e minha noiva Larissa Leles, que me apoiaram durante todo o processo de aprendizado e desenvolvimento.

Agradecimentos

Agradeço primeiramente a Deus por tudo que ele tem me proporcionado. Agradeço à minha família por todo apoio e cuidado, principalmente em ajudar a cuidar do meu filho Miguel Stemler enquanto eu me dedicava a este trabalho. Agradeço a minha noiva Larissa Leles, que é uma excelente advogada e foi a especialista responsável por verificar se os precedentes judiciais eram semelhantes ou não. Agradeço ao Conselho Nacional de Justiça (CNJ) por aprovar meu projeto e incentivar meus estudos, principalmente nas figuras da minha amiga Gabriela Soares, que ingressou comigo no CNJ há 11 anos atrás e continua sendo uma grande professora. Aos meus amigos Ricardo, Filipe, Lucas e Rondon, que são excelentes profissionais e me ensinam diariamente a ser uma pessoa melhor. Agradeço, também, ao professor Marcelo Ladeira por toda paciência e dedicação que teve comigo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

O Poder Judiciário Brasileiro possui um grande acúmulo de processos e um quantitativo considerável de casos repetitivos. Em 2015 o Código de Processo Civil foi reformulado e criou o Incidente de Resolução de Demandas Repetitivas (IRDR) e Incidente de Assunção de Competência (IAC). Esses institutos permitem aos tribunais de 2ª instância sobrestar o andamento de processos similares em 1ª instância, afetando um deles como processo principal para será julgado como paradigma para os demais. O julgamento nele proferido será aplicado para todos os casos sobrestados em 1ª instância, evitando a insegurança jurídica, já que reduz a chance de processos similares terem soluções judiciais diferentes por terem sido julgados por juízes distintos. Essa insegurança jurídica poderá ocorrer na 2ª instância, uma vez que tribunais diferentes podem julgar causas semelhantes de forma contrária. O objetivo deste trabalho é criar uma ferramenta para identificação automática dos Precedentes Judiciais semelhantes e testar a hipótese de que algoritmos que utilizam abordagens semânticas como o *Latent Semantic Indexing (LSI)* e *Latent Dirichlet Allocation (LDA)* apresentam resultados melhores do que aqueles que utilizam somente abordagem sintática como a função de ranqueamento *Okapi Best Matching 25 (BM25)*. Os modelos são avaliados de forma experimental, utilizando métricas de aferição de similaridade e o auxílio de um especialista. O melhor modelo é estendido utilizando entidades nomeadas para verificar se há melhora na performance. Precedentes semelhantes podem ser identificados com a utilização de técnicas baseadas em análise semântica de texto, mas essa abordagem apresenta custo maior do que a abordagem com o modelo BM25. Há melhora nos resultados com o uso de entidades nomeadas, mas com menor precisão ao identificar novos precedentes semelhantes em relação ao modelo BM25, que mostrou ser a melhor opção para a ferramenta em produção. A ferramenta apresenta grafos contendo os precedentes judiciais semelhantes, podendo o usuário verificar se há precedentes semelhantes com decisões divergentes e precedentes que deveriam estar suspensos devido a existência de recursos aos tribunais superiores de questão de mesmo tema.

Palavras-chave: Precedentes Judiciais, Demandas Repetitivas, Mineração de Texto, BM25, LSI, LDA

Abstract

The Brazilian Judiciary has too many lawsuits and a considerable number of repetitive cases. The Code of Civil Procedure was reformulated in 2015 and created the *Incidente de Resolução de Demandas Repetitivas (IRDR)* and *Incidente de Assunção de Competência (IAC)* institutes. These precedents allow the courts of Appeal to suspend similar lawsuits in the first instance, affecting one of them as the main process to be judged as a paradigm for the others. The judgment of this case has to be applied to all cases overturned in the first instance, avoiding legal uncertainty, since it reduces the chance of similar cases being judged differently. This legal uncertainty can occur in the second instance, since different courts can judge similar causes in the opposite way. The objective of this work is to create a tool for automatic identification of similar Precedents and to test the hypothesis that algorithms that use semantic approaches such as Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) present better results than those that use only syntactic approach such as the Okapi Best Matching 25 (BM25) ranking function. The models are evaluated experimentally, using similarity gauging metrics and the assistance of a specialist. The best model is extended using named entities to check for performance improvement. Similar precedents can be identified using techniques based on semantic text analysis, but this approach is more costly than the BM25 approach. There is improvement in the results with the use of named entities, but with less precision when identifying similar new precedents in relation to the BM25 model, which proved to be the best option for the tool in production. The tool presents graphs containing similar judicial precedents, and the user can verify if there are similar precedents with divergent decisions and precedents that should be suspended due to the existence of appeals to the higher courts of issue of the same subject.

Keywords: Judicial Precedents, class action, Text Mining, BM25, LSI, LDA

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ABJ Associação Brasileira de Jurimetria.

BERT Bidirectional Encoder Representations from Transformers.

BM25 Okapi Best Matching 25.

BNPR Banco Nacional de Dados de Demandas Repetitivas e Precedentes Obrigatórios.

BOW Bag of Words.

CNJ Conselho Nacional de Justiça.

CPC Código de Processo Civil.

CRISP-DM Cross Industry Standard Process for Data Mining.

DPJ Departamento de Pesquisas Judiciárias.

EM Expectation–maximization.

IAC Incidente de Assunção de Competência.

IRDR Incidente de Resolução de Demandas Repetitivas.

LDA Latent Dirichlet Allocation.

LSI Latent Semantic Indexing.

PLSI Probabilistic Latent Semantic Indexing.

RG Repercussão Geral.

RR Recursos Repetitivos.

STF Supremo Tribunal Federal.

STJ Superior Tribunal de Justiça.

SVD Singular-Value Decomposition.

TEMAC Teoria do Enfoque Meta Analítico Consolidado.

TST Tribunal Superior do Trabalho.

Capítulo 1

Introdução

O Poder Judiciário brasileiro tem como órgão de cúpula o Supremo Tribunal Federal (STF), ao qual compete primordialmente a guarda da Constituição Federal. Paralelamente àquele, encontra-se o Conselho Nacional de Justiça (CNJ), responsável pelo controle da atuação administrativa e financeira do Poder Judiciário. Visando aprimorar a Gestão Judiciária Brasileira, o CNJ lança desde o ano de 2006 o relatório *Justiça em Números* [1], publicado pelo Departamento de Pesquisas Judiciárias (DPJ), sendo a principal fonte de estatísticas oficiais do Poder Judiciário, detalhando a estrutura e a litigiosidade no país. Segundo o relatório, o total de processos no Poder Judiciário aumentou gradativamente nos últimos anos, atingindo 80 milhões de processos em tramitação ao final do ano de 2017 para aproximadamente 18 mil magistrados em mais de 15 mil unidades judiciárias, dentre turmas, varas e juizados especiais. Cada juiz proferiu no ano de 2017, em média, 7 sentenças por dia. Comparada aos países europeus, a produtividade dos magistrados brasileiros é a terceira maior de um total de 31 países, atrás somente da Dinamarca e da Áustria [2]. Já com relação à carga de trabalho dos juízes, que correspondente ao total de processos que tramitou no ano por magistrado, o Brasil passa a ocupar a segunda posição, logo após a Dinamarca.

Em outra vertente, o estudo *100 Maiores Litigantes* [3] identificou que 100 instituições apresentaram mais de 30% do total de processos ingressados no ano de 2011. Nessa mesma linha, a Associação Brasileira de Jurimetria (ABJ) em parceria com o CNJ apontou no estudo *Os Maiores Litigantes nas Ações Consumeristas na Justiça Estadual: Mapeamento e Proposições* [4] que 20 empresas concentram mais de 50% dos processos judiciais de ações consumeristas na Justiça Estadual, ou seja, determinadas instituições detêm o maior número de processos em trâmite no Poder Judiciário. Ocorre que, dentro desta quantidade de processos, existem inúmeros que tratam da mesma matéria jurídica.

A questão que se tem verificado na prática é que esses processos repetidos estão sendo julgados por juízes ou tribunais distintos, e em muitos deles, têm se apresentado solu-

ções contrárias em casos semelhantes. Decisões diferentes em situações análogas viola diretamente o princípio da segurança jurídica.

No intuito de amenizar esse impasse o Poder Judiciário tem criado mecanismos ou instrumentos que possibilitem que julgadores distintos adotem o mesmo posicionamento jurisdicional em casos semelhantes, garantindo assim igualdade aos litigantes. Estes mecanismos utilizados pelo Poder Judiciário fazem parte do Sistema de Precedentes Judiciais, tais como: Súmulas Simples/Vinculantes, Incidente de Resolução de Demandas Repetitivas (IRDR), Incidente de Assunção de Competência (IAC), Recursos Repetitivos (RR), Repercussão Geral (RG).

Vale lembrar que o intuito do Judiciário adotar a observância a Precedentes Judiciais não é apenas o de garantir segurança jurídica aos julgamentos dos juízes e Tribunais, mas também o de desafogar o judiciário e assegurar maior celeridade processual.

O primeiro mecanismo para fixação de precedentes veio por meio da Repercussão Geral (RG), que foi instituído pela Emenda Constitucional nº 45, de 2004, que criou a possibilidade de sobrestamento de processos das demais instâncias até ulterior julgamento da matéria com repercussão geral no STF. Como o próprio nome já diz, considera-se repercussão geral aquela matéria que possui grande relevância jurídica, política, social ou econômica, com repercussão para toda coletividade, ou seja, um caso afetado para julgamento servirá como direcionamento para os demais casos semelhantes a ele.

Em seguida, a Lei 11.672 de 2008 alterou o Código de Processo Civil (CPC) e introduziu um novo procedimento denominado Recursos Repetitivos (RR). Esse instituto se parece com o da Repercussão Geral quanto à determinação de que os processos de mesma tese jurídica tenham o trâmite suspenso (sobrestado) até a deliberação da matéria pelo STJ. Diferenciam-se na medida em que, na Repercussão Geral, basta um tema de grande relevância para o julgamento do Recurso, enquanto que no Recurso Repetitivo é preciso que repetidamente existam recursos com matérias semelhantes para que um deles seja afetado e julgado, servindo como direção para os demais casos.

O instituto do Recurso Repetitivo foi estendido ao Tribunal Superior do Trabalho (TST) com a edição da Lei nº 13.015/14, aplicando-se ao recurso de revista, no que couber, as normas relativas ao julgamento dos recursos extraordinários e especial repetitivos.

As demandas em massa tiveram atenção especial na reformulação do Código de Processo Civil (CPC), definido pela Lei nº 13.105, de 16 de março de 2015, com a adição de uma nova hipótese de suspensão do processo, o Incidente de Resolução de Demandas Repetitivas (IRDR). Esse incidente pode ser instaurado quando há, simultaneamente, efetiva repetição de processos na 1ª instância (varas e juizados especiais), que contenham controvérsia sobre a mesma questão de direito e risco de ofensa à isonomia e à segurança jurídica [5]. Ou seja, ele tem por objetivo evitar julgamentos divergentes em uma mesma

questão de direito analisada por diferentes juízes do mesmo tribunal repetidamente. O IRDR pode ser instaurado pelo juiz, pelo relator, pelas partes, pelo Ministério Público ou pela Defensoria Pública.

O CPC também criou o Incidente de Assunção de Competência (IAC), que é admissível quando o processo envolve relevante questão de direito, com grande repercussão social e sem repetição em múltiplos processos.

O Artigo 979 do CPC cita que os tribunais manterão banco eletrônico de dados atualizados com informações específicas sobre questões de direito submetidas ao incidente, comunicando-o imediatamente ao Conselho Nacional de Justiça, órgão encarregado de manter o cadastro nacional desses incidentes. A instauração e o julgamento do incidente serão sucedidos da mais ampla e específica divulgação e publicidade, por meio de registro eletrônico no Conselho Nacional de Justiça.

O Código de Processo Civil entrou em vigor em março do ano de 2016, iniciando em outubro daquele mesmo ano o monitoramento dos precedentes judiciais por intermédio do cadastro nacional do CNJ. As informações recebidas estão sendo publicadas no endereço eletrônico do CNJ[1] desde maio de 2017. O painel apresenta dados consolidados de todos os tribunais do Poder Judiciário com relação aos processos repetitivos e aos precedentes obrigatórios, assim como o quantitativo de processos sobrestados vinculados a cada tema, os assuntos dos processos e os tipos de recursos. A análise descritiva dos dados consta no relatório publicado pelo DPJ [6].

Há no cadastro Banco Nacional de Dados de Demandas Repetitivas e Precedentes Obrigatórios (BNPR), instituído pela Resolução n. 235/2016 do CNJ, mais de 2.500 precedentes judiciais cadastrados e mais de 1 milhão de processos sobrestados. Esses processos aguardam, em geral, julgamento da RG pelo STF e do RR pelo STJ ou pelo TST. Por estarem em vigor há pouco mais de três anos e serem relativamente novos, espera-se que o crescimento do número de IRDR e IAC seja superior a um crescimento linear, culminando, consequentemente, em um aumento significativo de processos sobrestados. Os institutos IRDR e IAC podem ser julgados, em regra, por 86 tribunais distintos, divididos em 5 Tribunais Regionais Federais, 27 Tribunais de Justiça, 3 Tribunais de Justiça Militar, 24 Tribunais Regionais do Trabalho e 27 Tribunais Regionais Eleitorais. A Figura 1.1 ilustra o organograma do Poder Judiciário com a indicação do local de julgamento de cada mecanismo.

Uma verificação inicial dos precedentes judiciais na presente pesquisa identificou dois IRDR's com temas semelhantes no Tribunal de Justiça de Minas Gerais referentes ao direito de férias prêmio em municípios mineiros diferentes. Há também a identificação

[1]Disponível em: `http://www.cnj.jus.br/pesquisas-judiciarias/paineis`

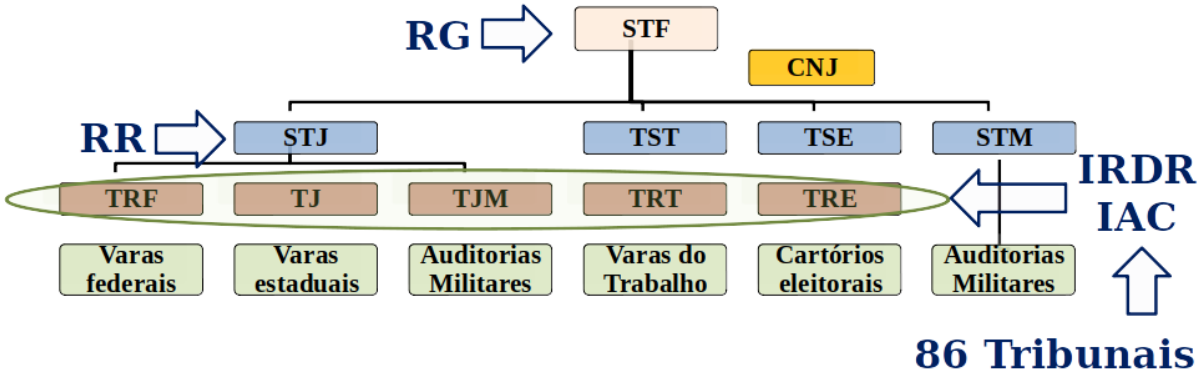

Figura 1.1: Organograma do Poder Judiciário.

de um IRDR que trata do cálculo do ICMS nas tarifas de uso dos sistemas elétricos de transmissão e de distribuição de energia em 8 Tribunais de Justiça, no STJ e no STF.

1.1 Definição do problema

Ao finalizar o julgamento do Incidente de Resolução de Demandas Repetitivas ou Incidente de Assunção de Competência pelo Tribunal, a decisão nele proferida será aplicada para todos os demais casos similares em 1ª instância, diminuindo assim a insegurança jurídica, já que reduzirá a chance de processos similares terem soluções judiciais diferentes. No entanto, essa insegurança jurídica poderá ser levada para o segundo grau (ou 2ª instância), na medida em que tivermos Tribunais diferentes julgando causas semelhantes por IRDR ou IAC de forma contrária.

Uma forma de evitar essa insegurança jurídica em 2ª instância seria a criação de uma ferramenta que apontasse aos tribunais os precedentes judiciais semelhantes aos seus. Dessa forma, o tribunal pode observar o que já foi julgado nos demais e aplicar a mesma tese já decidida ou, em caso de julgamento em divergência, indicar aos Tribunais Superiores que o tema necessita de uniformização da jurisprudência, evitando, assim, a insegurança jurídica.

1.2 Justificativa do tema

O CNJ possui papel fundamental ao dar publicidade às informações dos tribunais, possibilitando que haja a comparação de decisões proferidas em precedentes judiciais semelhantes. Cabe ao DPJ produzir relatórios periódicos a respeito da metodologia de tratamento de casos repetitivos e de formação concentrada de precedentes obrigatórios.

O IRDR e o IAC visam reduzir a insegurança jurídica na 1ª instância, diminuindo a chance de os juízes proferirem decisões divergentes em casos semelhantes. Entretanto, essa mesma insegurança jurídica poderá ocorrer na segunda instância, uma vez que tribunais diferentes podem julgar IRDR e IAC de temas semelhantes de forma diversa. Verifica-se, portanto, a importância da identificação dos temas de precedentes judiciais semelhantes, com o intuito de apontamento de decisões divergentes entre os tribunais. Essa identificação traz uma contribuição importante para minimizar a insegurança jurídica na 2ª instância.

Além disso, a identificação dos precedentes judiciais semelhantes possibilita a verificação de que um mesmo tema de IRDR ou IAC já pode estar submetido a julgamento em um recurso extraordinário (STF) ou especial (STJ) e, nestes casos, todos os IRDR's e IAC's de mesmo tema deverão ser suspensos até o julgamento do mérito pela instância superior.

1.3 Hipóteses a investigar

As hipóteses da pesquisa são que os precedentes judiciais semelhantes podem ser identificados com a utilização de técnicas baseadas em análise semântica de texto e que o resultado obtido pode ser melhorado com a utilização de entidades nomeadas. Espera-se que o resultado possa ser permanentemente aprimorado devido a entrada diária de novos registros e com a supervisão de um especialista.

1.4 Objetivos

O principal objetivo deste trabalho é criar uma ferramenta para identificação automática dos Precedentes Judiciais semelhantes. Os objetivos específicos são:

1. avaliar o uso de técnicas baseadas em análise semântica para estimar similaridade entre precedentes;
2. analisar valores de similaridade por cosseno para identificação de precedentes semelhantes;
3. validar os grupos de precedentes indicados como similares com apoio de um especialista humano;
4. medir a performance dos modelos comparando os agrupamentos propostos com os considerados válidos pelo especialista;
5. verificar ganhos de performance ao considerar entidades nomeadas;
6. desenvolver uma prova de conceito para utilização do algoritmo pelo CNJ.

