Recuperação de Informação em processos do MPSC

Rodrigo G. de Miranda, Luana da Silva, Ingrid K. de Souza e Vinícius M. de Sousa

Samaia IT - Integrando Soluções
E-mail(s): rodrigo@samaiait.com.br, luana@samaiait.com.br, ingrid@samaiait.com.br, vinicius@samaiait.com.br

Resumo

De acordo com a Constituição Federal (art. 129), os Ministérios Públicos (MPs) têm como funções a promoção de ações penais públicas, zelar pelo respeito entre os Poderes Públicos, exercer controle externo da atividade policial, por meio de pareceres públicos. O sistema de busca de pareceres do Ministério Público de Santa Catarina (MPSC) possibilita que o cidadão acesse os pareceres através do número do processo, de modo que o público em geral, *i.e.*, pessoas que não sejam técnicas da área ou estejam envolvidas em algum processo, têm dificuldade em acessar informações de seu interesse. Como primeiro passo para otimizar a recuperação de informação, foi proposta a identificação de tópicos dentro de assuntos, usando *topic modeling.* Técnicas de *topic modeling* auxiliam na recuperação de informação quando não se sabe como procurar a informação que se procura dentro da coleção, ou quando não se sabe a informação explicitamente. O algoritmo utilizado para a modelagem foi o LDA, *Latent Dirichlet Allocation,* o qual é um modelo probabilístico generativo. Neste artigo, foi proposto um estudo de caso de classificação de pareceres com uso de tópicos. Primeiramente foi realizada a extração de conteúdos dos arquivos PDF dos processos. Com os conteúdos e metadados extraídos foi realizado o pré-processamento dos arquivos, que inclui: (i) agrupamento dos documentos; (ii) *tokenização* por palavras, (iii) filtragem palavras mais relevantes, (iv) conversão de todo texto para letras minúsculas e (v) RSLP *stemming* (Removedor de Sufixos da Língua Portuguesa). Com os documentos pré-processados, foi realizada a modelagem usando o algoritmo LDA para *topic modeling*. Com a aplicação da metodologia apresentada foram encontrados quinze tópicos (ex: compras *online*, educação, aumento de preços, danos morais e materiais, ajustamento de conduta) relacionados ao assunto de práticas abusivas dos processos judiciais, melhorando a identificação de processos similares sem que precise do número do processo para acessá-lo.

Palavras-chave: mineração de texto, recuperação de informação, ministério público, topic model, processo judicial

1 Introdução

O crescimento na variedade e quantidade de dados disponibilizados por meio de mídias eletrônicas gerou uma grande demanda por maneiras mais eficientes e eficazes de organizar e pesquisar informações (SCHÜTZE *et al*, 2008). A primeira definição de Recuperação de Informação foi dada por Calvin Mooers (1951):

> "a recuperação de informação é o nome para o processo ou método por meio do qual um possível usuário da informação é capaz de converter sua necessidade em uma lista de citações a documentos armazenados contendo informações úteis a ele. A recuperação de informação embarca os aspectos intelectuais da descrição da informação e suas especificidades de pesquisa, e também qualquer sistema, técnicas ou máquinas que são empregadas para executar a operação. É crucial para a documentação e organização do conhecimento."

No contexto atual, Schütze (2008) a define como:

> "a busca de material (normalmente documentos) de natureza não estruturada (normalmente texto) que satisfaça uma necessidade de informação de grandes repositórios (normalmente armazenadas em computadores)." (SCHÜTZE *et al*, 2008)

Além disso, também pode ser referido como o processo de se obter e apresentar um conjunto de informações similares dentro de um sistema que contenha uma grande quantidade de informações, desde documentos, até imagens, áudios, mapas ou vídeos. Encontra aplicações nas mais diversas esferas da sociedade, desde sistemas de busca em bibliotecas, até sistemas de recomendação como os que o Google utiliza.

Dentro da imensidão de dados que existe estão os milhões de pareceres jurídicos. Os pareceres jurídicos são documentos que podem ser emitidos por um órgão público, como o Ministério Público, e também por juristas particulares, como advogados ou consultores jurídicos. Os MPs usam pareceres públicos para exercerem suas funções de: defender os interesses sociais, como o direito à vida, à saúde, à moradia, à liberdade, à educação, dentre outros; fiscalizar as leis, atuando na defesa da ordem jurídica e do regime democrático; assim como defender o patrimônio cultural, o meio ambiente e direitos e interesses da coletividade (https://mpsc.mp.br/o-ministerio-publico/entenda-o-ministerio).

