UNIVERSIDADE FEDERAL DE SÃO CARLOS
Bacharelado em Engenharia de Computação

Matteus Guilherme de Souza

Comparação de Tópicos nos Discursos dos Deputados Federais nos Anos de 2014 e 2022 com Métodos de Aprendizado de Máquinas

São Carlos-São Paulo

2019

Matteus Guilherme de Souza

Comparação de Tópicos nos Discursos dos Deputados Federais nos Anos de 2014 e 2022 com Métodos de Aprendizado de Máquinas

Trabalho de Conclusão de Curso apresentado ao Bacharelado em Engenharia de Computação para obtenção do título de Bacharel em Engenharia de Computação

Orientação Prof. Dr. Alan Demétrius Baria Valejo

São Carlos-São Paulo

2019

Dedico este trabalho a meus pais Edson e Verônica, a meus avós e todos que me apoiaram não importando a situação e necessidade neste período nada fácil.

Agradecimentos

Primeiramente agradeço a meus pais, que se esforçaram grandiosamente durante minha vida para que eu tivesse acesso a uma educação de qualidade e pudesse então entrar na UFSCar para realizar minha graduação. Eu sei o quanto esta jornada universitária não foi fácil para nenhum dos lados envolvidos.

Aos meus queridos avós, que, mesmo sem saber, mudaram toda a trajetória desta família ao enfrentar enormes dificuldades para que meus pais pudessem ter uma educação formal de qualidade. Fazendo que eu tivesse a oportunidade de estudar nas melhores escolas de minha cidade Campinas.

Agradeço aos meus diversos professores de Colégio, que me acompanharam durante o meu descobrimento das Ciências tanto Exatas quanto Humanas, em especial da Matemática durante meu aprendizado formal. Gostaria de agradecer em especial a Professora Otília, professora de Ensino Médio, que sempre viu o bom conhecimento em mim estabelecido para a área das Ciências Exatas mesmo eu não sendo o aluno mais esforçado da sala.

Aos meus colegas de faculdade, que estiveram comigo no dia a dia desta incrível e difícil jornada. Gostaria de agradecê-los imensamente, vocês proporcionaram felicidade para este que escreve.

A professora Sylvia Iasulaitis, que me introduziu ao mundo da pesquisa durante minha jornada na Universidade e que me apresentou o professor que faz a interface perfeita entre as Computação e as Ciências Sociais, meu orientador, Alan Demétrius Baria Valejo.

Agradeço especialmente ao meu orientador, Alan, por ter me orientado neste começo de jornada acadêmica, ter aceito ser meu orientador de Iniciação Científica e agora meu orientador de Trabalho de Conclusão de Curso. Agradeço por todos os conselhos e conhecimentos transmitidos durante esses anos de trabalho conjunto para o avanço da ciência.

Gostaria de agradecer por último o grupo de pesquisa Interfaces de modo geral, o qual foi uma fonte inesgotável de conhecimento e de avanço em minha carreira acadêmica e meus conhecimentos gerais, sem este grupo este trabalho nunca teria acontecido.

"Afortunado realmente é o homem conhece precisamente a si mesmo e tem uma noção correta entre o que ele pode conseguir e o que ele pode usar.
(Henri Cartier-Bresson)

Resumo

A Câmara dos Deputados do Brasil é um órgão político e institucional brasileiro a nível federal, sendo um dos componentes, junto ao Senado Federal, do Poder Legislativo. Dentro da Câmara dos Deputados do Brasil são elaboradas as leis que regem o Brasil. Para tal são realizados, dentro deste espaço, discursos onde os deputados discutem as propostas de lei criadas por eles e seus diversos pares. Os discursos possuem portanto, uma enorme riqueza do ponto de vista de dados a serem utilizados, dados os quais podem ser amplamente utilizados pela população e por cientistas de diversas áreas para o entendimento da representatividade e dos temas discutidos dentro deste espaço. Tais dados são então disponibilizados para a população de forma gratuita e aberta, porém de maneira bruta. Surgindo então, espaço para exploração científica e tecnológica de métodos, ferramentas, análises e metodologias que tragam informações relevantes a partir de tais dados. Isto é muito importante para que a população de maneira geral entenda melhor o que acontece dentro do processo democrático de direito brasileiro. Imerso neste contexto há o surgimento de métodos de Aprendizado de Máquinas, os quais ajudam a modelagem e análise de tais problemas. Este estudo implementa neste cenário, um *pipeline* para a análise dos tópicos, ou termos que aparecem de maneira conjunta, em discursos de Deputados Federais e a tentativa de encontrar tópicos semelhantes do ponto de vista sociopolítico entre diferentes períodos da política brasileira utilizando-se de métodos de Aprendizado de Máquina. Facilitando portanto uma análise das mudanças no cenário político brasileiro, em especial, observando os diferentes atores políticos brasileiros ao longo do tempo. O *pipeline* proposto é uma ferramenta para que cientistas sociais possam de maneira facilitada, ter acesso e realizar análises em tais dados. O *pipeline* aqui proposto pode então ser dividido em 3 partes: A coleta dos discursos antes brutos e sua organização em um formato estruturado e de fácil acesso, a extração dos tópicos dos discursos dos Deputados Federais utilizando-se do algoritmo *Propagation on Bipartite Graphs* e por fim uma análise manual reduzida dos tópicos coletados na etapa anterior de modo a encontrar tópicos que sejam similares em comparação com os encontrados em outro período temporal. Este estudo propõe como um estudo de caso o estudo de 3 partidos, o Partido dos Trabalhadores, o Movimento Democrático Brasileiro e o Partido Liberal. Para o primeiro é possível ver uma volta da necessidade de defesa de direitos antes considerados estabelecidos e pétreos, como a demarcação de terras indígenas. O Partido Liberal por sua vez, sofreu uma mudança grande em seus discursos, se orientando seus discursos para direita do cenário político brasileiro.