1.5 Contribuição esperada

A criação de uma ferramenta para identificação automática de precedentes judiciais constituirá uma inovação para o CNJ, pois atualmente não existe nenhum mecanismo nesse sentido e essa identificação não é realizada atualmente. Ademais, a utilização da técnica de similaridade de textos representa uma contribuição técnica no órgão, uma vez que tais técnicas não são utilizadas atualmente.

Como consequência, os tribunais poderão identificar precedentes judiciais semelhantes já julgados em outras cortes e utilizá-las como referência ou encaminhar a demanda aos tribunais superiores, uniformizando, assim, os entendimentos com o intuito de tornar o Poder Judiciário mais confiável e célere, reduzindo como consequência o estoque processual.

O presente trabalho está organizado da seguinte forma: o Capítulo 1 traz uma breve contextualização da necessidade de criação de uma ferramenta que auxilie os tribunais na identificação de precedentes judiciais semelhantes existentes em outros tribunais. O Capítulo 2 apresenta a fundamentação teórica e revisão do estado da arte contendo os principais estudos que versam sobre a criação dos instrumentos de Incidentes de Resolução de Demandas Repetitivas (IRDR) e Incidentes de Assunção de Competência (IAC) e sobre as técnicas de identificação de textos similares. O Capítulo 3 apresenta os experimentos realizados. O Capítulo 4 apresenta os resultados obtidos com a aplicação das técnicas de mineração de texto. Por fim, o último Capítulo traz as conclusões obtidas deste estudo e sugere a realização de trabalhos futuros ligados ao tema em questão.

Capítulo 2

Fundamentação teórica e revisão do estado da arte

Este Capítulo está subdividido em 2 etapas: a primeira refere-se a uma revisão qualitativa de amplo escopo sobre os precedentes judiciais com a utilização da base de dados Google Acadêmico. A segunda etapa apresenta uma revisão qualitativa e quantitativa sobre os artigos que versam sobre similaridade de textos publicados nas bases de dados *Scopus*[1] e *Web of Science*[2] nos últimos 5 anos.

2.1 Precedentes Judiciais

A Constituição Federal da República Federativa do Brasil de 1988 é a lei máxima brasileira, devendo as demais legislações estarem de acordo com o que ali foi prescrito pelo Constituinte Originário. Além das leis, tem-se também as fontes normativas secundárias, como doutrina, costumes, princípios gerais do direito e jurisprudência. A jurisprudência é considerada como o "conjunto de decisões uniformes e constantes dos tribunais resultante da aplicação de normas a casos semelhantes" [7].

Nos países de *common law*, como os Estados Unidos, por exemplo, os precedentes têm função primária como fonte do Direito, e sua aplicação é obrigatória a casos semelhantes posteriores. Quando estamos diante do sistema *civil law*, adotado pelo ordenamento jurídico brasileiro, considera-se a lei como fonte primária/principal do direito, e as outras fontes subordinam-se a ela [8].

No *civil law* os precedentes julgados não possuem poder vinculativo. Por esta razão, a não ser por alteração legislativa, como é o caso da súmula vinculante, a jurisprudência não pode ser concebida como fonte primária do Direito. Todavia, é incontestável que o

[1] `https://www.scopus.com/home.uri`
[2] `https://www.webofknowledge.com/`

conjunto de decisões a respeito de uma matéria, com as mesmas características, tende a persuadir o juiz, que se inclina a decidir do mesmo modo [8].

O Brasil adota o sistema *civil law*, entretanto, é de fácil percepção que há uma mistura dos dois sistemas na medida em que a lei não tem sido a única fonte a ser observada, mas também, a jurisprudência tem se tornado uma fonte normativa por excelência. Dessa forma, os precedentes não só têm sido utilizados como direcionamento em decisões judiciais, como também devem ser observados nos casos dos precedentes obrigatórios.

Nas últimas duas décadas, foram implementadas inúmeras reformas processuais de valorização do direito jurisprudencial, desde a criação de enunciados de súmulas, Súmula Vinculante, e julgamento liminar de demandas repetitivas.

O novo Código de Processo Civil mantém essas reformas e evidencia o importante papel que o direito jurisprudencial exerce no ordenamento jurídico brasileiro com o delineamento de um microssistema de litigiosidade repetitiva que encampa, entre seus preceitos, novo regramento dos precedentes no Brasil [9].

O sistema de Precedentes Judiciais organizados pelo Código é uma manifestação do princípio da igualdade, já que garante que pessoas numa mesma situação recebam do judiciário o mesmo tratamento no julgamento dos seus processos. Além disso, tende a promover a segurança jurídica e diminuir o volume de causas no Judiciário, já que, sabendo dos precedentes em seu caso concreto, as pessoas tendem a evitar demandas temerárias. A observância de precedentes tem acarretado uma maior estabilidade jurídica no país, na medida em que os julgadores seguem pontualmente a mesma linha de raciocínio para situações semelhantes, garantindo segurança jurídica às decisões em situações análogas.

Precedente, segundo Joyce Mendes [8], "é a decisão judicial tomada à luz de um caso concreto, cujo núcleo essencial pode servir como diretriz para o julgamento posterior de casos análogos". E ainda, segundo Henry Black, "como um caso sentenciado ou decisão de uma corte considerada como fornecedora de um exemplo ou de uma autoridade para um caso similar ou idêntico posteriormente surgido, ou para uma questão similar de direito" [10].

Os precedentes judiciais podem ser considerados, portanto, decisões jurídicas que servem de suporte, de subsídio, para o julgamento de processos similares posteriores a eles. O sistema de precedentes obrigatórios impõe ao juiz da causa o dever de observar o precedente judicial antes de proferir o seu julgamento.

Neste trabalho, destacamos os institutos do Incidente de Resolução de Demandas Repetitivas (IRDR), Incidente de Assunção de Competência (IAC), Recursos Repetitivos (RR) e Repercussão Geral (RG), uma vez que estes são os precedentes cadastrados no BNPR.

O Incidente de Resolução de Demandas Repetitivas encontra-se disciplinado no art.

976, incisos I e II do CPC, assim mencionando: "Art. 976. É cabível a instauração do incidente de resolução de demandas repetitivas quando houver, simultaneamente: I - efetiva repetição de processos que contenham controvérsia sobre a mesma questão unicamente de direito; II - risco de ofensa à isonomia e à segurança jurídica".

O IRDR é dirigido ao presidente do Tribunal (2ª instância) para que este, sobrestando os processos que estão tramitando naquele juízo acerca do assunto afeto ao IRDR, possa julgá-lo, fixando o precedente judicial que será aplicado a todos estes processos que foram sobrestados e sejam semelhantes ao julgado.

A grosso modo, pode-se observar que o procedimento utilizado pelo IRDR é o mesmo do Recurso Repetitivo. Conforme dispõe o art. 1.036 do CPC, "Art. 1.036. Sempre que houver multiplicidade de recursos extraordinários ou especiais com fundamento em idêntica questão de direito, haverá afetação para julgamento de acordo com as disposições desta Subseção, observado o disposto no Regimento Interno do Supremo Tribunal Federal e no Superior Tribunal de Justiça".

Apesar da semelhança, observa-se que o Recurso Repetitivo é julgado pelo STJ ou TST, sobrestando os feitos do país inteiro em igual ou semelhante questão a ele, devendo seu julgamento ser observado por todos. No IRDR, no entanto, o sobrestamento do feito ocorre apenas em relação aos processos naquele Tribunal ao qual o incidente foi submetido (art. 982, inciso I, do CPC), e a aplicação do julgado será apenas aos processos daquele juízo.

Em outra vertente, tem-se o Incidente de Assunção de Competência, que se encontra regulamentado no art. 947 do CPC, conforme: "Art. 947. É admissível a assunção de competência quando o julgamento de recurso, de remessa necessária ou de processo de competência originária envolver relevante questão de direito, com grande repercussão social, sem repetição em múltiplos processos".

Ao contrário do IRDR e do RR, que exigem a repetição de processos com matéria de direito semelhante, o IAC exige que seja apenas uma questão de grande relevância, com repercussão social e sem repetição de processos.

Também é possível observar uma leve semelhança entre o IAC e o instituto da Repercussão Geral. No caso do último, a exigência é um pouco maior, pois não cabe só uma questão de relevância social, mas sim uma questão de relevância jurídica, política ou econômica, conforme disposição do art. 102, § 3º da CF. A Repercussão Geral (RG) é julgada pelo STF e o IAC pelo tribunal a que foi submetido o incidente.

Com a entrada em vigor do Código de Processo Civil, foi imbuído ao CNJ a responsabilidade de manter um Banco Eletrônico de Dados atualizados com informações específicas acerca destes incidentes: "Art. 979. A instauração e o julgamento do incidente serão sucedidos da mais ampla e específica divulgação e publicidade, por meio de registro

eletrônico no Conselho Nacional de Justiça. § 1º Os tribunais manterão banco eletrônico de dados atualizados com informações específicas sobre questões de direito submetidas ao incidente, comunicando-o imediatamente ao Conselho Nacional de Justiça para inclusão no cadastro".

Em seguida, O CNJ edita a Resolução 235 de 13/07/2016, que dispõe sobre a padronização de procedimentos administrativos decorrentes de julgamentos de repercussão geral, de casos repetitivos e de incidente de assunção de competência nos tribunais. Em seu artigo 5º menciona:

> *Fica criado, no âmbito do CNJ, banco nacional de dados com informações da repercussão geral, dos casos repetitivos e dos incidentes de assunção de competência do Supremo Tribunal Federal (STF), do STJ, do TST, do TSE, do STM, dos Tribunais Regionais Federais, dos Tribunais Regionais do Trabalho e dos Tribunais de Justiça dos Estados e do Distrito Federal.*
>
> *§ 1º O banco nacional de dados será alimentado continuamente pelos tribunais, com a padronização e as informações previstas nos Anexos I a V desta Resolução.*
>
> *§ 2º O CNJ disponibilizará as informações para toda a comunidade jurídica, separando em painéis específicos os dados relativos à repercussão geral, aos recursos repetitivos, ao incidente de resolução de demandas repetitivas e ao incidente de assunção de competência admitidos e julgados pelos tribunais.*
>
> *§ 3º A gestão das informações a que se refere o § 2º deste artigo, bem como a criação do Número Único dos Temas (NUT) de IRDR e de IAC são da competência da Comissão Permanente de Gestão Estratégica, Estatística e Orçamento do CNJ, com o apoio técnico do Departamento de Pesquisas Judiciárias (DPJ).*
>
> *§ 4º O Número Único dos Temas de IRDR e de IAC conterá as informações previstas nos §§ 4º e 5º do art. 1º da Resolução CNJ 65/2008, seguidas de um algarismo identificador do respectivo incidente, além de um número sequencial único gerado por ordem cronológica de cadastro, que será vinculado à descrição do tema, enviada pelos Tribunais Regionais Federais, Tribunais Regionais do Trabalho e pelos Tribunais de Justiça dos Estados e do Distrito Federal.*

Os artigos que versam sobre o sistema de precedentes judiciais vinculantes no CPC, IRDR e IAC, visam principalmente discutir as principais questões processuais decorrentes

desses novos institutos, tais como uniformização do julgado, segurança jurídica e celeridade processual. Dessa forma, esses novos institutos podem auxiliar na redução da morosidade na tramitação dos processos judiciais e minimizar o discrepante número de julgados divergentes sobre uma mesma questão de direito na Justiça Brasileira [11] [12] [13] [14].

Outros países também possuem mecanismos para lidar com as demandas repetitivas [15], sendo possível identificar institutos semelhantes no:

- direito americano: criou a *class action*, que agrega em uma única ação diversas lides de mesmo conteúdo. Esse mecanismo permite que um sujeito autorizado pelo juíz conduza o processo em nome da classe inteira;
- direito austríaco: possui o *Testprozess* ou demanda modelo/teste, que permite que um grupo de pessoas com pedido e interesse similares possam ingressar com somente uma única demanda;
- direito dinamarquês: apresenta a *class action* semelhante ao direito americano, devendo o pressuposto de direito ser substancialmente idêntico entre as demandas, assim como ocorre no IRDR brasileiro;
- direito alemão: apresenta como demanda coletiva o *Musterprozessführung* ou causa piloto, onde a parte propõe uma ação com a finalidade de utilizar a solução Jurisprudencial como referência para a resolução de diversas outras controvérsias, e não somente para resolver o caso específico. A comissão responsável por reformar o CPC reconheceu no documento Exposição de Motivos do CPC/2015 que o IRDR foi inspirado na sistemática do direito alemão;
- direito português: apresenta solução semelhante ao do sistema alemão somente no âmbito do contencioso administrativo;
- direito canadense e israelense: possuem mecanismo *class action* semelhante ao norte-americano.

Não foram encontrados na literatura especializada, no entanto, artigos que apliquem técnicas de mineração de texto na identificação de precedentes judiciais semelhantes ou que fazem relação desses novos institutos com a insegurança jurídica na 2ª instância.

Essas técnicas têm sido utilizadas em dados do Poder Judiciário para calcular as taxas de reforma de decisão em câmaras de direito criminal do Tribunal de Justiça de São Paulo por meio de expressões regulares [16]; para construir processos de descoberta de conhecimento em textos na área do direito [17]; para predizer se decisões judiciais serão condenatórias ou não [18]; ou para melhorar os métodos de procura de documentos judiciais [19].

2.2 Técnicas de mineração de texto

Os precedentes judiciais constantes no BNPR não são estruturados de maneira que se possa identificar de forma trivial precedentes semelhantes, ou seja, utilizar a descrição do precedente para realizar tal tarefa.

Algoritmos de mineração de textos podem ser utilizados para resolver esse problema, sendo a principal característica dos dados textuais a esparsidade e a alta dimensionalidade [20], pois a matriz de documentos por atributos (palavras ou termos) é geralmente grande e formada em sua grande maioria por valores nulos.

A identificação dos precedentes semelhantes é considerada uma forma de recuperação de informação, do inglês *Information Retrieval* (IR), uma vez que deseja-se extrair automaticamente informações associadas a estes dados que apresentam natureza não estruturada [21].

Cada documento da base de dados apresenta um precedente e é delimitado de forma a apresentar a sua descrição conforme o tribunal, tipo e número do precedente. Após a delimitação dos documentos, podem ser realizadas as etapas de *tokenization*, exclusão dos *stopwords*, normalização e *stemming*, conforme descritos a seguir:

- *tokenization*: consiste em separar, inicialmente, a descrição do precedente por palavras, desconsiderando os símbolos e pontuações.
- *stop words*: são as palavras muito comuns que agregam, em princípio, pouco valor na identificação dos documentos semelhantes, como artigos e preposições.
- normalização: método que permite que palavras com diferenças superficiais possam se corresponder, seja por apresentar acento opcional, como na forma verbal dêmos/demos, seja por alteração do novo acordo ortográfico, como ocorrido na palavra contra-regra para contrarregra, ou por erros de ortografia. Também deve-se levar em conta a acentuação, uma vez que seu uso pode alterar significativamente o significado da palavra, como por exemplo secretária e secretaria. Além disso, é importante que sejam criados relacionamentos entre *tokens* não normalizados por intermédio de uma lista de sinônimos, acrônimos e sinônimos (*thesaurus*). Parte dos acrônimos podem ser identificados pelo uso de letras maiúsculas no texto.
- *stemming*: consiste em um processo de transformar, em sua grande maioria reduzir, as palavras flexionadas a uma forma concisa, em geral excluindo os sufixos e/ou prefixos e as flexões verbais, como por exemplo, as palavras inclusão, incluir, incluso podem ser transformadas para *incluir*.

O resultado desse tratamento é uma lista de termos por documento. Uma forma de ponderar os termos em um documento é considerar a frequência que ocorrem, cuja notação

é $tf_{t,d}$, onde t é o termo e d o documento. Essa forma de representação dos documentos onde a ordem dos termos é desconsiderada e cada termo considera o número de ocorrências é conhecida na literatura como modelo *bag of words* [21].

Como forma de atenuar o efeito de termos que ocorrem frequentemente em diversos documentos, a ponderação dos termos frequentes deve ser reduzida conforme o aumento da frequência dos termos na coleção de documentos. Esse método é denominado de *inverse document frequency* (idf) e geralmente definida como o logaritmo da frequencia total de um termo nos documentos em relação à frequência do termo em um determinado documento t, conforme a fórmula:

$$idf = log(N/df_t) \tag{2.1}$$

A combinação da definição da frequência dos termos com o inverso da frequência dos documentos é denominada de ponderação *tf-idf*$_{t,d}$. Há diversas formas de ponderação dos termos e documentos baseadas na função *tf-idf*, sendo um dos principais esquemas de ponderação denominado *Okapi BM25*, que será a utilizada neste estudo com a denominação de BM25. Este método é considerado como o estado da arte na recuperação de informações e leva em consideração no cálculo o tamanho dos documentos [21].

A combinação dos documentos e dos termos como uma matriz é conhecida como *term-document matrix*. Uma das principais formas de calcular similaridade entre cada par de documentos é pelo cálculo da similaridade cosseno, que, por intermédio do modelo de vetor espacial, realiza a divisão do produto dos vetores (V) de dois documentos (d) em relação ao produto de suas distâncias euclidianas, conforme notação a seguir:

$$sim(d_1, d_2) = (\vec{V}(d_1).\vec{V}(d_2))/(|\vec{V}(d_1)||\vec{V}(d_2)|) \tag{2.2}$$

Os precedentes cadastrados no banco de dados não possuem classes específicas que permitam a verificação de similaridade entre eles, não se tratando portanto, de um problema de classificação supervisionado. Este estudo baseia-se no método de similaridade de documentos por agrupamento (*clustering*). O método utilizado é considerado como aprendizado não supervisionado, uma vez que nenhum modelo de rotulação ou informações adicionais são dados ao algoritmo de aprendizagem.

Murugan e Karthika [22] realizaram uma recente revisão de literatura das técnicas e métodos de mineração de texto. Foram descritos os passos envolvidos no processo global de mineração de texto, tais como: pré-processamento e transformação do texto, seleção dos atributos, aplicação das técnicas de mineração de texto e avaliação.

Na mesma linha, Kweku-Muata [23] ilustra em seu estudo como um modelo de processo genérico de mineração de dados pode ser adaptado para realizar análise de agrupamentos.

As análises realizadas foram baseadas no modelo *Cross Industry Standard Process for Data Mining (CRISP-DM)* [24], que é o modelo de referência geralmente utilizado pelos especialistas para a mineração de dados.