Atualmente, o sistema de busca de pareceres do MPSC exige um nível alto de conhecimento acerca do processo por parte do usuário, o que acaba dificultando o acesso ao canal pelo qual o MPSC se comunica oficialmente, o que restringe o público que utiliza o sistema. O sistema atual tem a interface apresentada na **Figura 1**. Um sistema com mais facilidade de acesso e pesquisa melhoraria a busca por informações pela sociedade, ampliando a transparência que deve existir na administração pública. Isso evita o deslocamento desnecessário das pessoas para escritórios de advocacia, a fim de compreender questões simples sobre temas de seu interesse.

Este trabalho tem como objetivo dar o primeiro passo para melhorar a recuperação de informação em processos no MPSC. Para esse fim, é apresentado um estudo de caso com o propósito de expor uma metodologia de agrupamento de processos em tópicos. Os tópicos são agrupados dentro de assuntos gerais, com o propósito de criar uma rede de relação entre todos os pareceres gerados pelo MP. O estudo foi realizado com dados relacionados ao assunto Práticas Abusivas, visto ser um tema abrangente que contém diversos tipos de práticas que podem ser agrupadas em tópicos.

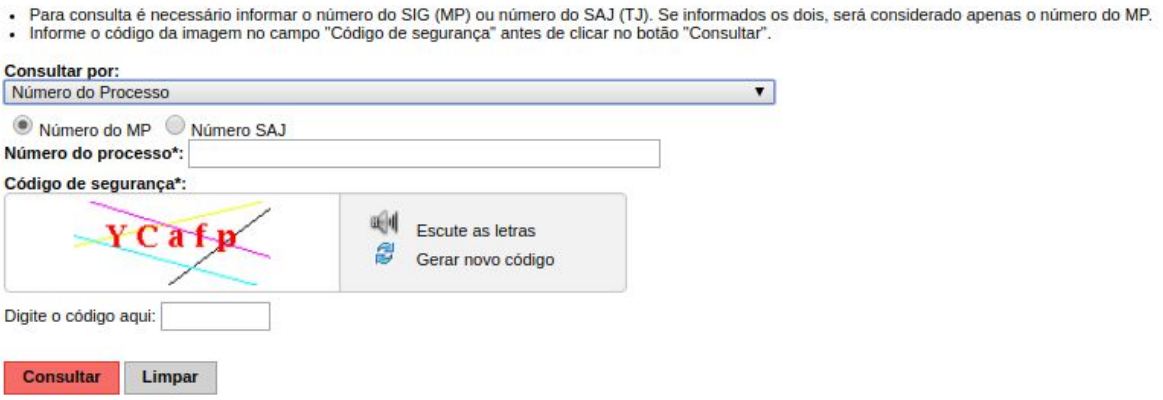

Figura 1: Sistema de busca de processos do MPSC
Fonte: https://www.mpsc.mp.br/servicos/procedimentos-e-processos

A seção 2 apresenta a fundamentação teórica sobre recuperação de informação, assim como uma descrição do LDA. A seção 3 apresenta a metodologia de processamento dos pareceres públicos para que os tópicos possam ser encontrados. Em seguida, a seção 4 descreve o estudo de caso e por fim tem-se as considerações finais.

2 Fundamentação

No campo de Recuperação de Informação, a utilização de modelos probabilísticos, como *topic modeling,* tem se tornado uma escolha popular para o aprendizado não-supervisionado de tópicos latentes dentro de coleções de documentos (CHANG, 2009). O objetivo desse modelo é encontrar descrições curtas de documentos dentro de grandes coleções de documentos. Tais descrições facilitam atividades como classificação, sumarização, similaridade e relevância dos documentos que, por sua vez, auxiliam a recuperação de informações. Assim, esse método pode ser usado em diversos campos cujos dados são discretos, como em dados genéticos, imagens, redes sociais, processamento de textos (BLEI, 2012), áreas da saúde (SONG *et al.*, 2017), identificação de linhas de pesquisa sobre transporte público (SUN & YIN, 2017) e até análise da evolução dos temas da agenda política do parlamento europeu (GREENE & CROSS, 2017). Dentro do *topic modeling*, o modelo mais difundido, e usado neste artigo, é o modelo probabilístico generativo de *Latent Dirichlet Allocation* (LDA).