Há, em geral, duas formas de realizar os resultados de uma clusterização [25], a primeira é com o uso de medidas estatísticas, a segunda é com a classificação dada por um especialista. Assim, é possível calcular a precisão obtida, que é o número de precedentes agrupados corretamente em relação ao total agrupado, e *recall*, também conhecido como revocação, que é o número de precedentes agrupados corretamente em relação ao total que deveria ter sido agrupado. Logo, é possível calcular a métrica *F-measure*, que é a média harmônica dessas métricas.

$$precisao = \frac{(relevantes \cap recuperados)}{(recuperados)} \tag{2.3}$$

$$recall = \frac{(relevantes \cap recuperados)}{(relevantes)} \tag{2.4}$$

$$F\text{-}measure = \frac{(2 * precisao * recall)}{(precisao + recall)} \tag{2.5}$$

Sinoara et al. [26] realizaram um extenso mapeamento da literatura sobre estudos de mineração de texto relacionados à semântica e identificaram que o método mais utilizado é o *Latent Semantic Indexing (LSI)*. Outra técnica que é comumente usada para modelagem de tópico é *Latent Dirichlet Allocation (LDA)*. A Figura 2.1 apresenta as principais tarefas de mineração de textos identificadas nesse mapeamento. Observa-se que o tópico similaridade de documentos, que é o foco deste trabalho, está presente em somente 4% dos estudos analisados.

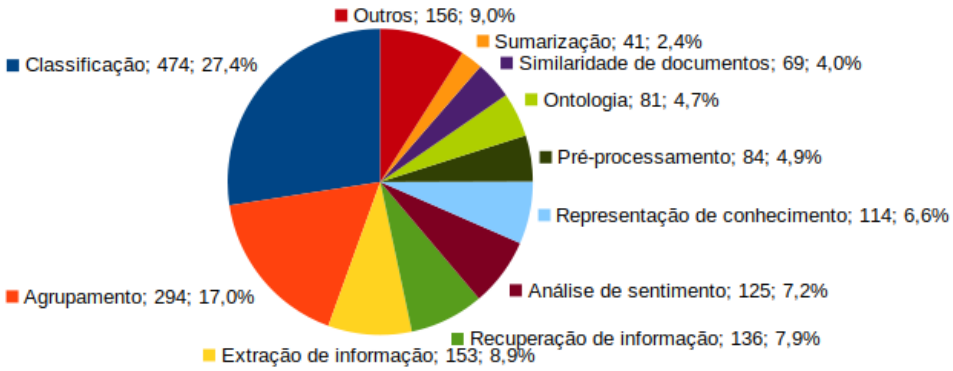

Figura 2.1: Tarefas de mineração de textos.

Fonte: Adaptado de Sinoara et al. [26]

Allahyari et.al [27] realizaram um levantamento das técnicas de mineração de texto baseadas em classificação, agrupamento e técnicas de extração. Segundo esse estudo, a maneira mais comum de representar o documento é com *Bag of Words (BOW)*, considerando o número de ocorrências de cada termo, independente da ordem de ocorrência. Tem-se, ainda, que as 3 principais técnicas de redução de dimensionalidade utilizadas na mineração de texto são *Probabilistic Latent Semantic Indexing (PLSI)*, *Latent Semantic Indexing (LSI)* e *Latent Dirichlet Allocation (LDA)*, corroborando o estudo de Sinoara [26].

A técnica LSI utiliza *Singular-Value Decomposition (SVD)*, que é a decomposição de uma matriz m x n em k dimensões. Dessa forma, uma grande matriz de termos por documentos é reduzida a uma matriz de documentos por k dimensões, que são combinações lineares da matriz original. Essa técnica permite a associação de termos com significados semelhantes, levando-se em consideração a questão semântica dos termos e não somente a sintática. Assim, reduz o ruído causado pela sinonímia e polissemia, lidando latentemente com semântica de texto [28].

A técnica PLSI é semelhante ao LSI, entretanto, essa técnica apresenta uma base mais sólida na inferência estatística com a utilização de probabilidade na decomposição fatorial, mais precisamente com o uso do algoritmo *Expectation–maximization (EM)*, que é um método iterativo para encontrar estimativas de máxima verossimilhança em um modelo com variáveis latentes não observadas [29].

A técnica LDA é um modelo bayesiano hierárquico de três níveis, no qual cada item de uma coleção é modelado como uma mistura finita sobre um conjunto subjacente de tópicos. [30]. Esse algoritmo pressupõe que cada documento é uma mistura de um pequeno número de tópicos e que a presença de cada palavra é atribuível a um dos tópicos do documento.

Já Mikolov et.al [31] propõem duas novas técnicas para computar representações vetoriais de palavras a partir de conjuntos de dados muito grandes. Os resultados obtidos por essas técnicas foram comparados a outros modelos baseados em redes neurais utilizando-se a similaridade de palavras. Foram observadas melhorias consideráveis na precisão e com um custo computacional muito menor para medir semelhanças de palavras sintáticas e semânticas em um conjunto de dados de 1,6 bilhão de palavras.

Ainda segundo os autores, os métodos usados para extração de vetores de palavras (*Word2Vec*) aprendidas por redes neurais apresentam resultados significativamente melhores do que o *LSI*. Além disso, o *LDA* apresenta um custo computacional muito alto em grandes conjuntos de dados.

Jacob Devlin et al. [32] introduziram um novo modelo denominado *Bidirectional Encoder Representations from Transformers (BERT)*, que apresenta bons resultados em diver-

sas tarefas de processamento de linguagem natural, como resposta a perguntas e inferência de palavras faltantes. Os pesquisadores do *GOOGLE* disponibilizam bases pré-treinadas que podem ser utilizadas com esse modelo[3].

Além da pesquisa qualitativa, o levantamento do estado da arte também foi realizado utilizando o método Teoria do Enfoque Meta Analítico Consolidado (TEMAC). Esse método é dividido em 3 etapas [33]:

- preparação da pesquisa: escolha da base de dados de pesquisa, palavra-chave; área; e espaço temporal da pesquisa;
- apresentação e interrelação dos dados: analisar a evolução do tema ano a ano, autores com mais publicações e os mais citados, os países que mais publicaram, frequência de palavras-chave;
- detalhamento, modelo integrador e validação por evidências: realizada a co-ocorrência de palavras-chave para identificar as principais similaridades dos artigos recuperados pela pesquisa.

A pesquisa foi realizada nas bases de dados *Web of Science* e *Scopus* no dia 09 de março de 2019 utilizando o termo *document similarity* e considerando os artigos publicados na área de Ciência da Computação.

A pesquisa retornou 341 resultados na base de dados *Web of Science* e 645 na base *Scopus*. Tendo em vista o excesso de artigos duplicados ao considerar as duas bases, optou-se por analisar somente os estudos da base *Scopus*, pois ela apresenta quase o dobro de artigos, o artigo mais citado possui 3,5 vezes mais citações e foram publicados no ano de 2018 2,6 vezes mais artigos do que na base *Web of Science*.

O estudo *Some experiments in the generation of word and document associations*, de Gerard Salton, publicado no ano de 1962 [34], foi o mais antigo cadastrado em ambas as bases. Este estudo menciona que, a maioria dos problemas de recuperação automática de informações na época baseava-se em procedimentos que usam a frequência de ocorrência de certas palavras ou classes de palavras. O número de publicações por ano permaneceu abaixo de 10 publicações até o ano de 2003, apresentando posterior crescimento até atingir 63 publicações em 2018. Os Estados Unidos da América constam como o país com maior número de publicações na área, com 129 artigos, seguido pela China, com 113 publicações. Entretanto, ao considerar o número de citações, os artigos chineses foram citados 1.283 vezes, enquanto que os americanos 1.171 .

Com relação aos autores, Oren Kurlend (israelense), Xiaojun Wan (chinês) e Jianwu Yang (chinês) apresentaram os maiores quantitativos de artigos publicados com, respectivamente, 10, 8 e 6 publicações na área. O brasileiro Danilo Medeiros Eler, da Universidade

[3] `https://github.com/google-research/bert`

Estadual Paulista (UNESP) publicou 4 artigos e consta entre os 10 autores com maior número de publicações. O espanhol Paolo Rosso consta como o autor mais citado, com 39 citações, seguido pelo italiano Alessandro Moschitti (29 citações) e o japonês Masao Fuketa (27 citações).

A Figura 2.2 apresenta o correlacionamento das palavras-chave postas nos artigos que versam sobre similaridade de documentos, onde os nós representam a quantidade de palavras-chaves inseridas pelos autores e as arestas a indicação que as palavras-chaves são informadas conjuntamente nos artigos. É possível observar que há semelhanças deste trabalho com o grupo de estudos que utilizam as técnicas LSI e LDA, sendo ainda bastante utilizada a similaridade cosseno para identificação de documentos semelhantes. Logo, este trabalho utiliza as técnicas acima mencionadas, baseando-se no modelo de referência CRISP-DM para mineração de dados.

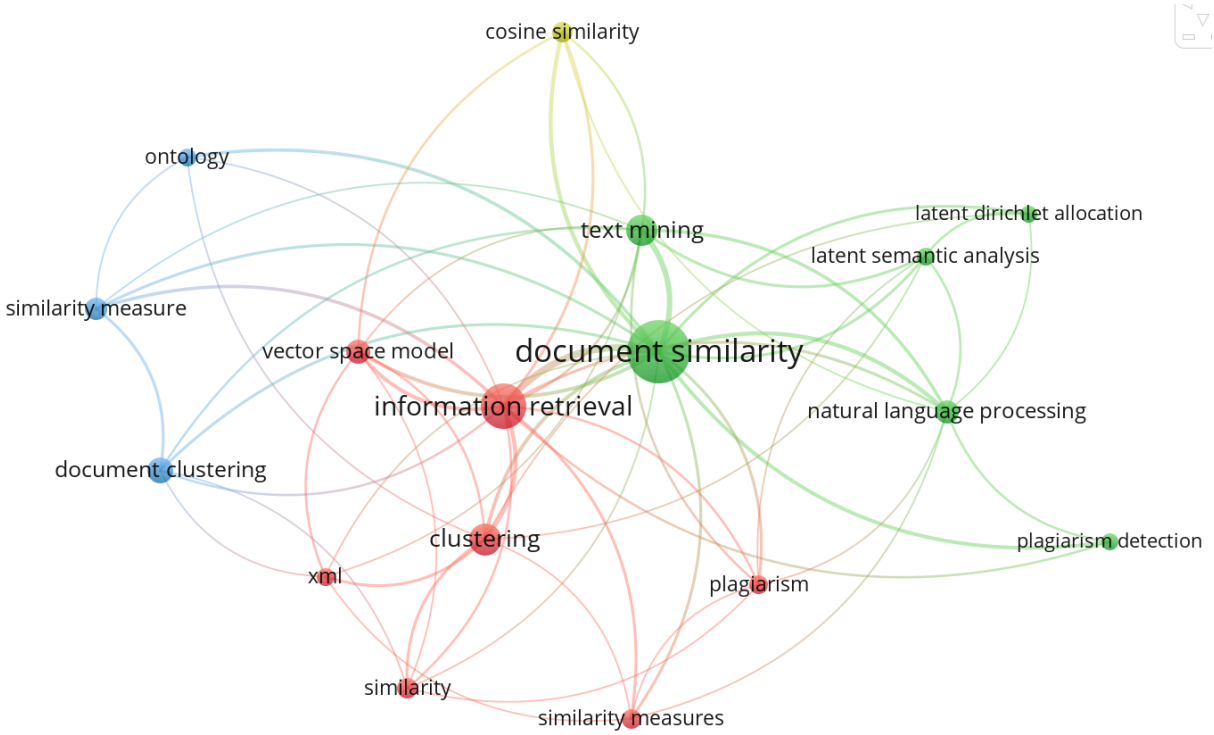

Figura 2.2: Co-ocorrência das palavras-chave.

Freire et.al [35] apresentam um *framework* que combina componentes com *parsers* morfológicos e sintáticos, bases de conhecimento e lexicais, algoritmos de aprendizagem automática e de alinhamento e ainda cálculo da similaridade, conforme Figura 2.3. Esse *framework* é adaptado no presente estudo para extrair entidades nomeadas na descrição dos precedentes, sendo utilizado o banco de dados BNPR como base, os modelos BM25, LSI e LDA como algoritmos de identificação de similaridade entre termos, o tesauro do

STF como base de dados léxica, a extração das normas com expressões regulares e os pacotes *quanteda* [36] e *udpipe* [37] como ferramentas de *tokenization* e *POS-Tag*.

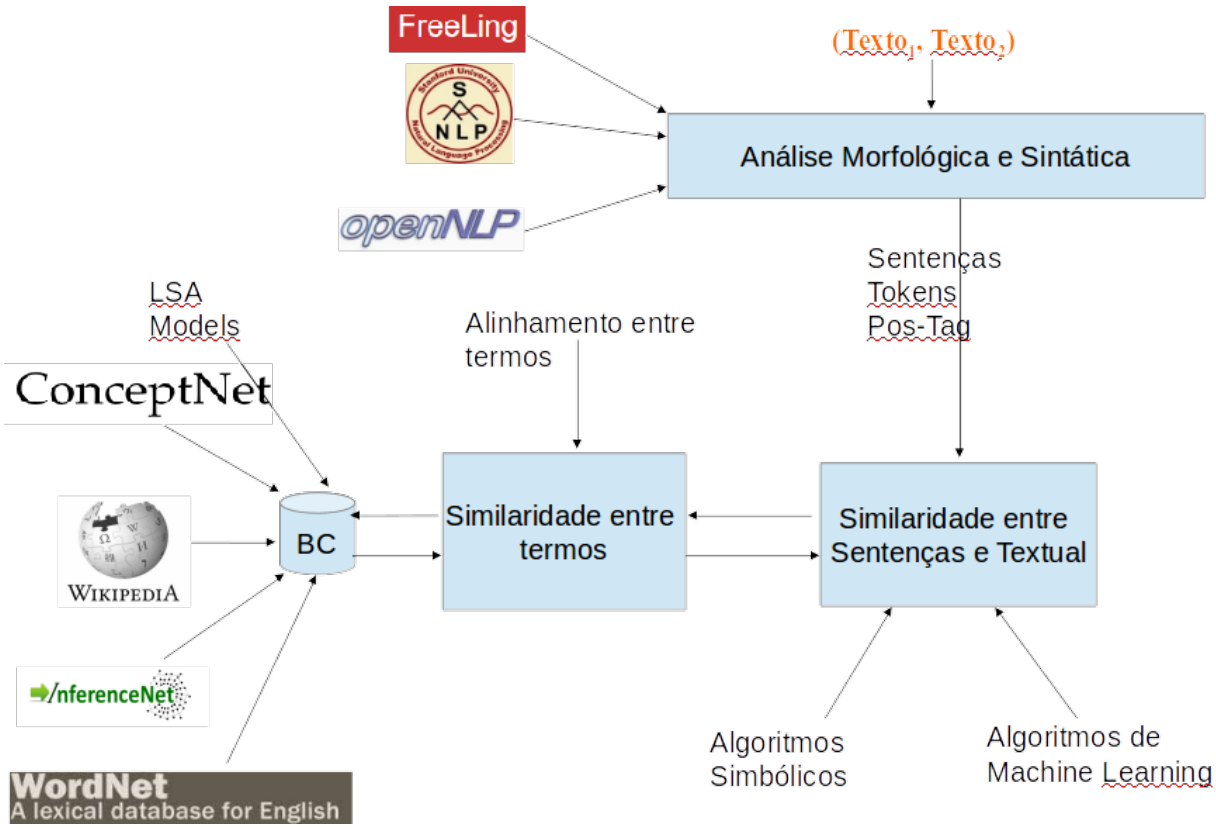

Figura 2.3: Framework para identificação de similaridade de sentenças.

Fonte: Adaptado de Freire et al. [35]

Capítulo 3

Solução Proposta

Este Capítulo apresenta os métodos utilizados no desenvolvimento deste estudo, tanto para o levantamento do estado da arte, quanto para a etapa de recuperação de informações. Há ainda um detalhamento das etapas dos experimentos realizados.

3.1 Metodologia

O levantamento do estado da arte é realizado por intermédio de pesquisa nas bases de referência *Scopus* e *Web of Science* utilizando o método Teoria do Enfoque Meta Analítico Consolidado (TEMAC). O TEMAC é dividido em 3 etapas [33]: i) preparação da pesquisa, ii) apresentação e interrelação dos dados, e iii) detalhamento, construção do modelo integrador e validação por evidências.

A preparação se dá com a escolha da base de dados, palavras-chave, área, e do espaço temporal para a pesquisa. A interrelação dos dados é realizada analisando a evolução do tema ano a ano, os autores com mais publicações e os mais citados, os países que mais publicaram, e a frequência de palavras-chave. Por fim, a co-ocorrência de palavras-chave é utilizada para identificar as principais similaridades dos artigos recuperados pela pesquisa.

Para mineração dos dados, o *Cross Industry Standard Process for Data Mining (CRISP-DM)* [24] é utilizado como modelo de referência (Figura 3.1).

A fase de **entendimento do negócio** visa entender o problema e os recursos disponíveis, traçar os objetivos e definir um plano de ação. Os dados utilizados neste estudo são públicos e estão disponíveis no portal do CNJ, compondo a base do BNPR.[1]

A fase de **entendimento dos dados** tem por objetivo a familiarização com as informações constantes na base de dados, realizando as primeiras análises descritivas e identificando possíveis inconsistências.

[1] `http://www.cnj.jus.br/pesquisas-judiciarias/demandas-repetitivas`

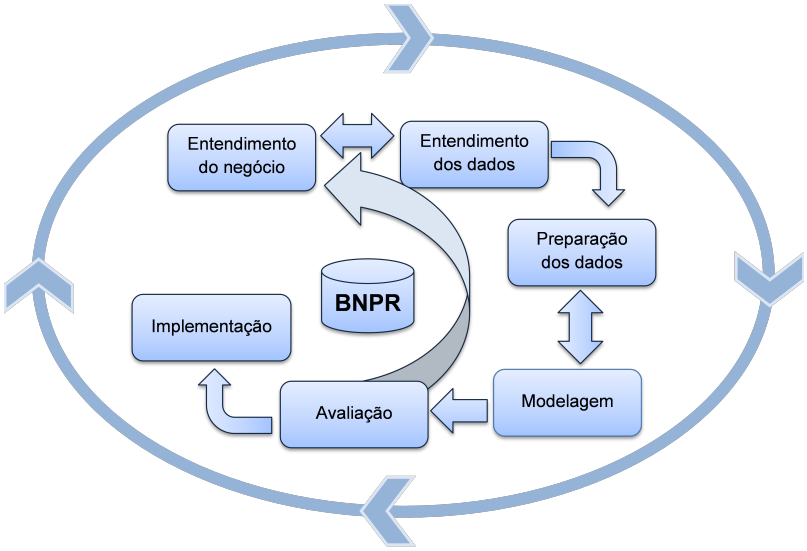

Figura 3.1: Fases do CRISP-DM.