Por meio da LDA é inferida uma associação entre palavras dos documentos e tópicos. Os tópicos, por sua vez, representam os assuntos e temas que compõem os documentos. Assim, todas as palavras pertencem a algum tópico e todo documento pode ser representado por um conjunto de tópicos, cada um deles com maior ou menor probabilidade de comporem o documento. Por fim, o objetivo é descobrir padrões no uso da palavra e conectar documentos que exibem padrões similares aos mesmos tópicos.

Segundo Blei (2003), a LDA parte do princípio que os dados (o *corpus*, denominado por ***D***) de entrada são gerados a partir de uma mistura de tópicos aleatórios, e que cada tópico é caracterizado por uma distribuição de palavras. Essa distribuição, que dá nome ao método, é a distribuição de *Dirichlet*. A palavra é a unidade básica dos dados, e é chamada de *token*. O vetor de *tokens* discretos (vocabulário) da coleção de documentos tem tamanho V. O documento (**w**) é uma sequência de N palavras, o corpus é uma coleção de M documentos e θ representa uma mistura de tópicos. Desta forma, assume-se que o modelo generativo para cada documento **w** num *corpus* ***D***:

1. Escolhe N ~Poisson (ξ)
2. Escolhe θ ~ Dir (α)
3. Para cada uma das N palavras w_n:
    a. Escolhe um tópico z_n ~ Multinomial (θ).
    b. Escolhe uma palavra w_n de $p(w_n|z_n,\beta)$, uma probabilidade multinomial condicionada no tópico z_n.

Muitas simplificações são feitas para esse modelo básico. Primeiramente, assume-se que a dimensão k da distribuição de Dirichlet (e, consequentemente, o número de tópicos, **z**) é conhecida e fixa. Em segundo lugar, a probabilidade das palavras é parametrizada por uma matriz β com dimensões $k \times V$ onde $\beta_{ij} = p(w^j = 1|z^i = 1)$ sendo inicialmente tratada como uma quantidade fixa a ser estimada. Por fim, a distribuição de Poisson não é crítica para o cálculo e N é independente das variáveis (θ e z) geradas pelos dados. Para início do cálculo, são estabelecidos hiperparâmetros pelo usuário, α e β, que indicam respectivamente as densidades de tópicos por documentos e de palavras por tópicos.

Desta maneira, o modelo de representação gráfica do LDA é ilustrado pela **Figura 2**, com todas as variáveis citadas representadas pelos círculos.

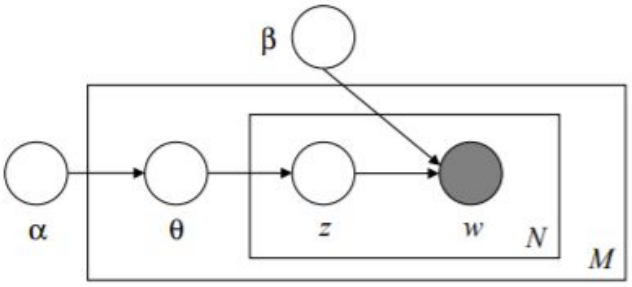

Figura 2: Modelo de representação gráfica do LDA

Blei (2012) relatou que os algoritmos de *topic modeling* geralmente se enquadram em duas categorias: por amostragem ou variacional. Para este artigo, foi usado o algoritmo por amostragem *WarpLDA* baseado no princípio de *Monte-Carlo Expectation Maximization (MCEM)* e simplifica o *design* do sistema de maneira distribuída com alocação de memória mais eficiente, por meio da distribuição de matrizes esparsas. O seu grande diferencial de outros algoritmos derivados do LDA é sua rapidez de processamento que pode alcançar 11G de tokens por segundo com 256 máquinas (CHEN *et al*, 2016). Em outras palavras, o *LDA* assume que cada palavra do vocabulário pertence a um tópico. O tópico é atribuído aleatoriamente a cada palavra. O algoritmo itera sobre as palavras modelando as probabilidades de que um conjunto de palavras pertença ao mesmo tópico, até que um critério de convergência (ou número de iterações) seja atingido.

A parametrização do algoritmo é feita por meio da matriz palavras por documento ($Vx\theta$), os hiperparâmetros (α e β), o número de tópicos (k), o número de iterações, precisão de convergência e número de checagem para convergência.