A base de dados do BNPR possui informações por processo judicial, sendo aproximadamente 2.500 processos em grau de recurso e mais de 2,2 milhões de processos aguardando o julgamento dos processos paradigmas. A base de dados utilizada possui informações referentes a:

- número do precedente: faz parte da chave de identificação do precedente judicial em conjunto com o tribunal e o tipo de instituto;
- tribunal: faz parte da chave de identificação do precedente judicial em conjunto com o tipo de instituto e o número do precedente;
- tipo de instituto: pode conter os valores RG, RR, IRDR e IAC. Faz parte da chave de identificação do precedente judicial em conjunto com o tribunal e o número do precedente;
- assunto do processo: estabelecido pelas Tabelas Processuais Unificadas do CNJ. Essa informação é induzida no modelo para verificar se há melhora no agrupamento de precedentes;[2]
- tipo do processo: apresenta duas possibilidades: paradigma, processo principal que sobe em grau de recurso; e sobrestado, que é processo aguarda decisão do processo principal. Somente os processos paradigmas são utilizados no modelo;
- situação do precedente: campo que denota se o processo paradigma foi admitido, recusado, julgado, julgado em definitivo, arquivado, está pendente, entre outros.

[2] `http://www.cnj.jus.br/programas-e-acoes/tabelas-processuais-unificadas`

É utilizado no modelo para verificar se o precedente judicial foi admitido, está sobrestado (suspenso) ou se foi julgado, desde que apresente uma tese.

- título do precedente: descrição da questão pelo qual o recurso foi interposto. Essa é a principal variável na identificação dos precedentes semelhantes, uma vez que os termos desse campo são utilizados como atributos do modelo. Desse campo também são extraídas as palavras-chave, os atos normativos citados e as entidades nomeadas;
- tese: descrição do julgamento do precedente. Não é utilizada no modelo para identificação de precedentes semelhantes, mas é utilizada posteriormente para verificação de precedentes semelhantes com teses divergentes;

A fase de **preparação dos dados** consiste em construir uma base de dados mais trabalhada, como:

- excluir da base de dados os precedentes cancelados, recusados ou sem descrição do precedente;
- excluir partes dos títulos dos precedentes do Tribunal de Justiça de São Paulo que estão em formato padrão, mantendo somente a descrição dos precedentes;
- gerar o *corpus* contendo as descrições dos precedentes por documento;
- gerar os *tokens*, removendo pontuações, símbolos, *stopwords*, palavras com menos de 3 letras;
- excluir as palavras que aparecem somente uma vez em toda a base de dados, pois atributos únicos não trazem ganho ao modelo.
- gerar a matriz de atributos por documento, que nesse caso é a matriz de termos por precedente.

Na fase da **modelagem** são aplicadas as técnicas que utilizam abordagem sintática e semântica. O processo de criação de modelos está representado na Figura 3.2. Após a preparação dos dados são criadas matrizes precedente-atributo aplicando frequência de palavras (BOW), BOW ponderada com BM25, LSI, LSI ponderada com BM25 e LDA. A partir dessas matrizes são calculados cossenos entre dois precedentes, gerando matrizes de similaridade entre eles. Então é aplicado o ponto de corte do valor do cosseno para determinar quais precedentes serão efetivamente considerados como similares. Estes são apresentados como grafos. O melhor dentre os modelos gerados é avaliado considerando uma normalização de termos para compor as matrizes precedente-atributo utilizando entidades nomeadas.

As entidades nomeadas utilizadas no modelo são obtidas com:

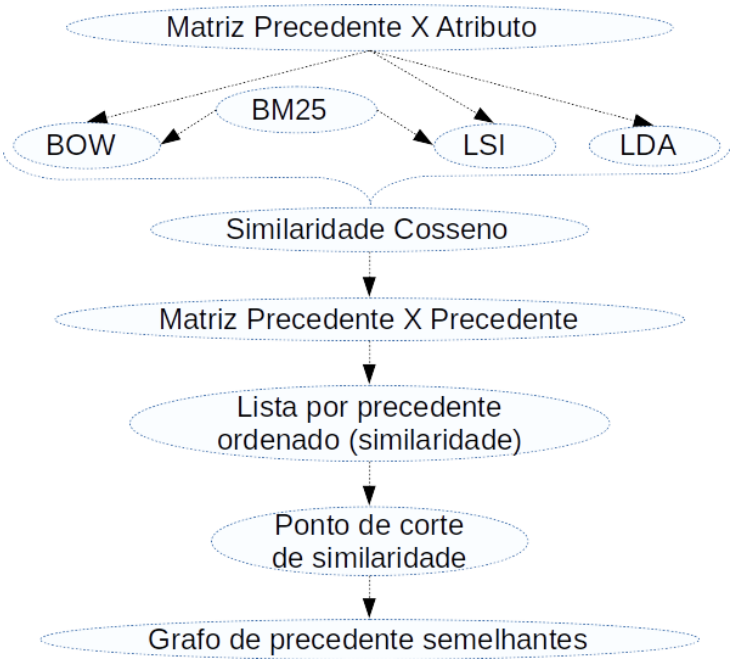

Figura 3.2: Fluxograma de modelagem do modelo.

- **tesauro do Supremo Tribunal Federal (STF)**: dicionário contendo entidades nomeadas e sinônimos;[3]
- **normas**: identificação por meio de expressões regulares de artigos de leis, decretos, protocolos, estatutos, emendas constitucionais, medidas provisórias, portarias, instruções normativas e códigos;
- **expressões com a primeira letra maiúscula**: extraídas da descrição dos precedentes por intermédio de expressões regulares e uso de n-gramas;
- **substantivos próprios**: os termos são etiquetados usando *POS-tagging* (do inglês *Part-Of-Speech tagging*) aplicando o pacote R de processamento de linguagem natural *udpipe*[37]. Os substantivos próprios são extraídos utilizando operadores de proximidade.

A fase de **avaliação** é importante para identificar qual modelo apresenta resultados melhores que os demais. Essa avaliação é realizada de forma experimental, utilizando métricas de aferição de similaridade e o auxílio de um especialista, que verificará se existe ou não similaridade entre os precedentes de forma manual. A Resolução Nº. 235/2016 do CNJ prevê que os tribunais e o CNJ devem organizar, como unidade permanente, o Núcleo de Gerenciamento de Precedentes (Nugep).

[3]Acessado em: 01/06/2019. http://www.stf.jus.br/portal/jurisprudencia/pesquisarVocabularioJuridico.asp

A métrica *F-measure* é utilizada como suporte para definir o ponto de corte dos grupos a serem rotulados pelo especialista, para comparar os modelos, para definir o número de dimensões dos modelos que utilizam redução de dimensionalidade, e para definir o ponto de corte que define os agrupamentos do modelo final. A Figura 3.3 ilustra as etapas da fase de avaliação.

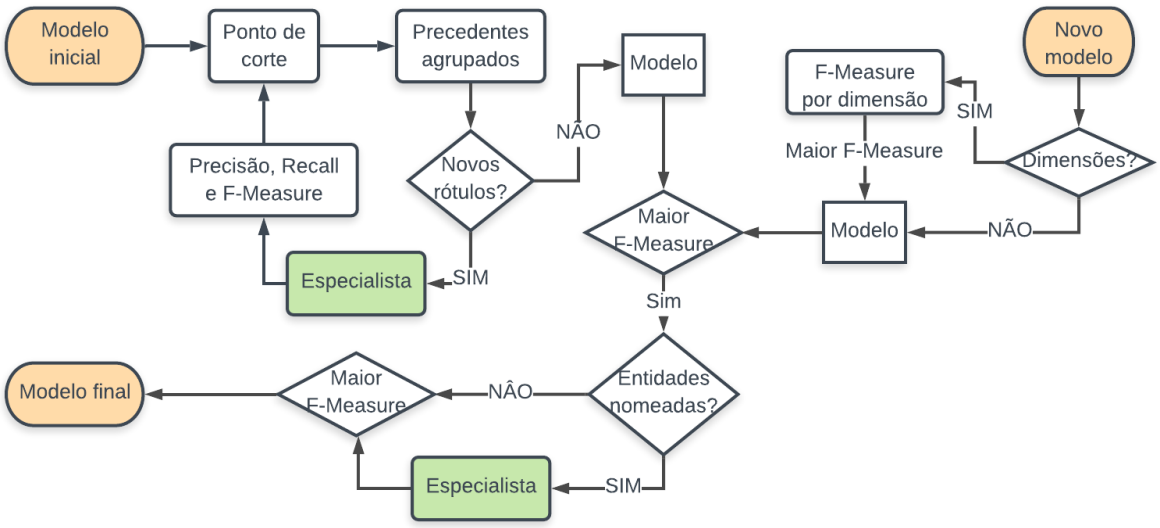

Figura 3.3: Fluxograma de avaliação do modelo.

Por fim, na fase de **implantação**, uma prova de conceito é apresentada ao CNJ com o intuito de utilização do modelo na instituição. Como a base de dados é atualizada diariamente, o modelo é incrementado e verifica-se o percentual de precedentes agrupados corretamente. O fluxograma de atualização pode ser visualizado na Figura 3.4. Observa-se que o modelo agrupa os precedentes e, caso não pertençam a algum grupo existente ou não apresentem alta similaridade, o especialista deve rotulá-los para que as métricas de avaliação sejam devidamente calculadas. Já com relação às entidades nomeadas, elas somente são consideradas no modelo após a aprovação do especialista.

Tendo em vista que um precedente "A" pode ser semelhante a "B", e "B" ser semelhante a "C", são utilizados grafos para visualização e identificação de grupos de precedentes semelhantes.

O software R é utilizado para realizar as análises de mineração de texto [38]. A ferramenta *Quanteda (Quantitative Analysis of Textual Data)* [36] é adicionada ao R para realizar a análise de maneira mais rápida e eficiente, uma vez que esta ferramenta é implementada em C++, realiza o processamento em paralelo nos processadores da máquina e apresenta gerência de memória eficiente na estruturação dos dados.

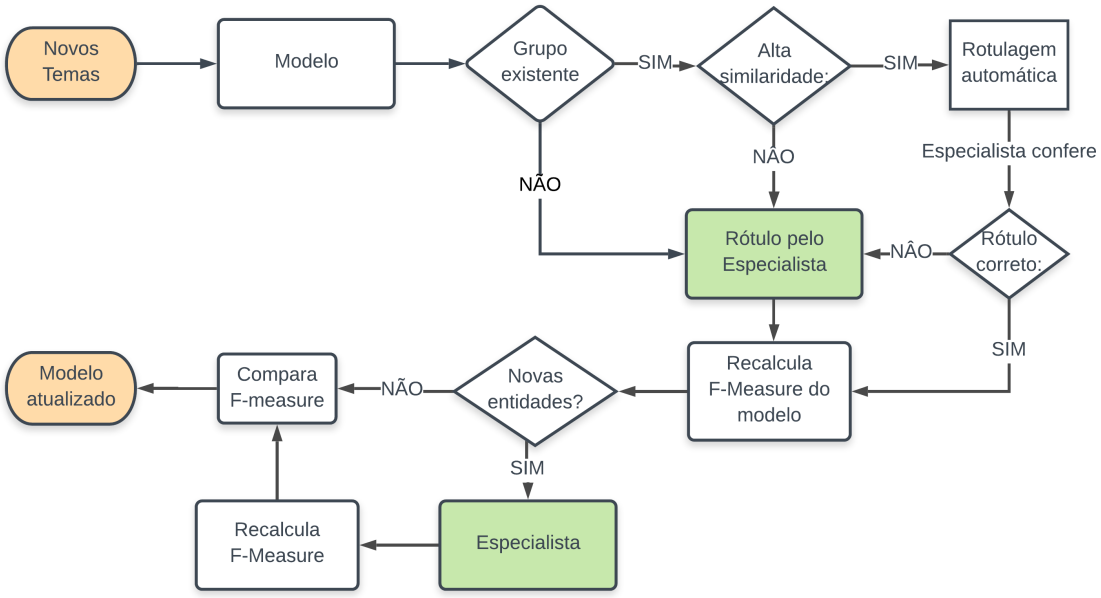

Figura 3.4: Fluxograma de atualização do modelo.

3.2 Experimentos

Esta seção detalha os passos utilizados no agrupamento dos precedentes judiciais.

1. ***Download* dos dados.** Entrar no site do CNJ onde constam os painéis públicos (`http://www.cnj.jus.br/pesquisas-judiciarias/paineis`) e clicar no painel de *Demandas Repetitivas*. Em seguida, filtrar pelos tipos de incidente RG, RR, IRDR e IAC e clicar na aba superior em *Pesquisa textual*. Logo após, clicar com o botão direito do cursor na tabela que contém os temas de demandas repetitivas e precedentes obrigatórios e exportar os dados no formato *csv*. Os dados utilizados nessa pesquisa foram extraídos no dia 14 de junho de 2019.

2. **Importar e transformar em banco de dados.** Importar a base de dados no software R e agrupar o campo assunto de forma que os assuntos de determinado precedente estejam em uma única célula separados por ponto e vírgula. Dessa forma, cada linha da base transformada representa um precedente único, totalizando 2.472 precedentes.

3. **Limpeza da base de dados.** Excluir os registros cuja situação esteja como cancelado ou recusado e que não apresentem a descrição do precedente. Após a limpeza o total de precedentes baixou para 2.287. Os precedentes ingressados no ano de 2019 são utilizados somente na etapa final de atualização do modelo, restando, portanto, 2.260 precedentes.

4. **Tratamento do campo descrição do precedente.** O campo descrição dos precedentes contém de 24 a 7.395 caracteres. Em apenas 4 dos 45 tribunais da base de dados esse campo apresenta mais de mil caracteres. O TJSP apresenta 5 precedentes com mais de três mil caracteres, destoando dos demais tribunais, conforme observado na Figura 3.5. A descrição do tema propriamente dito destes precedentes é obtida com a utilização de expressões regulares, uma vez que no TJSP esse campo de descrição apresenta diversas outras informações. Após esse tratamento, a variável de descrição dos precedentes varia de 24 a 2.597 caracteres. O novo acordo ortográfico entrou em vigor em janeiro de 2016 com o objetivo de facilitar a comunicação entre os países de lingua portuguesa, com isso, diversas palavras não são mais acentuadas. Assim, os acentos são desconsiderados neste trabalho e todas as letras são alteradas para minúsculas.

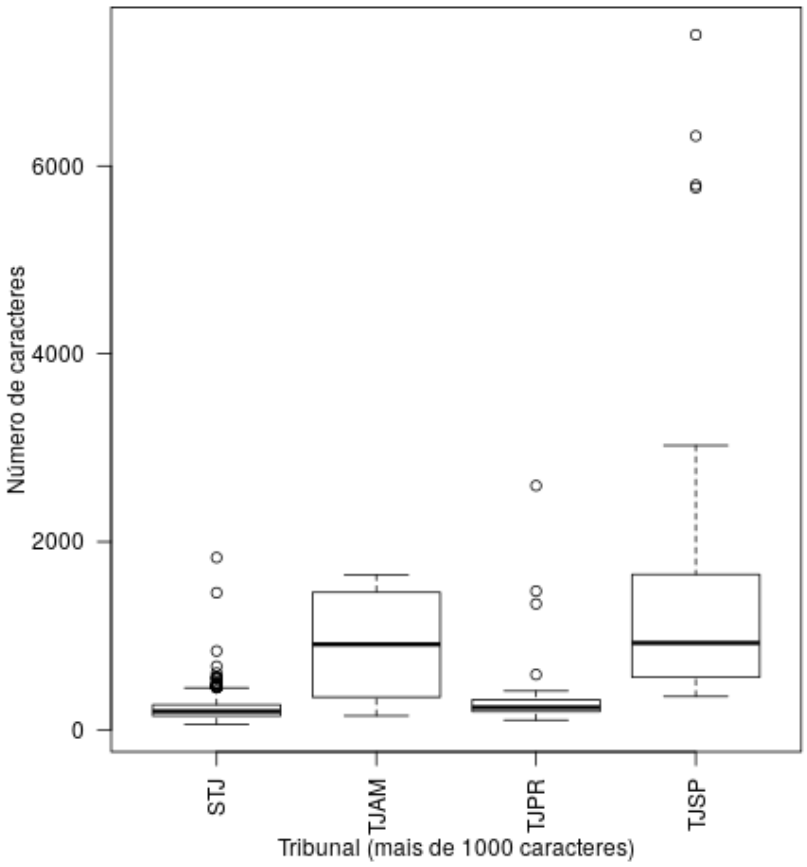

Figura 3.5: Boxplot do número de caracteres dos títulos dos precedentes.

5. **Análise léxica: construir *corpus e tokens*.** A remoção de sinais de pontuações e símbolos resulta em 2.260 documentos e 6.964 atributos. A exclusão de *stopwords* e palavras com menos de 4 caracteres resulta em 6.305 atributos. A exclusão de

palavras que aparecem somente uma única vez na base de dados resulta em 3.370 atributos.

6. **Ponderação dos termos.** Os termos são ponderados conforme a função de ordenamento BM25 com parâmetros padrões de $k = 1{,}2$ e $b = 0{,}75$.

7. **Similaridade dos documentos.** A similaridade cosseno é utilizada na identificação de precedentes semelhantes utilizando-se o cálculo do cosseno do ângulo entre 2 vetores. Documentos semelhantes apresentam valores próximos a 1.

8. **Ponto de corte de similaridade.** Como cada documento apresenta um vetor contendo o valor da similaridade em relação aos demais, deve ser estipulado um ponto de corte de forma que se possa identificar os documentos semelhantes.

9. **Rotulagem dos dados.** Os agrupamentos realizados são encaminhados ao especialista, que verifica se os precedentes foram agrupados de forma correta e insere um rótulo a cada precedente judicial.

10. **Métricas de avaliação.** A partir dos dados rotulados, é possível calcular precisão, *recall* e a métrica *F-measure*.

11. **Definição do número de dimensões.** A métrica *F-measure* é utilizada para definir o número ideal de dimensões. Escolhe-se o modelo que apresentar o maior valor da métrica *F-measure*;

12. **Comparação dos modelos.** Os modelos são comparados com base no valor da métrica *F-measure* obtida para cada um deles.