Muito embora a LDA seja uma excelente ferramenta na mineração de texto, sem um pré-processamento o algoritmo processará todas as palavras e símbolos dos documentos fornecidos. Isso pode dificultar e aumentar o tempo de tratamento de dados, pois os documentos podem estar 'sujos' e conterem muitas palavras sem um sentido próprio como sinais de pontuação, *stopwords* (artigos, preposições, verbos de ligação, entre outras), nome próprios de pessoas ou lugares, além de diferenciar palavras com desinências nominais de número. Assim, todo o processo de encontrar tópicos com temas relevantes se torna impreciso, pois muitas dessas palavras se agrupam não representando um assunto ou significado. Em vista disso, é necessário um pré-processamento de qualidade de texto.

3 Metodologia

Para realizar a mineração de texto foi utilizado *topic modeling* com o algoritmo de *LDA*. A fim de melhorar a caracterização dos tópicos nos processos judiciais foi necessário pré-processar os dados de entrada. A **Figura 3** mostra em detalhes as etapas realizadas.

3.1 Pré-processamento (PP) de texto

O PP dos pareceres é o passo inicial da metodologia do artigo. Primeiramente, o procedimento consistiu em extrair o texto dos documentos PDF e agrupá-los pelo número do processo. Para que pudessem ser aplicadas as funções de PP, e posteriormente serem compilados pelo algoritmo LDA, foi usada a técnica de *tokenização,* que consistiu em caracterizar os pareceres por um vetor de palavras.