13. **Indução com entidades nomeadas.** As entidades nomeadas são identificadas utilizando-se o tesauro do STF, as normas e nomes próprios extraídos dos textos dos precedentes. Na indução dos modelos também são considerados os assuntos (metadados) cadastrados no processo.

14. **Comparação de modelos induzidos considerando entidades nomeadas.** Os modelos são comparados com os anteriores levando-se em consideração o valor da métrica *F-measure*.

15. **Incremento do modelo.** Quando a base de dados BNPR é atualizada, o modelo que apresenta maior valor da métrica *F-measure* é escolhido.

16. **Visualização do modelo.** Gráficos de rede / grafos dinâmicos são utilizados para melhor visualização e identificação de grupos de precedentes semelhantes.

Capítulo 4

Resultados e análise

Esse Capítulo apresenta os resultados obtidos pela aplicação de técnicas de mineração de texto e processamento de linguagem natural, sendo subdividido em abordagem sintática, abordagem semântica, indução do modelo com entidades nomeadas, modelo final e aprendizagem incremental do modelo.

Inicialmente, os precedentes judiciais ingressados até o final do ano de 2018 são agrupados utilizando-se a função de ranqueamento *Okapi Best Matching 25 (BM25)*, que é bastante utilizada no processamento de linguagem natural. Os agrupamentos são então analisados pelo especialista e são adicionadas palavras-chave para cada precedente agrupado.

Em seguida, são aplicadas as técnicas LSI e LDA, que utilizam abordagem semântica ao tender a agrupar termos sinônimos em uma mesma dimensão. Os resultados obtidos são então comparados aos do BM25 por meio da métrica *F-measure*.

O modelo de melhor resultado é escolhido para ser induzido com a utilização de entidades nomeadas identificadas na descrição dos precedentes, como normas e nomes próprios. O tesauro disponibilizado pelo Supremo Tribunal Federal (STF) também é utilizado como um dicionário.

Os resultados obtidos são novamente comparados aos do modelo anterior. A partir do modelo final, é possível destacar os efeitos do Código de Processo Civil (CPC) em relação aos precedentes judiciais.

Por fim, o modelo é incrementado com os dados dos precedentes ingressados no ano de 2019, sugerindo novos agrupamentos e identificando novas entidades nomeadas.

4.1 Abordagem sintática

Após a realização do pré-processamento das descrições dos precedentes, construção do *corpus*, *tokenization*, remoção de *stopwords*, normalização e seleção de atributos, é gerada uma matriz de precedentes por atributos de tamanho 2.260 por 3.370. Os elementos dessa matriz são ponderados por intermédio da função de ranqueamento BM25.

A função de similaridade cosseno é então utilizada na matriz ponderada para gerar uma matriz de similaridade entre precedentes (2.260 por 2.260). Assim, é possível identificar os precedentes com maior similaridade em relação aos demais, sendo mais semelhantes os que apresentam valores próximos a 1.

Como forma de reduzir o custo computacional, cada precedente foi comparado aos 20 de maior similaridade. Dessa forma, analisou-se a relação de 2.260 documentos em relação a 20 (total de 45.200 comparações), ao invés de 2.259 (total de 5.105.340 comparações).

A Figura 4.1 apresenta o histograma dos valores de similaridade observados no vigésimo ponto de maior similaridade. Verifica-se que a maior parte dos precedentes são comparados aos que possuem ponto de corte acima de 0,12, enquanto que todos são comparados aos precedentes com ponto de corte acima de 0,3. O ponto de corte 0,3 agrupa mais de 260 precedentes e somente a metade é agrupada corretamente, conforme rotulação realizada pelo especialista. A escolha de comparar cada precedente somente aos 20 de maior similaridade é, portanto, adequada.

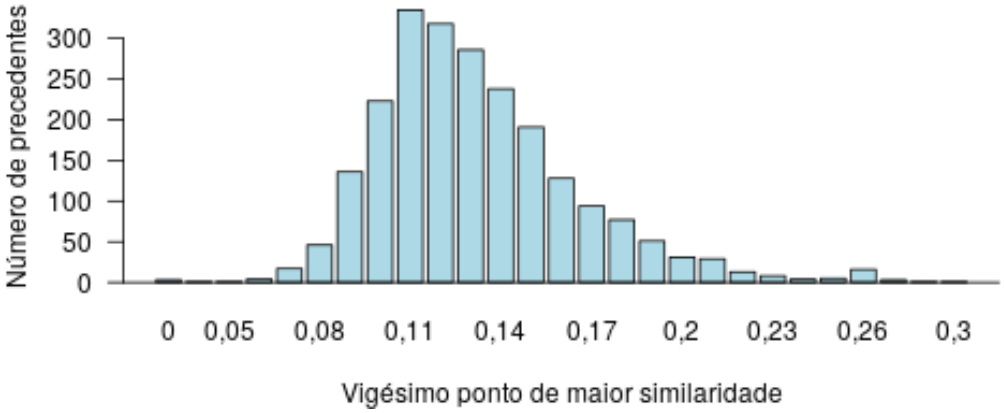

Figura 4.1: Histograma dos valores de similaridade de *rank* 20.

É necessário escolher um ponto de corte dos valores de similaridade de forma que os documentos com valores superiores ao ponto escolhido possam ser considerados semelhantes. A Figura 4.2 apresenta a frequência dos valores de similaridade obtidos em cada par de precedentes. O valor 0,64 foi escolhido como ponto de corte inicial devido ao intervalo considerável posterior a este ponto (0,64 a 0,75).

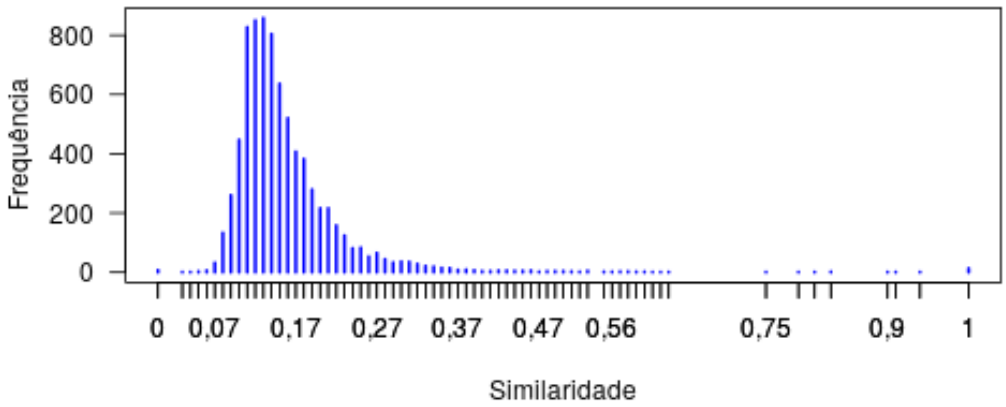

Figura 4.2: Histograma dos valores de similaridade para cada par de precedentes.

Ao considerar o ponto de corte inicial, são agrupados 27 precedentes judiciais. Ocorre que um precedente "A" pode ser semelhante ao precedente "B", enquanto que "B" pode ser semelhante ao precedente "C", logo, mesmo o precedente "A" não sendo diretamente semelhante ao precedente "C", considerando o valor de similaridade calculado, eles pertencem ao mesmo agrupamento e são considerados similares neste estudo. Por isso são utilizados grafos na definição dos agrupamentos, conforme observado na Figura 4.3.

Os resultados iniciais são encaminhados ao especialista, sendo criadas palavras-chave para cada precedente agrupado. Além disso, também são rotulados os precedentes identificados por intermédio de expressões regulares que são semelhantes aos previamente agrupados mas não constaram no agrupamento inicial.

Após essa rotulagem, todos os precedentes identificados no agrupamento inicial foram agrupados corretamente, ou seja, precisão de 100%. Entretanto, há precedentes relevantes que devem constar nesses grupos e não apareceram, por isso a métrica *recall*, que representa o percentual de documentos agrupados corretamente em relação ao total de documentos relevantes que deveriam ser agrupados, foi de 73%. A métrica *F-measure* combina precisão e *recall* em uma única métrica por meio do cálculo da média harmônica delas, ponderando uniformemente os dois índices.

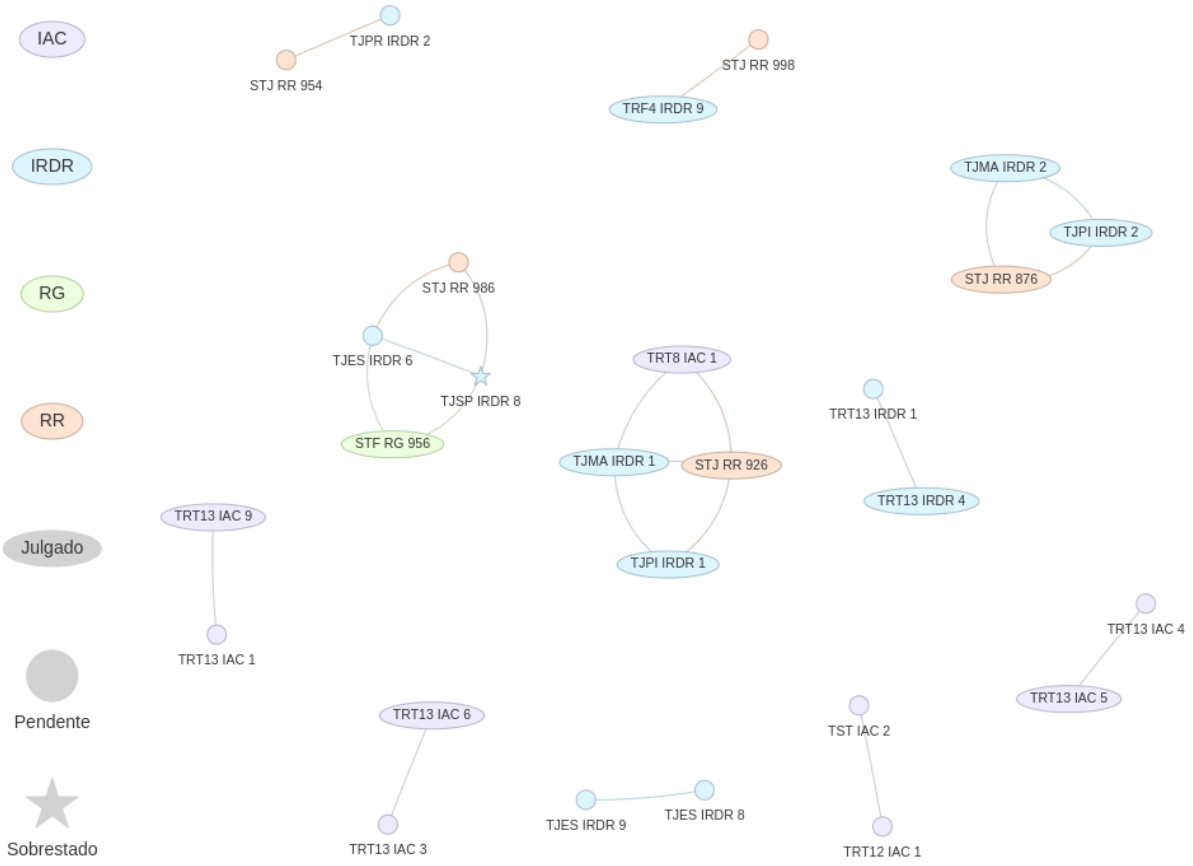

Figura 4.3: Grafo de precedentes agrupados com ponto de corte inicial.

A Figura 4.4 apresenta as métricas *F-measure*, precisão, *recall* e percentual de precedentes não rotulados para cada ponte de corte da similaridade cosseno avaliado. Verifica-se no ponto de corte 0,39 que quase todos os documentos rotulados foram agrupados corretamente, gerando *recall* de 95%, entretanto, 73% dos precedentes agrupados não foram rotulados.

Ao escolher o novo ponto de corte (de 0,39), 132 precedentes apresentam, em princípio, similaridade em relação a outros. Os resultados são novamente encaminhados ao especialista e rotulados. O total de agrupamentos corretos passa de 11 para 32, de um total de 50 grupos identificados.

Como são verificados mais alguns grupos semelhantes, optou-se por rotular os 287 precedentes agrupados considerando o ponto de corte 0,31. Esse quantitativo representa 12,7% do total de documentos da base de dados analisada, contendo 51,2% do total de IAC, 38,5% do total de IRDR, 8,5% do total de RG e 5,1% do total de RR.

Apesar de terem sido rotulados mais do que o dobro de precedentes apresentados no ponto de corte 0,39, foram identificados somente 7 novos agrupamentos de um total de 79

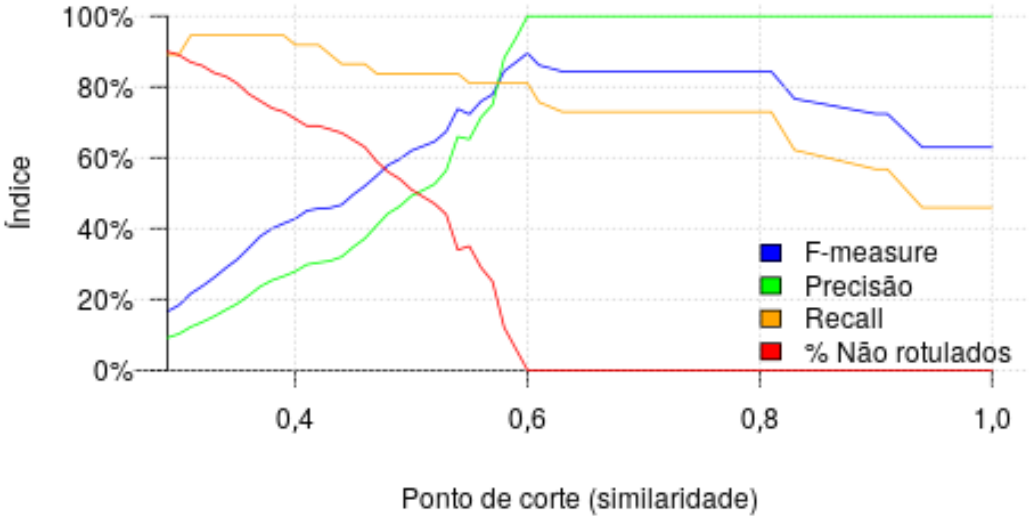

Figura 4.4: Métricas de avaliação por ponto de corte de similaridade após rotulagem inicial.

Tabela 4.1: Métricas de performance associadas aos principais pontos de corte: BM25.

Ponto de corte	**Precedentes**	***F-measure***	**Precisão**	***recall***
0,57	40	45,5%	100%	29,4%
0,438	105	64,7%	74,3%	57,4%

grupos identificados. Ou seja, esse modelo não auxilia em novas classificações a partir de 287 documentos agrupados, pois muitos precedentes foram agrupados em conjunto apesar de serem diferentes.

Os novos resultados obtidos em cada ponto de corte podem ser visualizados na Figura 4.5. O ponto de corte 0,57 é considerado o mais conservador, pois apresenta precisão de 100% e agrupa 40 precedentes semelhantes, conforme observado na Tabela 4.1. Já o ponto de corte 0,438 agrupa 105 documentos e apresenta o maior valor da métrica *F-measure* (64,7%), com precisão de 74,3% e *recall* de 57,4% em relação ao número de precedentes agrupados.

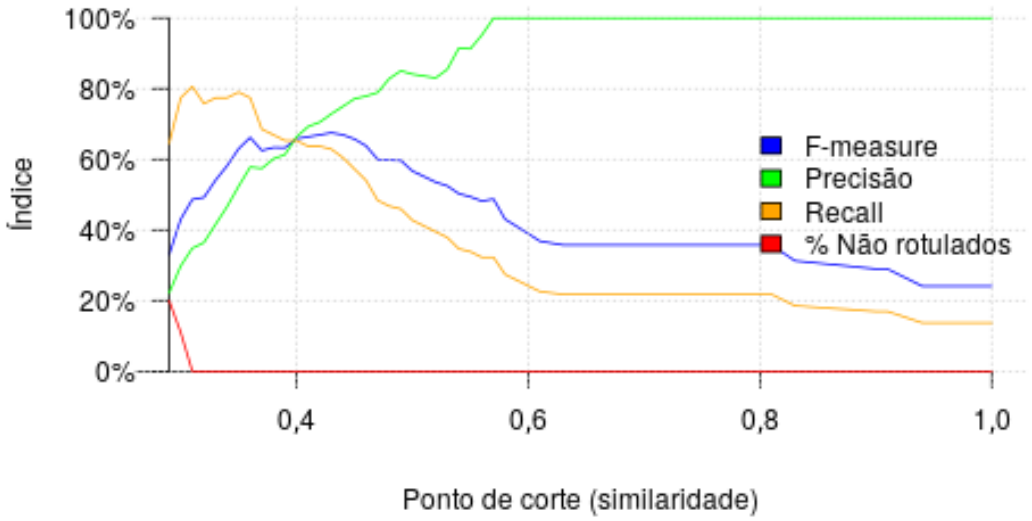

Figura 4.5: Métricas de avaliação por ponto de corte de similaridade: Modelo BM25.

4.2 Abordagens semânticas

As técnicas de LSI e LDA são aplicadas neste Capítulo para verificar se há melhora no resultado ao aplicar a modelagem por tópicos, uma vez que a redução dos termos em dimensões tende a agrupar palavras com significados semelhantes, como sinônimos e polissemia.

4.2.1 Indexação semântica com LSI

A técnica LSI é aplicada tanto na matriz de atributos por precedentes, contendo a contagem de termos, como na matriz ponderada pela função BM25. Essa técnica utiliza da decomposição de valores singulares, do inglês *Singular-Value Decomposition (SVD)*, para gerar uma matriz denominada *low-rank*, que mapeia cada atributo/documento em um espaço k-dimensional, onde k representa os maiores autovalores da matriz inicial.

Com o intuito de identificar o valor ótimo de dimensões (k) de cada modelo, a função de similaridade cosseno é utilizada nas matrizes *low-rank* geradas para k variando a cada 50 dimensões até 1.000. Dessa forma, calcula-se a métrica *F-measure* máxima obtida em cada modelo, conforme observado na Figura 4.6. Nota-se que os resultados obtidos pelo modelo ponderado pela função BM25 é superior em todas as dimensões.

O modelo LSI com BM25 de 200 dimensões foi o escolhido por apresentar a maior métrica *F-measure* (67,7%) e agrupar 100 precedentes. Já o modelo LSI convencional apresentou os melhores resultados com 500 dimensões (*F-measure* de 54,2%), conforme a Tabela 4.2.