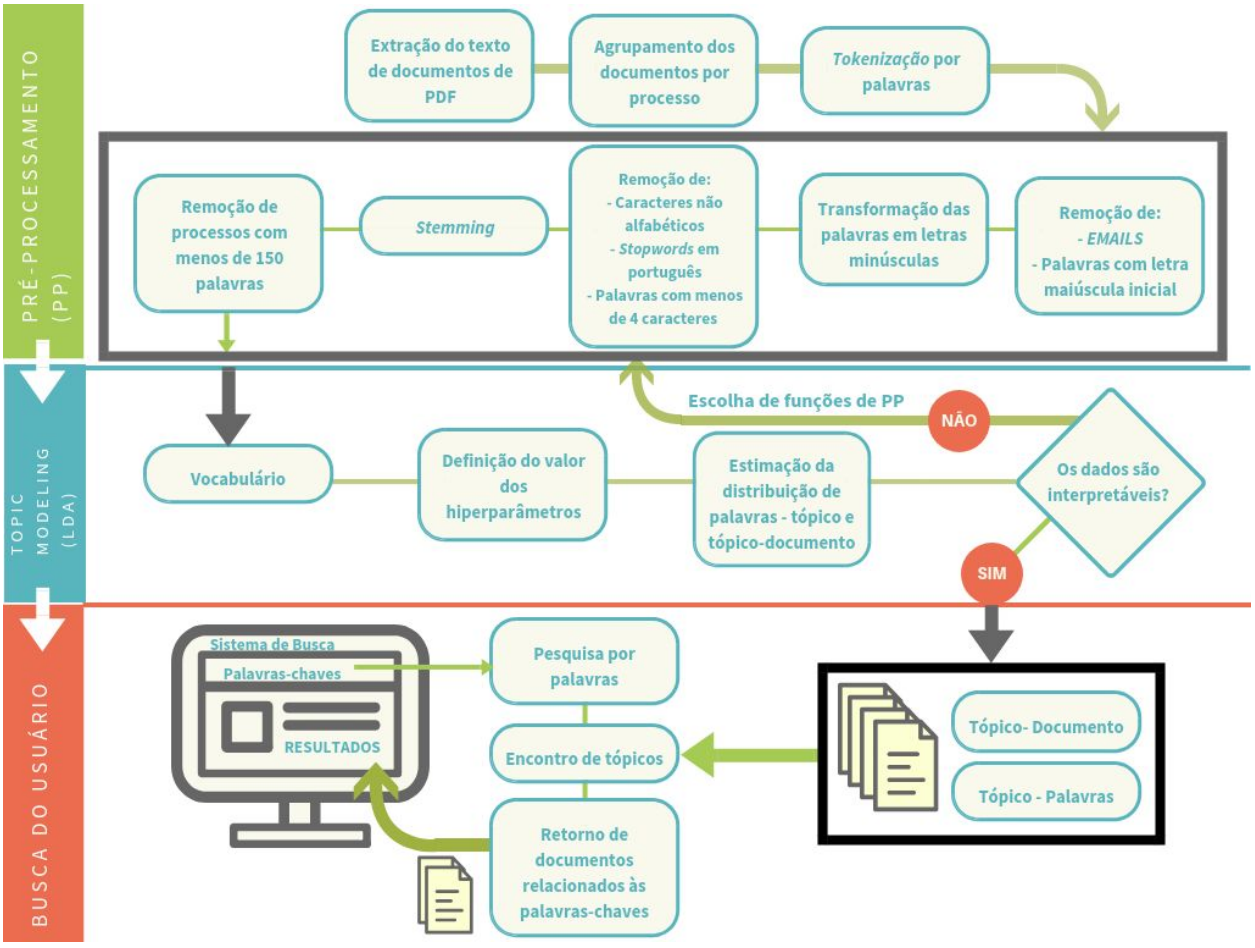

Figura 3: Etapas de processamento

Os passos seguintes foram iterativos, já que era necessário avaliar se os dados gerados eram interpretáveis. O objetivo deste passo foi 'limpar' os dados de entrada para que, nos dados de saída, as palavras fossem relevantes e representassem o tópico verdadeiramente. Dentre as funções geradas, em que foi observado melhor tratamento dos dados, estiveram a filtragem de palavras a partir de remoção de *e-mails* e palavras com letra maiúscula inicial (representando substantivos próprios, como nomes de pessoas ou de cidades). Após a primeira filtragem, houve a conversão de todo texto para letras minúsculas e remoção de: caracteres não alfabéticos (como

numerais, encontrados em processos, datas, prazos), *stopwords* em português e palavras de até 3 caracteres. Outra função foi o RSLP *Stemmer* (Removedor de Sufixos da Língua Portuguesa) que remove os sufixos e depois completa as palavras com aquela mais usada no texto de mesmo radical. Por fim, a última função proposta deste passo contou com a remoção de processos com menos de 150 palavras, pois foi observado que quando um parecer tinha essa quantidade de palavras ou menos, havia muita linguagem técnica e nenhum desenvolvimento de algum tema, refletindo em dados menos interpretáveis.

Depois destas funções, o vocabulário de palavras que comporia a matriz de entrada palavra-documento estava pronto e junto a ele foi necessário a definição dos hiperparâmetros que foram configurados conforme **Tabela 1**.

Parâmetros	Configurações
Número de tópicos	15
α	0,1
β	0,01
Número de iterações	1000
Precisão Convergência	0,001
Número de checagem para convergência	25

Tabela 1: Valores de parâmetros de entrada no algoritmo de WarpLDA

Por fim, houve a estimação da distribuição das matrizes de palavras-tópico e tópico-documento pelo algoritmo de *WarpLDA*. Se estas matrizes apresentassem dados interpretáveis com significados de palavras conectadas por um tema, esses dados seriam guardados e passíveis de serem consumidos por um sistema de busca. Este sistema a partir da inserção de palavras-chaves pelo usuário, procuraria estas palavras nas matrizes, relacionaria aos tópicos correspondentes e, como resultado da pesquisa, retornaria os documentos associados a esses tópicos.

4 Estudo de Caso

Esta seção do artigo tem como objetivo descrever o estudo de caso feito por meio da metodologia apresentada na seção 3.

4.1 Descrição do corpus

O corpus de documentos analisados contém documentos pertencentes ao assunto 'Práticas Abusivas' do Ministério Público de Santa Catarina. Total de 24.110 documentos e, depois de agrupados por processo resultam em 3.201 processos. Cada processo tem em média 7,5 documentos e 1974 palavras. A **Figura 4** apresenta a distribuição conjunta do número de documentos por processo e o número de palavras por processo, junto com suas respectivas distribuições marginais. Cada hexágono representa a contagem de processos com os respectivos números de documentos e palavras. Regiões mais claras indicam uma contagem maior. Percebe-se através da figura que a maior concentração de documentos encontra-se em regiões com menos documentos e menos palavras por processo (canto inferior esquerdo).

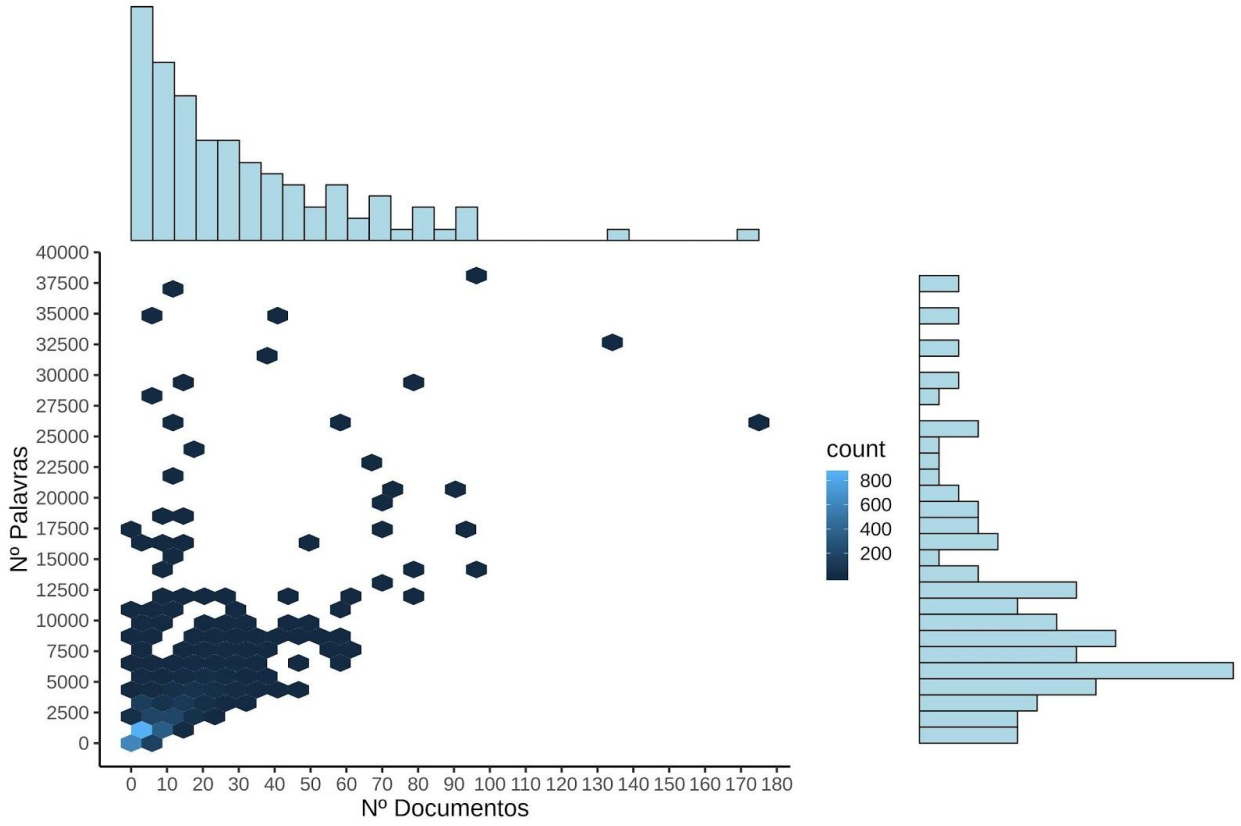

Figura 4: Distribuição conjunta de documentos e palavras por processo

4.2 Resultados

O primeiro resultado foi encontrar quinze tópicos dentro do assunto práticas abusivas. Os tópicos foram nomeados como: acidente de trânsito, cobrança indevida, compras online, educação, imóveis, indenizações, isenções e benefícios, plano de saúde, produto impróprio para venda, regularização em construção civil, serviços públicos essenciais, sistema financeiro e dois tópicos que não foram identificados. A **Tabela 2** apresenta as quinze primeiras palavras que compõem cinco tópicos utilizados como exemplo**.**

PRÁTICAS ABUSIVAS					
	"SERVIÇOS PÚBLICOS ESSENCIAIS"	**"EDUCAÇÃO"**	**"ISENÇÕES E BENEFÍCIOS"**	**"PLANO DE SAÚDE"**	**"PRODUTO IMPRÓPRIO PARA VENDA"**
1	água	aluno	ingressar	médica	alimentares
2	tarifa	ensino	entrar	plano	embalagem
3	transporte	escolar	descontados	seguros	fabricante
4	usuário	matriculada	estudo	cobertura	validade
5	esgoto	acadêmico	benefício	credenciada	mantendo
6	faturamento	estudo	culturais	paciente	rotulagem
7	abastecimento	certificado	anos	cooperativas	desacordo
8	aumentado	educacional	cento	hospitalares	sanitária