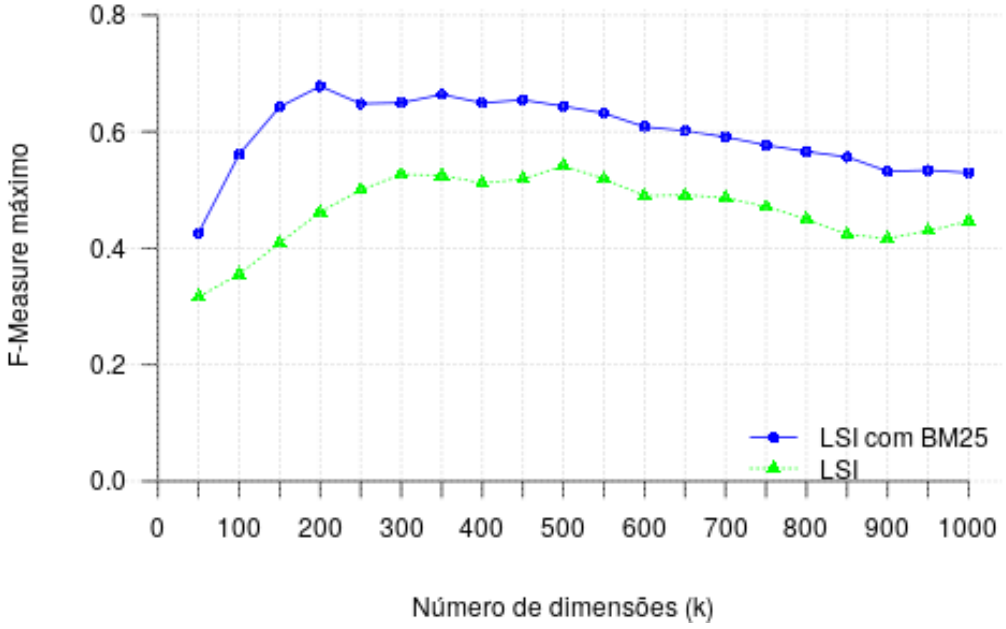

Figura 4.6: *F-measure* máxima por dimensão k dos modelos LSI.

Tabela 4.2: Métricas de performance associada aos principais pontos de corte: LSI.

Modelo	**Ponto de corte**	**Precedentes**	***F-measure***	**Precisão**	***recall***
LSI	0,52	104	54,2%	62,5%	47,8%
LSI com BM25	0,72	100	67,8%	80%	58,9%

Os resultados obtidos em cada ponto de corte para o modelo LSI com BM25 podem ser visualizados na Figura 4.7. Assim como verificado no modelo que utilizou somente o BM25, é possível identificar um ponto de corte mais conservador, 0,9, com precisão de 100% e 33 precedentes agrupados. A partir do ponto de corte 0,7, começam a aparecer precedentes que não foram rotulados.

O ponto de corte 0,628 agrupa 198 precedentes, com 25 não rotulados. Estes precedentes agrupados e ainda não rotulados são encaminhados ao especialista e verifica-se que nenhum foi agrupado corretamente.

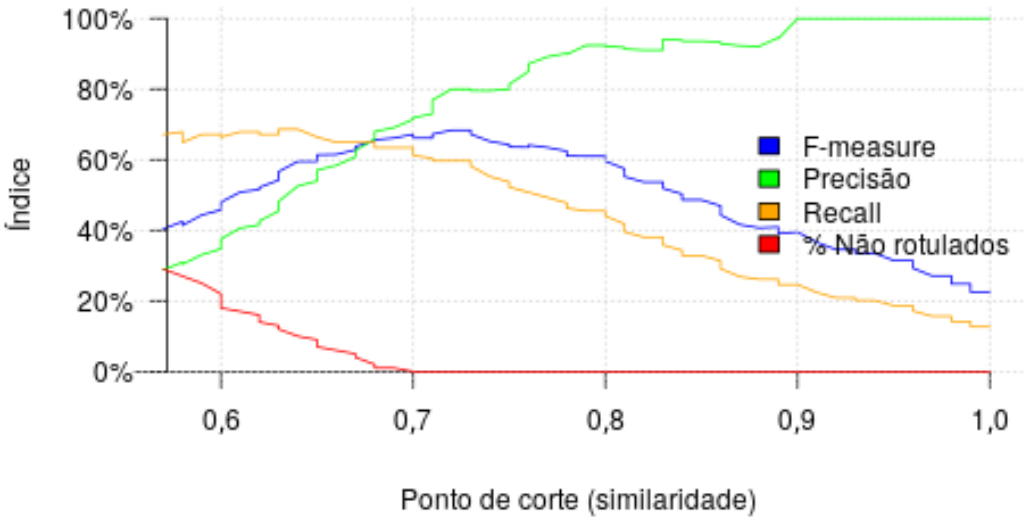

Figura 4.7: Métricas de avaliação por ponto de corte: Modelo LSI com BM25.

4.2.2 Modelagem de tópicos com LDA

O modelo *Latent Dirichlet Allocation (LDA)* é um modelo Bayesiano misto para variáveis discretas, por isso não foi possível utilizar o BM25 como ferramenta de ponderação dos elementos da matriz.

Assim como no modelo LSI, é necessário escolher o número de dimensões/tópicos resultantes do modelo. A função de similaridade cosseno é utilizada nas matrizes de precedentes por fatores, com k variando a cada 50 fatores até 500. Em seguida, é calculada a métrica *F-measure* máxima obtida em cada modelo.

Os resultados são gerados somente para o modelo que utiliza o algoritmo de *Gibbs sampling*, pois ele utiliza menos memória do que o algoritmo *Variational Expectation-Maximization (VEM)*. Além disso, os valores padrões de 50/fatores para α e de 0,1 para β foram utilizados como parâmetros da distribuição *a priori*.

Observa-se da Figura 4.8 que o modelo com 150 tópicos apresentou a maior métrica *F-measure*, entretanto, o valor máximo obtido é muito inferior aos observados nos modelos LSI e BM25. Essas métricas seriam provavelmente incrementadas caso fossem realizados ajustes nos parâmetros dos modelos LDA. Tais ajustes não foram feitos devido ao elevado custo de processamento e de tempo gasto na geração dos resultados. Ademais, estudos recentes mostram que o modelo LDA não é consistente quando aplicado a textos curtos[39].

Há de se destacar que foi necessária aproximadamente uma hora para gerar os resultados dos modelos com tópicos variando a cada 50 até o total de 500. Já o modelo LSI levou 20 minutos para os modelos com dimensões variando a cada 50 até o total de 1.000. Parte da diferença pode estar no fato do pacote R *Quanteda*[36], utilizado no LSI, ser con-

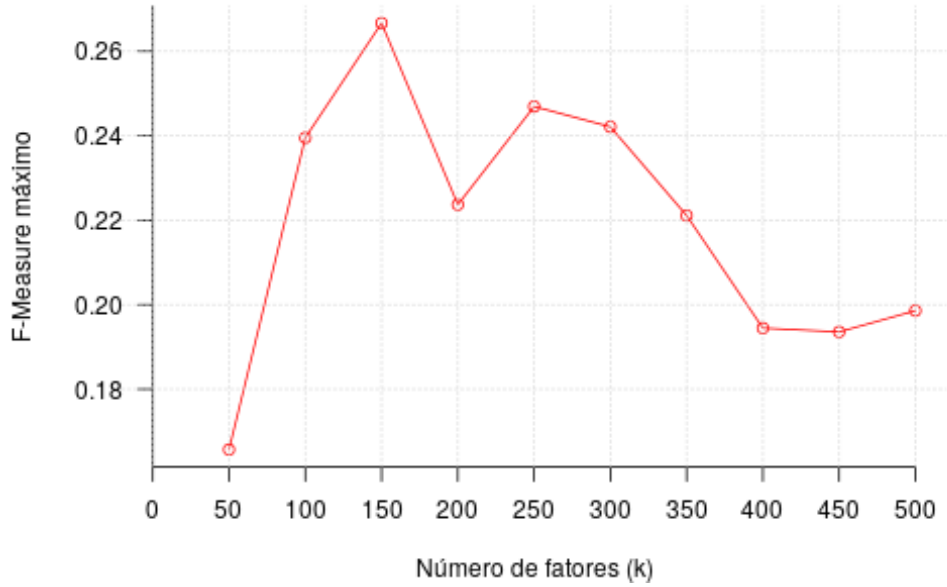

Figura 4.8: *F-measure* média por tópico k do modelo LDA.

siderado mais rápido e eficiente que os demais pacotes, pois utiliza *C++* e processamento *multithreading*.

Os resultados obtidos em cada ponto de corte para o modelo LDA podem ser visualizados na Figura 4.9. Verifica-se que o modelo apresenta precisão de 100% no ponto de corte 0,98, com 10 precedentes agrupados. Em seguida, há grande queda de precisão e aumento de dados agrupados e não rotulados.

O algoritmo LDA pressupõe que cada documento é uma mistura de um pequeno número de tópicos e que a presença de cada palavra é atribuível a um dos tópicos do documento. Esse modelo parece não se ajustar aos dados dos precedentes judiciais, pois, a maior parte dos termos utilizados não é atribuível a um único tópico. Além disso, a grande maioria dos precedentes judiciais são diferentes, tendo o especialista identificado 136 precedentes semelhantes de um total de 2.260, ou seja, somente 6% do total.

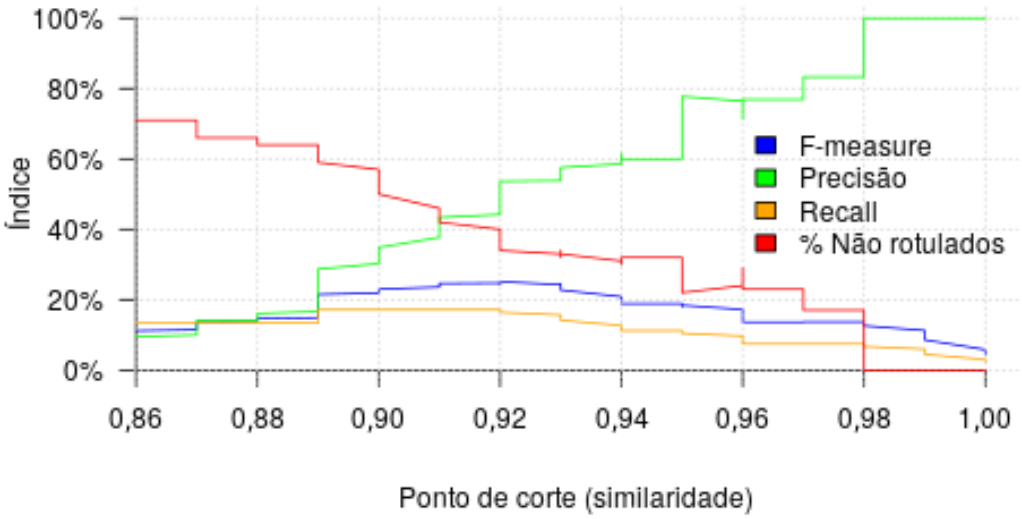

Figura 4.9: Métricas de avaliação por ponto de corte: Modelo LDA.

4.3 Indução de modelo com o uso de entidades nomeadas

A métrica *F-measure* é utilizada como forma de comparar os modelos gerados em relação ao número de precedentes agrupados. Dessa forma, não é necessário escolher um ponto de corte de similaridade de forma arbitrária, pois, dado que o especialista identificou 136 precedentes judiciais semelhantes, basta selecionar o modelo de maior *F-measure* que agrupa quantitativo de precedentes próximo a esse ponto.

Verifica-se da Figura 4.10 que os modelos LSI apresentam os maiores valores *F-measure* ao agrupar entre 100 e 120 precedentes, isso ocorre devido ao fato dos modelos agruparem de 0 a 100 precedentes com alta precisão e deixando de agrupar outros que deveriam ter sido agrupados ou agruparem mais de 120 precedentes com baixa precisão e agrupando quase todos os precedentes identificados pelo especialista que deveriam ter sido agrupados. O modelo LSI com BM25 permanece com indicador acima dos demais modelos até o agrupamento de 140 precedentes. Entretanto, o valor da métrica *F-measure* altera conforme o número de dimensões escolhido.

O modelo BM25 já é bem mais simples e apresenta valores da métrica *F-measure* bem próximos aos obtidos pelo LSI, inclusive com métricas superiores quando mais de 140 precedentes são agrupados. Por isso, esse modelo é o escolhido para verificar se há melhora nos resultados ao considerar entidades nomeadas.

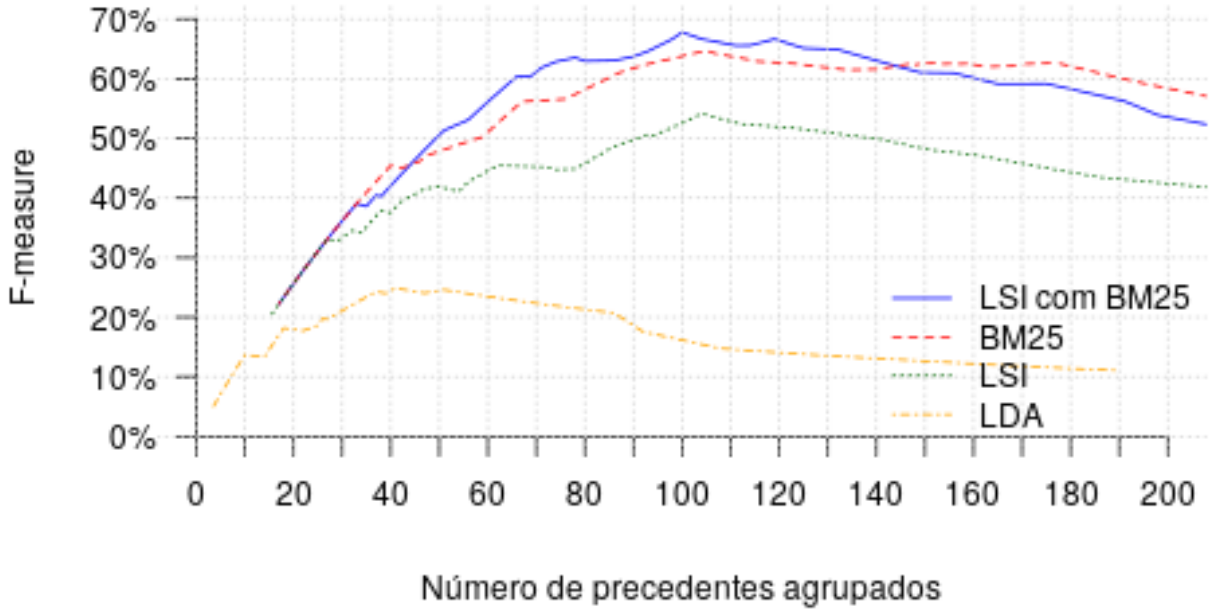

Figura 4.10: Valor da métrica *F-measure* utilizando BM25, LSI, LSI com BM25 e LDA.

Com o intuito de verificar se há ganhos ao usar entidades nomeadas, o modelo BM25 é induzido com:

- artigos e normas extraídos das descrições dos precedentes judiciais;
- tesauro do STF;
- assuntos cadastrados nos precedentes;
- entidades nomeadas identificadas como nomes próprios via *POS-tagging* ou por apresentar palavras em sequência contendo a primeira letra maiúscula.

A lematização também é testada nos termos de forma a ignorar tempo verbal, gênero, plural e demais flexões nas palavras. A Figura 4.11 traz os valores da métrica *F-measure* obtidos pelo modelo BM25 induzido conforme os itens descritos anteriormente. Observa-se que somente os modelos induzidos com o tesauro do STF e com as normas apresentaram resultados melhores do que o do BM25.

Ao todo, foram identificadas nas descrições dos precedentes judiciais 514 normas (137 termos), 1.566 nomes próprios (214 termos), 2.854 palavras compostas no tesauro do STF (274 termos) e 5.596 assuntos (663 termos). O modelo induzido com os assuntos dos processos (metadados na base BNPR) pode não ter obtido bons resultados devido ao fato dos tribunais cadastrarem nos processos os assuntos com níveis mais abrangentes das Tabelas Processuais Unificadas do CNJ. Observa-se na Figura 4.12 que os quatro assuntos mais frequentes são de níveis abrangentes, englobando direito processual cível e do trabalho, direito tributário, direito administrativo e direito civil.

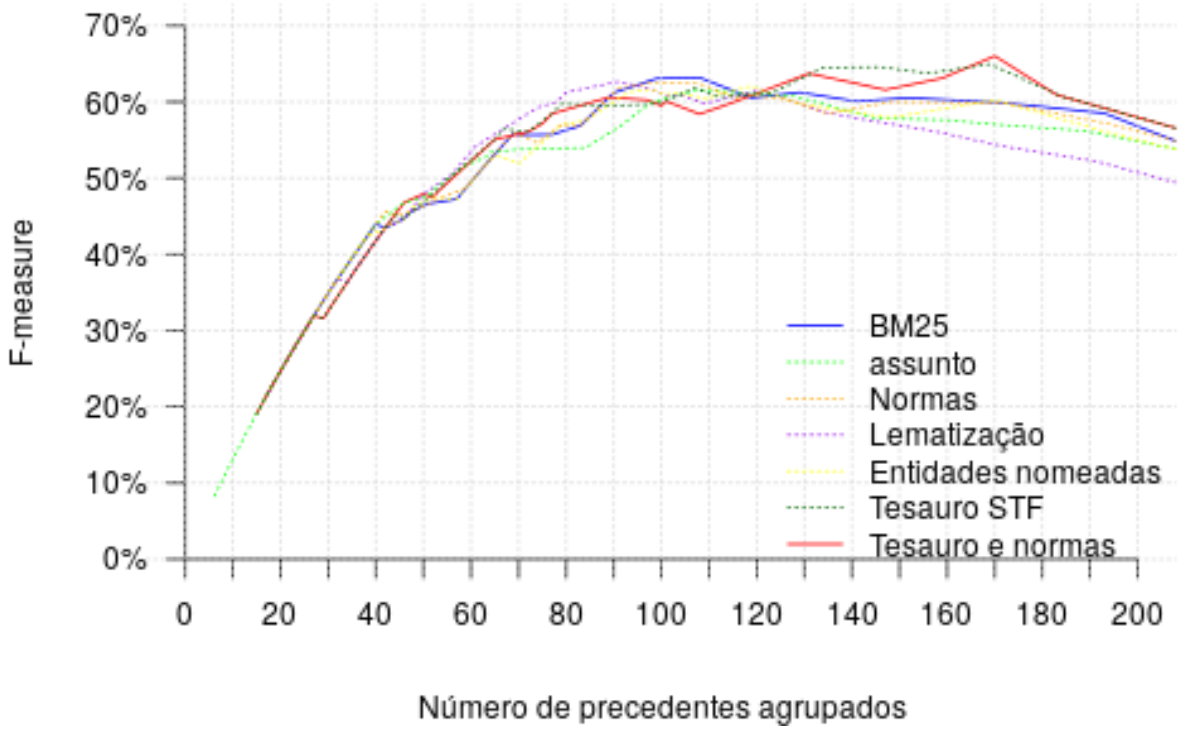

Figura 4.11: Métrica *F-measure* utilizando BM25 induzido.

Foram identificadas 409 entidades nomeadas que constam no tesauro do STF ou referem-se às normas extraídas dos precedentes. Esses termos aparecem 3.369 vezes nos textos dos precedentes. Ao induzir o modelo BM25 com essas entidades, verifica-se um pequeno ganho, pois o modelo BM25 apresenta métrica *F-measure* máxima de 64,7% ao agrupar 105 precedentes, enquanto que o modelo induzido com entidades nomeadas apresenta *F-measure* de 66%, entretanto, agrupando 65 precedentes a mais.

Como forma de aperfeiçoar ainda mais o modelo de maneira que a rotulação realizada pelo especialista sobressaia sobre os demais termos, mas não os desconsiderando, optou-se por adicionar em duplicidade as palavras-chave dadas pelo especialista à descrição do precedente judicial, gerando novamente o modelo.

O resultado obtido apresenta resultados satisfatórios, agrupando 118 precedentes judiciais com métrica *F-measure* de 87,7% e precisão de 96,6%. Além disso, o modelo não apresentou métricas ainda melhores por ter agrupado temas parecidos, como por exemplo os precedentes que versam sobre honorários advocatícios para advogados dativos, defensores dativos ou da fazenda pública.

A Figura 4.13 mostra os valores da métrica *F-measure* obtidos em relação ao número de precedentes agrupados para os modelos BM25, induzido com entidades nomeadas e com as palavras-chave inseridas pelo especialista. Observa-se que os dois últimos modelos

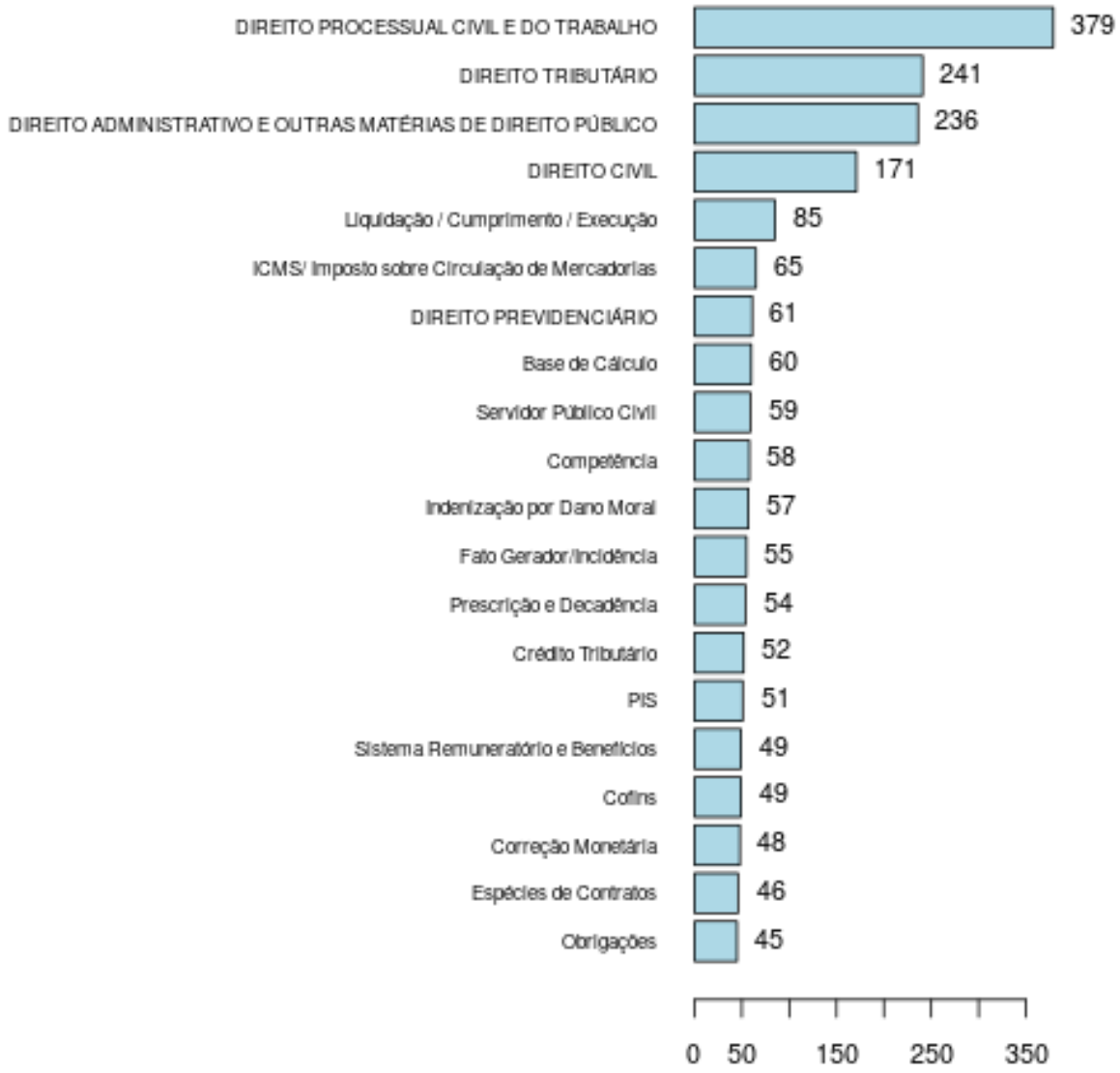

Figura 4.12: Assuntos mais frequentes dos precedentes judiciais.

apresentaram os maiores indicadores ao agrupar cerca de 125 precedentes.

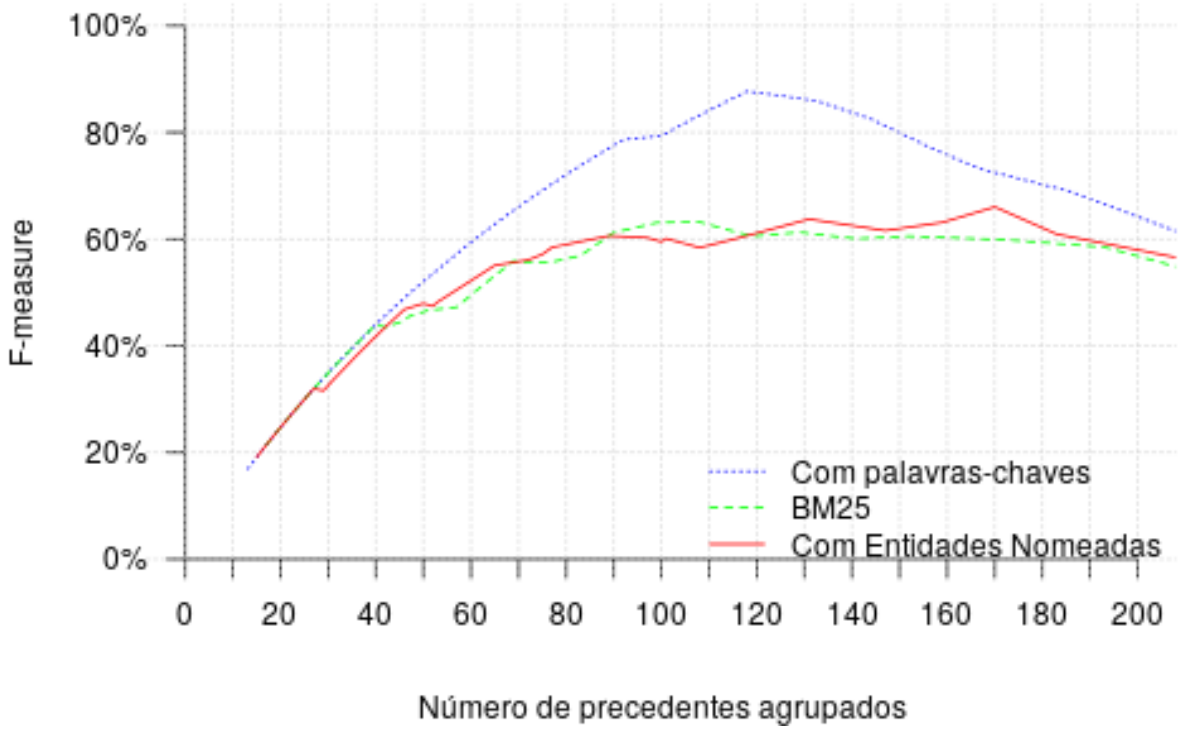

Figura 4.13: *F-measure* dos modelos com BM25.

4.4 Modelo final

A ferramenta resultante dessa pesquisa pode ser implementada no Conselho Nacional de Justiça (CNJ) para agrupar precedentes judiciais e o grafo dinâmico gerado pode ser disponibilizado ao público em geral em seu endereço eletrônico. A Figura 4.14 mostra os agrupamentos obtidos pelo modelo, destacando-se a precisão obtida de quase 97%.

Cada cor no grafo representa um tipo de precedente judicial: vermelho Repercussão Geral (RG), laranjado Recursos Repetitivos (RR), azul Incidente de Resolução de Demandas Repetitivas (IRDR) e verde Incidente de Assunção de Competência (IAC). Cada símbolo representa uma situação do precedente: elipse indica precedente já julgado e com uma tese firmada, estrela indica precedente sobrestado, e círculo indica precedente admitido. É possível identificar no grafo os precedentes semelhantes com decisões divergentes, pois as descrições dos precedentes e decisões proferidas podem ser visualizadas ao passar o *mouse* sobre o precedente.

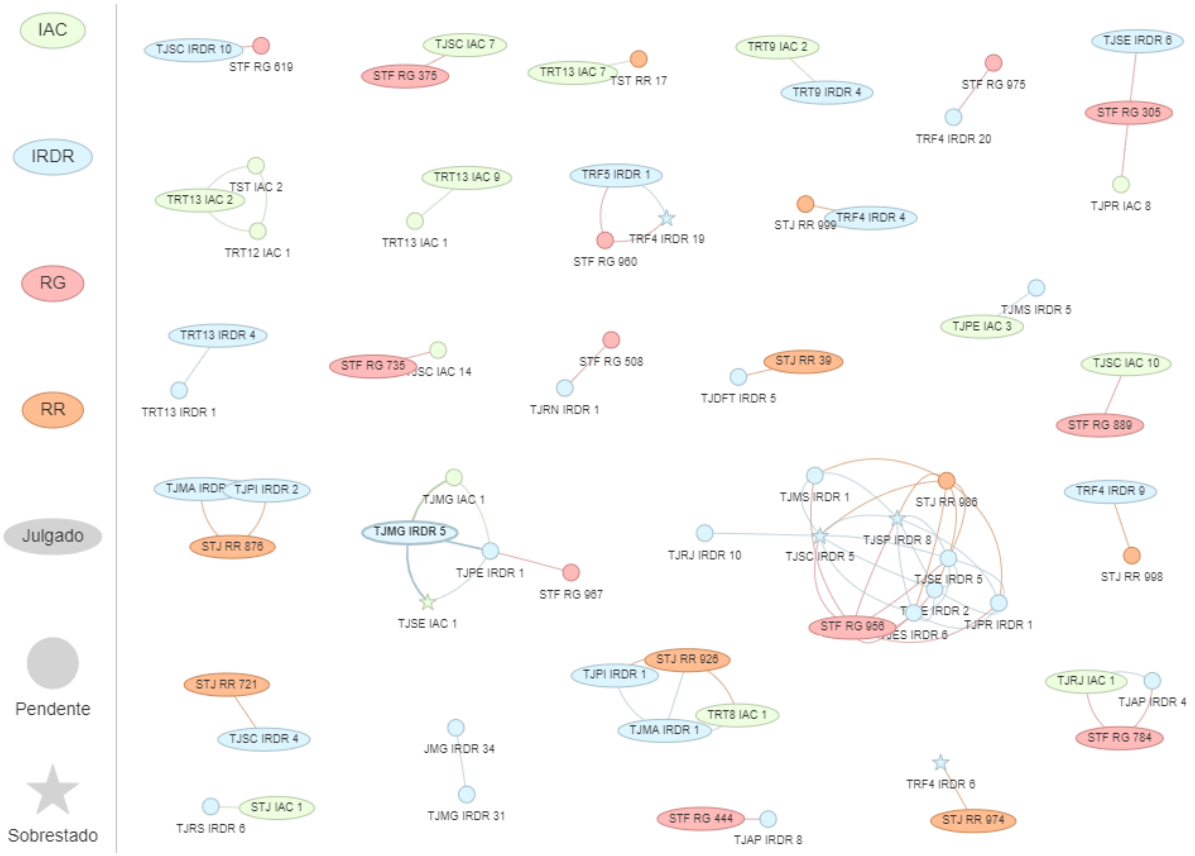

Figura 4.14: Grafo de precedentes judiciais agrupados: modelo final.

O Código de Processo Civil (CPC) traz instrumentos para não admitir ou suspender os precedentes que versem sobre questão com recurso afetado pelos tribunais superiores, conforme observado nos seguintes trechos:

> *"Art. 976, §4º: É incabível o incidente de resolução de demandas repetitivas quando um dos tribunais superiores, no âmbito de sua respectiva competência, já tiver afetado recurso para definição de tese sobre questão de direito material ou processual repetitiva."*
>
> *"Art. 982, §3º: Visando à garantia da segurança jurídica, qualquer legitimado mencionado no art. 977, incisos II e III , poderá requerer, ao tribunal competente para conhecer do recurso extraordinário ou especial, a suspensão de todos os processos individuais ou coletivos em curso no território nacional que versem sobre a questão objeto do incidente já instaurado."*

No entanto, poucos são os precedentes suspensos, apesar da similaridade com temas de Recursos Repetitivos e Repercussões Gerais. Esses institutos estão presentes em 3 de cada quatro agrupamentos identificados. Dos 399 IRDR e IAC na base de dados, 56 (14%) contêm precedentes semelhantes aos recebidos pelo STF e STJ.

Destaca-se o precedente sobre a legalidade da inclusão da Tarifa de Uso do Sistema de Transmissão (TUST) e da Tarifa de Uso do Sistema de Distribuição (TUSD) de energia elétrica na base de cálculo do Imposto sobre Circulação de Mercadorias e Serviços (ICMS) por apresentar IRDR instaurado em oito Tribunais de Justiça e ainda com recursos para o STF e para o STJ, conforme observado na Figura 4.15.

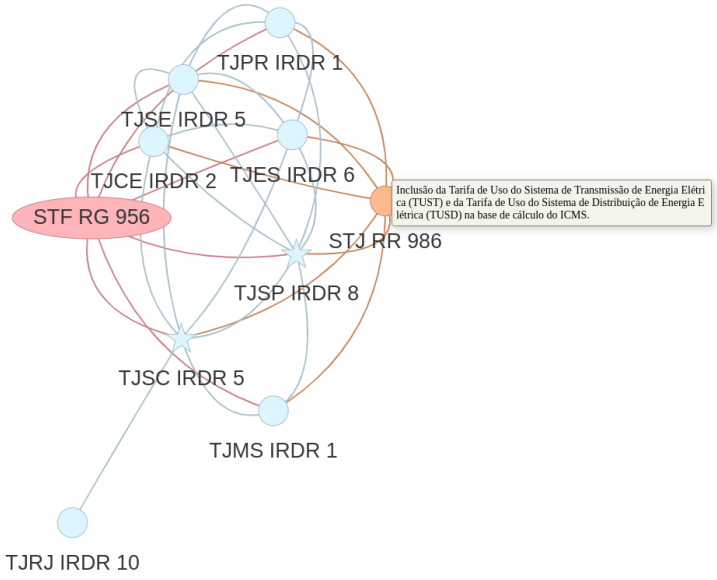

Figura 4.15: Grafo de precedentes agrupados sobre ICMS na tarifa de energia elétrica.

O STF recebeu em 20/04/2017 recurso oriundo do Tribunal de Justiça do Estado de São Paulo com esse tema e a questão foi julgada em 07/09/2017 da seguinte maneira: *"O Tribunal, por maioria, reconheceu a inexistência de repercussão geral da questão, por não se tratar de matéria constitucional, vencido o Ministro Marco Aurélio."*

Já o STJ recebeu em 15/12/2017 recurso oriundo dos Tribunais de Justiça do Rio Grande do Sul, Tocantins e Mato Grosso referente a esse tema. A questão foi afetada no STJ e ainda não foi julgada, entretanto, o tribunal determinou a suspensão nacional de todos os processos pendentes, individuais ou coletivos.

Observa-se que os três tribunais de origem com recurso no STJ não apresentam IRDR, já os demais, com exceção de São Paulo e Santa Catarina, podem estar em desconformidade com o CPC.

4.5 Aprendizagem incremental do modelo

A ferramenta proposta guarda o histórico dos resultados anteriores e compara com os obtidos pela nova base de dados. Os dados dos precedentes judiciais ingressados no ano de 2019 são utilizados para avaliação do modelo incrementado.

A nova carga contém 58 novos precedentes. Foram identificadas 29 novas entidades nomeadas, das quais 22 são corretas. As novas entidades nomeadas ficam armazenadas em um repositório até deliberação do especialista.

O modelo BM25 agrupou 7 novos precedentes judiciais, dentre os quais 5 (71,4%) foram agrupados corretamente segundo o especialista. Já o modelo induzido com o uso de entidades nomeadas agrupou 8 novos precedentes, sendo 4 (50%) agrupados corretamente. Destaca-se que todos os precedentes agrupados corretamente no segundo modelo também estiveram contemplados no primeiro modelo.

Apesar do modelo BM25 ter sido mais eficiente ao identificar os novos precedentes, o modelo induzido com entidades nomeadas ainda apresenta melhores resultados de forma geral, conforme observado na Figura 4.16. Esse modelo identificou 127 precedentes, com métrica *F-measure* de 65,5% e precisão de 71,7%, contra 110 precedentes identificados pelo BM25 (*F-measure* de 62,1% e precisão de 73,7%).