9	concessionária	mensalidades	deficiente	internação	depositando
10	ônibus	educação	disponibilizar	hospitalizaçã o	penais
11	município	aula	show	farmacêutico	vencimento
12	reajusto	graduação	idosos	gestão	impróprios
13	munícipes	disciplinados	concessão	escolher	animal
14	passageiros	superiores	cinquenta	prescrição	manipulação
15	reservatório	pólo	concedente	extração	vigente

Tabela 2: Tópicos gerados para o assunto Práticas Abusivas

Outro resultado alcançado foi encontrar processos similares de acordo com os tópicos que os compõem. Esse resultado pode ser verificado na **Figura 5,** que projeta os processos por tópico em duas dimensões. As dimensões são calculadas pelo método *T-SNE* (t-*Distributed Stochastic Neighbor Embedding*) (MATENM & HINTON 2008), que pode ser entendido como uma versão não-linear do método de análise de componentes principais.

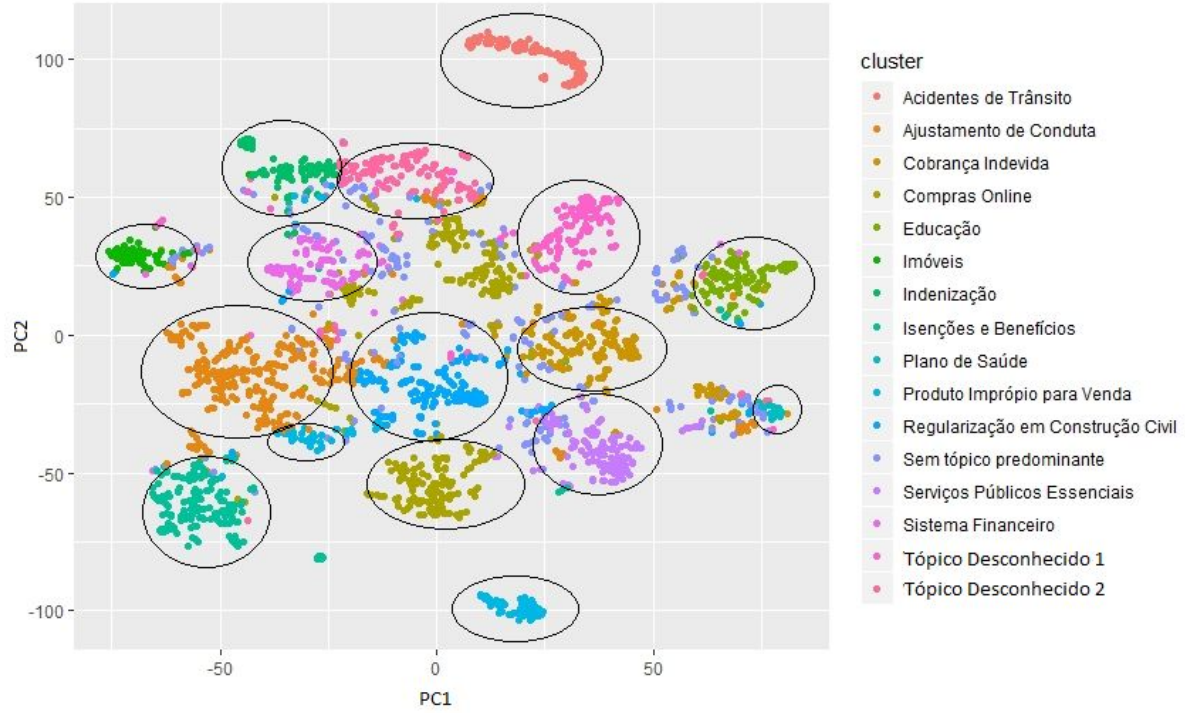

Figura 5: Tópicos gerados para o assunto Práticas Abusivas

A categorização de cada processo em um tópico foi feita em dois passos. O primeiro consiste em categorizar os processos como 'sem tópico predominante' quando nenhum tópico possui a probabilidade maior que 0,3 de compor o processo. O segundo foi atribuir como tópico predominante o tópico que teve a maior probabilidade de compor o processo. A boa categorização dos processos por tópicos auxilia na recuperação de informações em um sistema de busca, através da busca em um conjunto de processos similares. A **Tabela 3** apresenta o número de processos atribuídos a cada tópico, bem como o percentual que o número de processos representa no corpus de estudo.