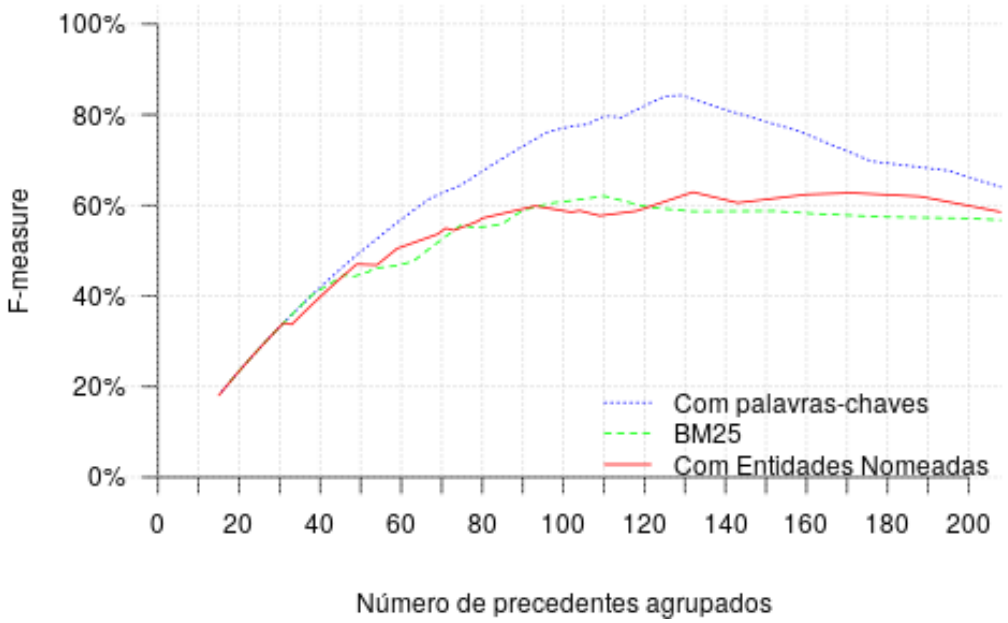

Figura 4.16: *F-measure* dos modelos incrementais com BM25.

O modelo final atualizado, que utiliza as palavras-chave identificadas pelo especialista, agrupa 9 precedentes a mais do que o modelo contendo somente os dados até o ano de 2018,

totalizando 129 precedentes agrupados. Destes, 91,5% estão agrupados corretamente. O grafo contendo os agrupamentos atualizados pode ser visualizado na Figura 4.17.

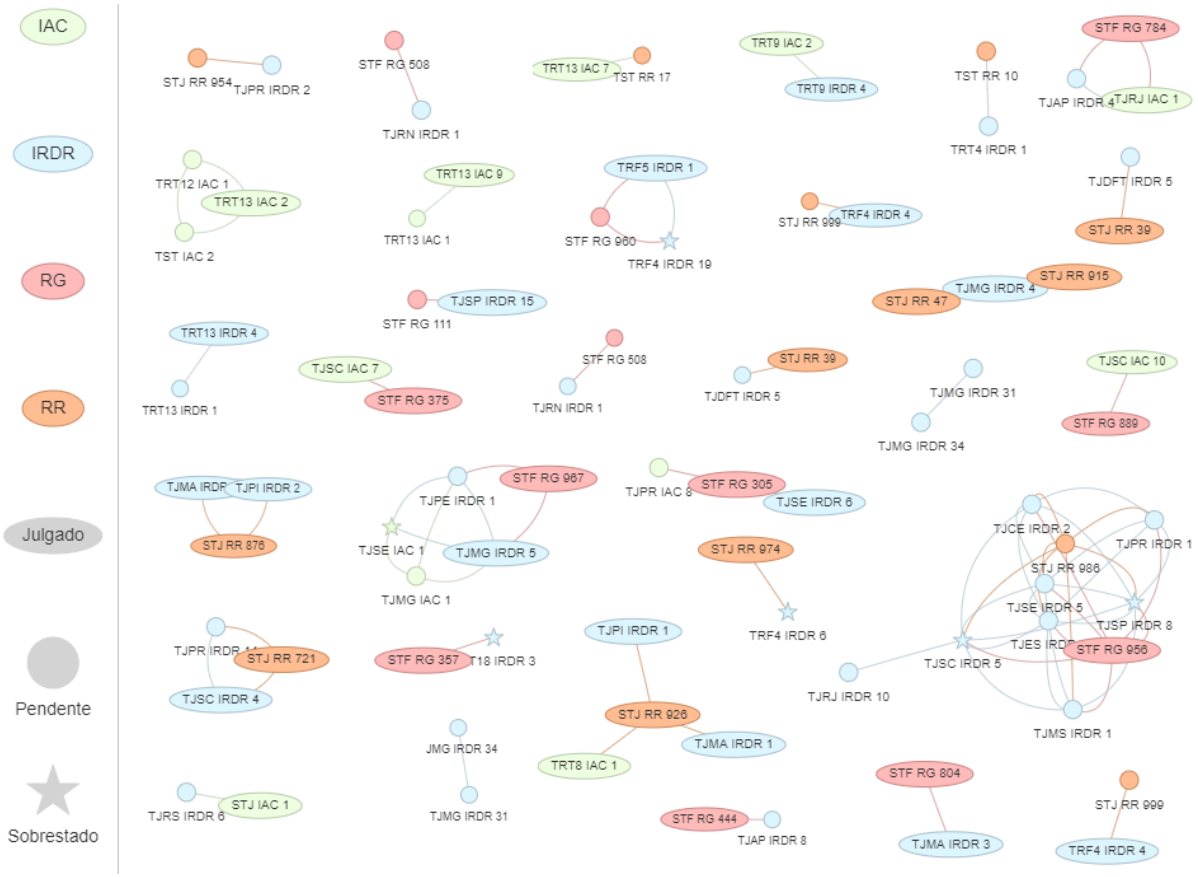

Figura 4.17: Grafo de precedentes judiciais agrupados: modelo incremental.

Capítulo 5

Conclusões

A ferramenta desenvolvida nessa pesquisa para agrupar precedentes judiciais apresenta resultados satisfatórios, uma vez que agrupa corretamente mais de 90% dos precedentes judiciais que são similares a temas de Incidente de Resolução de Demandas Repetitivas e Incidente de Assunção de Competência, e identifica novos temas com precisão superior a 70%. Essa ferramenta gera um grafo dinâmico que pode ser disponibilizado ao público em geral por meio do endereço eletrônico do CNJ e pode ser uma contribuição importante para minimizar a insegurança jurídica na 2ª instância. Essa ferramenta pode auxiliar aos tribunais identificarem precedentes já julgados para que em seu julgamento, a mesma tese já decidida seja aplicada, ou, em caso de julgamento em divergência, que o próprio Tribunal possa indicar aos tribunais superiores os recursos que necessitariam de uniformização da jurisprudência, para evitar a insegurança jurídica.

A hipótese da pesquisa de que os precedentes judiciais semelhantes podem ser identificados com a utilização de técnicas baseadas em análise semântica de texto não é rejeitada, uma vez que os resultados obtidos com a utilização do modelo LSI em conjunto com o BM25 foram ligeiramente superiores aos do modelo que utiliza somente abordagem sintática BM25. Entretanto, os resultados obtidos pelo modelo LSI variam conforme a escolha do número de dimensões, sendo necessário, na medida em que há incrementos na base de dados, verificação quanto ao número ideal de dimensões. Como essa verificação demanda um custo computacional considerável, é preferível a utilização do modelo BM25 por ser mais simples e com valores da métrica *F-measure* bastante próximos aos do modelo LSI.

Com relação à utilização de entidades nomeadas, a hipótese de que há melhora nos resultados com sua utilização é rejeitada, pois, apesar dos resultados obtidos serem bastante próximos aos do modelo BM25, o modelo com entidades nomeadas identifica corretamente somente 50% dos precedentes semelhantes ingressados na base de dados no ano de 2019, enquanto que o modelo BM25 apresenta precisão de 75%.

O presente trabalho atinge seu objetivo principal ao criar uma ferramenta para identificação automática dos Precedentes Judiciais semelhantes e seus objetivos específicos ao avaliar o uso de técnicas baseadas em análise semântica para estimar similaridade entre precedentes; ao analisar os valores de similaridade por cosseno para identificação dos precedentes semelhantes; validar os grupos de precedentes indicados como similares com apoio do especialista; ao medir a performance dos modelos comparando os agrupamentos propostos com os considerados válidos pelo especialista; ao verificar ganhos de performance ao considerar entidades nomeadas; e ao desenvolver uma prova de conceito para utilização do algoritmo pelo CNJ. O grafo dinâmico gerado ainda possibilita a identificação de Precedentes semelhantes com decisões divergentes e mostra os precedentes semelhantes aos Recursos Repetitivos (STJ) e às Repercussões Gerais (STF). Dessa forma, a ferramenta pode ser utilizada pelo CNJ.

Por fim, o modelo final apresenta resultados interessantes com relação à aplicação do Código de Processo Civil (CPC), pois ele traz instrumentos para não admitir ou suspender precedentes que versem sobre questão com recurso afetado pelos tribunais superiores, no entanto, poucos são os precedentes suspensos, apesar da similaridade com temas de Recursos Repetitivos e Repercussões Gerais. Esses dois institutos estão presentes em 3 a cada 4 agrupamentos identificados. Dos 399 IRDR e IAC na base de dados, 56 (14%) contêm temas semelhantes aos recebidos pelo STF e STJ. Destaca-se, inclusive, a existência de um tema com IRDR em 8 Tribunais de Justiça, inclusive com recursos no STF no STJ.

Como limitações do trabalho tem-se: a avaliação realizada por um único especialista; o pacote utilizado para gerar os resultados do modelo LDA pode divergir devido à implementação do algoritmo; ajustes nos parâmetros do modelo LDA e BM25 podem gerar resultados diversos; a identificação das entidades nomeadas pode ser realizada por meio de outros pacotes; e não há um tesauro único do Poder Judiciário, pois o tesauro do STF não contempla algumas peculiaridades dos diversos ramos de justiça.

Como trabalhos futuros, sugere-se o aperfeiçoamento na identificação automática de entidades nomeadas e de palavras-chave com base nas descrições dos precedentes judiciais. Dessa forma, a ferramenta se torna cada vez menos dependente de um especialista. Além disso, é importante que a ferramenta seja adaptada para que leia os textos das petições iniciais dos processos ingressados na 1º instância e sugira o sobrestamento ou não do processo devido à similaridade a precedentes existentes nos tribunais de instâncias superiores.

Referências

[1] *Justiça em Números 2018*. Relatório Técnico, Conselho Nacional de Justiça, 2018. http://www.cnj.jus.br/programas-e-acoes/pj-justica-em-numeros/relatorios, acesso em 01/3/2019. 1

[2] *Estudo Comparado Sobre Recursos, Litigiosidade e Produtividade: a prestação jurisdicional no contexto internacional*. Relatório Técnico, Conselho Nacional de Justiça, 2011. http://www.cnj.jus.br/images/pesquisas-judiciarias/relat_estudo_comp_inter.pdf, acesso em 02/6/2018. 1

[3] *100 Maiores Litigantes*. Relatório Técnico, Conselho Nacional de Justiça, 2012. http://www.cnj.jus.br/images/pesquisas-judiciarias/Publicacoes/100_maiores_litigantes.pdf, acesso em 02/6/2018. 1

[4] Associação Brasileira de Jurimetria: *Os Maiores Litigantes nas Ações Consumeristas na Justiça Estadual: Mapeamento e Proposições*. Relatório Técnico, Conselho Nacional de Justiça, 2017. http://www.cnj.jus.br/pesquisas-judiciarias/justica-pesquisa/publicacoes, acesso em 02/6/2018. 1

[5] Brasil: *Lei, nº 13.105, de 16 de março de 2015*. Código de Processo Civil, mar 2015. 2

[6] *Relatório do Banco Nacional de Dados de Demandas Repetitivas e Precedentes Obrigatórios*. Relatório Técnico, Conselho Nacional de Justiça, 2018. http://www.cnj.jus.br/files/conteudo/arquivo/2018/02/03a6c043d7b9946768ac79a7a94309af.pdf, acesso em 02/6/2018. 3

[7] Tella, María José Falcón y: *Lições de teoria geral do direito*. São Paulo: Revista dos Tribunais, 2011. 7

[8] Mendes, Joyce Barros: *Precedente judicial como fonte do direito*. 2015. 7, 8

[9] Nunes, Dierle e André Frederico HORTA: *Aplicação de precedentes e distinguishing no cpc/2015: Uma breve introdução*. Coleção Grandes Temas do Novo CPC. Salvador. Juspodvim, 2015. 8

[10] Nogueira, Cláudia Albagli: *O novo código de processo civil e o sistema de precedentes judiciais: pensando um paradigma discursivo da decisão judicial*. Revista Brasileira de Direito Processual–RBDPro, Belo Horizonte, ano, 22:185–210, 2015. 8

[11] Gomes, Marcos Paulo Pereira: *O sistema de precedentes judiciais vinculantes no novo cpc: uma busca por uniformização, segurança jurídica e celeridade processual.* da ADVOCEF, página 131. 11

[12] Antônio Pereira Gaio Júnior: *Incidente de Resolução de Demandas Repetitivas no projeto do novo CPC: Breves apontamentos.* Legis Augustus, 4(2):1–11, 2013, ISSN 2179-6637. 11

[13] Karol Araújo Durço: *O Incidente de Resolução de Demandas Repetitivas: Uma das propostas centrais do projeto de novo Código de Processo Civil.* Revista Eletrônica de Direito Processual, 8(8), 2016, ISSN 1982-7636. 11

[14] Jesus Silva, Jamyl de: *O Incidente de Resolução de Demandas Repetitivas no projeto do novo Código de Processo Civil: Segurança Jurídica e legitimidade democrática das decisões judiciais no Estado Constitucional de direito.* Dissertação de mestrado, Universidade de Brasíia, 2014. 11

[15] Artur César de Souza: *Resolução de demandas repetitivas: Comunicação de demanda individual incidente de resolução de demandas repetitivas recursos repetitivos.* Novo Processo Civil Brasileiro. Almedina Brasil, São Paulo, Brasil, 2015, ISBN 978-858-49-3091-3. 11

[16] Nunes, Marcelo G e Julio AZ Trecenti: *Reformas de decisão nas câmaras de direito criminal em são paulo.* Conjur, 15(11), 2015. 11

[17] Beppler, Márcio Duarte e Anita Maria da Rocha Fernandes: *Aplicação de text mining para a extração de conhecimento jurisprudencial.* Anais SULCOMP - Congresso Sul Brasileiro de Computação, 1, 2012, ISSN 2359-2656. 11

[18] Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro e Vasileios Lampos: *Predicting judicial decisions of the european court of human rights: A natural language processing perspective.* PeerJ Computer Science, 2:e93, 2016. 11

[19] Landthaler, Jörg, Bernhard Waltl, Patrick Holl e Florian Matthes: *Extending Full Text Search for Legal Document Collections Using Word Embeddings.* Em *JURIX*, páginas 73–82, 2016. 11

[20] Aggarwal, Charu C e ChengXiang Zhai: *Mining text data.* Springer Science & Business Media, 2012. 12

[21] Manning, Christopher D., Prabhakar. Raghavan e Hinrich.Salton1962 Shutze: *Introduction to information retrieval.* Cambridge University Press, 2008, ISBN 0521865719. `https://nlp.stanford.edu/IR-book/`. 12, 13

[22] Murugan, S e R Karthika: *A literature review on text mining techniques and methods.* International Journal of Computer Sciences and Engineering, 6(2):96–99, 2018. 13

[23] Osei-Bryson, Kweku Muata: *Towards supporting expert evaluation of clustering results using a data mining process model.* Information Sciences, 180(3):414–431, 2010. 13

[24] Shearer, Colin: *The crisp-dm model: the new blueprint for data mining*. Journal of data warehousing, 5(4):13–22, 2000. 14, 19

[25] Hotho, Andreas, Andreas Nürnberger e Gerhard Paaß: *A brief survey of text mining*. Em *Ldv Forum*, volume 20, páginas 19–62. Citeseer, 2005. 14

[26] Akemi Sinoara, Roberta, João Antunes e Solange Oliveira Rezende: *Text mining and semantics: a systematic mapping study Journal of the Brazilian Computer Society Text mining and semantics: a systematic mapping study*. Journal of the Brazilian Computer Society, 23(1):9, 2017. 14, 15

[27] Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez e Krys Kochut: *A brief survey of text mining: Classification, clustering and extraction techniques*. arXiv preprint arXiv:1707.02919, 2017. 15

[28] Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer e Richard Harshman: *Indexing by latent semantic analysis*. Journal of the American society for information science, 41(6):391, 1990. 15

[29] Hofmann, Thomas: *Unsupervised learning by probabilistic latent semantic analysis*. Machine learning, 42(1-2):177–196, 2001. 15

[30] Blei, David M, Andrew Y Ng e Michael I Jordan: *Latent dirichlet allocation*. Journal of machine Learning research, 3(Jan):993–1022, 2003. 15

[31] Mikolov, Tomas, Kai Chen, Greg Corrado e Jeffrey Dean: *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781, 2013. 15

[32] Devlin, Jacob, Ming Wei Chang, Kenton Lee, Kristina Toutanova Google e A I Language: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Relatório Técnico, ISBN 1810.04805V2. `https://github.com/tensorflow/tensor2tensor`. 15

[33] Mariano, Ari e Maíra Santos: *Revisão da Literatura: Apresentação de uma Abordagem Integradora*. ISBN 978-84-697-5592-1. 16, 19

[34] Salton, Gerard e Gerard: *Some experiments in the generation of word and document associations*. Em *Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall)*, páginas 234–250, New York, New York, USA, 1962. ACM Press. 16

[35] Freire, Jânio, Vládia Pinheiro e David Feitosa: *Flexsts-um framework para similaridade semantica textual*. Em *PROPOR-International Conference on the Computational Processing of Portuguese, Tomar, Portugal*, 2016. 17, 18

[36] Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller e Akitaka Matsuo: *quanteda: An R package for the quantitative analysis of textual data Software • Review • Repository • Archive*. `https://doi.org/10.21105/joss.00774`. 18, 23, 34

[37] Straka, Milan e Jana Straková: *Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe*. Relatório Técnico. `http://ufal.mff.cuni.cz/udpipe`. 18, 22

[38] Meyer, David, Kurt Hornik e Ingo Feinerer: *Text mining infrastructure in r*. Journal of statistical software, 25(5):1–54, 2008. 23

[39] Hajjem, Malek e Chiraz Latiri: *Combining IR and LDA Topic Modeling for Filtering Microblogs*. Procedia Computer Science, 112:761–770, jan 2017, ISSN 1877-0509. 34