Tópicos	n	%
Compras Online	326	13,55
Ajustamento de Conduta	323	13,42
Sem tópico predominante	238	9,89
Isenções e Benefícios	184	7,65
Cobrança Indevida	182	7,56
Regularização em Construção Civil	181	7,52
Tópico Desconhecido 2	154	6,4
Tópico Desconhecido 1	132	5,49
Serviços Públicos Essenciais	123	5,11
Educação	106	4,41
Produto Impróprio para Venda	105	4,36
Acidentes de Trânsito	104	4,32
Sistema Financeiro	97	4,03
Indenização	80	3,33
Imóveis	55	2,29
Plano de Saúde	16	0,67

Tabela 3: Número de processos e percentual de documentos no tópico

Essa sumarização é uma maneira de entender em alto nível os resultados do algoritmo LDA. Destaca-se que o tópico que foi atribuído a mais processos foi 'Compra Online', com 13,55% e 'Plano de Saúde' foi o atribuído a menos processos, com 0,67% do *corpus*.

Considerações Finais

O estudo de caso na obtenção de classificação das palavras em tópicos mostrou um bom resultado, pois a maioria das palavras selecionadas obtinham significado referente ao seus tópicos.

No entanto, algumas palavras como o advérbio 'claramente', ou verbos como 'disponibilizar' não possuíam um significado relevante que os diferisse de outros assuntos. Além disso, as palavras 'concessão' e 'concedente', mesmo vindo do mesmo radical foram consideradas distintas no tópico "Isenções e Benefícios".

Trabalhos Futuros

O algoritmo de LDA mostrou um ótimo desempenho na triagem de tópicos no estudo de caso Práticas Abusivas. No entanto, há alguns aspectos para trabalhos futuros que ainda podem ser melhorados, como os seguintes itens:

1. Limpeza mais detalhada de numerais escritos, sites eletrônicos, palavras sem sentido próprio, como preposições ou advérbios, ou seleção por classe gramatical, como somente de substantivos e adjetivos;
2. Agrupamento de palavras ou expressões compostas, tendo os exemplos de: abastecimento de água, ensino a distância, internet móvel, propaganda enganosa, pirâmide financeira, meia-entrada, cinquenta por cento;

3. Inclusão de siglas que podem ser importantes palavras dentro dos tópicos, mas que foram retiradas por possuírem letra maiúscula inicial, como: MEC, PROCON, DPVAT;
4. Criar tópicos específicos só para referência de leis e artigos para auxiliar como base de fundamentação de processos dentro do assunto inserido;
5. Análise de Sentimentos: verificar a eficácia empírica da utilização de métodos de análise de sentimentos dentro de um contexto jurídico.

O próximo passo é aplicar essa metodologia no tema mais abrangente de assuntos dos pareceres públicos e verificar o comportamento das palavras nos tópicos, bem como a inter-relação de temas.

Agradecimentos

Esse estudo foi financiado com o apoio de Samaia IT - Integradora de Sistemas.

Bibliografia Consultada

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993-1022, 2003.

BLEI, David M. Probabilistic topic models. **Communications of the ACM**, v. 55, n. 4, p. 77-84, 2012.

CHANG, Jonathan et al. Reading tea leaves: How humans interpret topic models. In: **Advances in neural information processing systems**. 2009. p. 288-296.

CHEN, Jianfei et al. Warplda: a cache efficient o (1) algorithm for latent dirichlet allocation. **Proceedings of the VLDB Endowment**, v. 9, n. 10, p. 744-755, 2016.

GREENE, Derek; CROSS, James P. Exploring the political agenda of the European parliament using a dynamic topic modeling approach. **Political Analysis**, v. 25, n. 1, p. 77-94, 2017.

MAATEN, Laurens van der; HINTON, Geoffrey. Visualizing data using t-SNE. **Journal of machine learning research**, v. 9, n. Nov, p. 2579-2605, 2008.

MOOERS, Calvin N. Zatocoding applied to mechanical organization of knowledge. **American documentation**, v. 2, n. 1, p. 20-32, 1951.

SCHÜTZE, Hinrich; MANNING, Christopher D.; RAGHAVAN, Prabhakar. **Introduction to information retrieval**. Cambridge University Press, 2008.

SONG, Chang-Woo; JUNG, Hoill; CHUNG, Kyungyong. Development of a medical big-data mining process using topic modeling. **Cluster Computing**, p. 1-10, 2017.

SUN, Lijun; YIN, Yafeng. Discovering themes and trends in transportation research using topic modeling. **Transportation Research Part C: Emerging Technologies**, v. 77, p. 49-66, 2017.