Palavras-chave: Aprendizado de Máquinas; Extração de Tópicos; Comparação de Cadeias de Caracteres; Aprendizado Não Supervisionado; Discursos; Deputados Federais.

Abstract

The Chamber of Deputies of Brazil is a Brazilian political and institutional body at the federal level, and is one of the components, along with the Federal Senate, of the Legislative Branch. Brazil's Chamber of Deputies is where the laws that govern Brazil are drafted. To this end, speeches are made in this space, where the deputies discuss the bills they and their various peers have created. The speeches therefore have an enormous wealth of data to be used, data that can be widely used by the population and scientists from various fields to understand the representativeness and issues discussed within this space. This data is then made available to the public free of charge and openly, but in a raw form. There is then room for scientific and technological exploration of methods, tools, analyses and methodologies that bring relevant information from such data. This is very important for the general population to better understand what is happening within the Brazilian democratic process under the rule of law. Immersed in this context is the emergence of Machine Learning methods, which help in the modeling and analysis of such problems. In this scenario, this study implements a *pipeline* for the analysis of topics, or terms that appear together, in speeches by Federal Deputies and the attempt to find similar topics from a socio-political point of view between different periods of Brazilian politics using Machine Learning methods. This makes it easier to analyze changes in the Brazilian political scene, especially by observing the different Brazilian political actors over time. The proposed *pipeline* is a tool for social scientists to easily access and analyze such data. The *pipeline* proposed here can then be divided into 3 parts: the collection of the previously raw speeches and their organization in a structured and easily accessible format, the extraction of topics from the speeches of the Federal Deputies using the *Propagation on Bipartite Graphs* algorithm and finally a reduced manual analysis of the topics collected in the previous stage in order to find topics that are similar in comparison with those found in another time period. This study proposes three parties as a case study: the Workers' Party, the Brazilian Democratic Movement and the Liberal Party. For the former, it is possible to see a return to the need to defend rights that were once considered established and fixed, such as the demarcation of indigenous lands. The Liberal Party, for its part, has undergone a major change in its discourse, moving the speeches to the right of the brazilian political scenery.

Keywords: Machine Learning, Topics Extraction, Character Sequence Comparison; Non-supervised Learning; Speeches; Federal Deputies.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Existe na atualidade uma abundância de dados não estruturados no formato textual e a cada ano um aumento exponencial na quantidade de dados introduzidos ao mundo. Faz-se necessário, então, em tal cenário, a criação de métodos de modelagem e análise de tais dados textuais, visto que os dados podem possuir grande quantidade de informação armazenada de forma latente. Uma vez que é inviável analisar manualmente essa quantidade de dados, é necessário, portanto, a criação de métodos computacionais que realizem tal modelagem e análise de modo automático.

Motivados pela necessidade de processamento de dados textuais, em 2003, emergiu uma área de pesquisa denominada Modelos Probabilísticos de tópicos (do inglês *Probabilistic Topic Models*)(BLEI, 2012). Modelos Probabilísticos de Tópicos são um conjunto de algoritmos que têm como objetivo a descoberta de temáticas ocultas em coleções de textos (corpus). A extração de tópicos possui grande importância, uma vez que pode ser utilizada para organizar e agrupar um subconjunto de textos pelos seus respectivos temas. Outro exemplo é a descoberta de temas ocultos em uma determinada coleção de textos, os quais poderiam passar despercebidos caso fossem analisados manualmente.

Atualmente, pesquisadores têm procurado meios para melhorar o desempenho de tais algoritmos, em especial, o espaço computacional usado por esses métodos, como o *Latent Dirilecth Allocation* (LDA)(BLEI; NG; JORDAN, 2003), já que tais métodos utilizam-se de estruturas matriciais para a realização dos cálculos necessários, levando a um alto custo computacional para que tais métodos sejam executados em grandes conjuntos de dados. Além disso, novos métodos tentam modelar o corpus de modo a agregar novos tipos de informações semânticas ou de contexto, a fim de melhorar a precisão dos tópicos encontrados. Com base na lacuna observada, pesquisas foram desenvolvidas focando na tentativa de utilizar grafos como estrutura de dados, de modo a armazenar os dados textuais de forma eficiente e incorporar as relações e os padrões de conectividade entre palavras presentes em um corpus, como proposto em (FALEIROS; VALEJO; LOPES, 2019; FALEIROS; ROSSI; LOPES, 2017). Temos, portanto, que tal área refletiu sucesso nas pesquisas recentes, sendo de grande importância a área de extração de tópicos em documentos textuais a partir de estruturas de grafos.

A Câmara dos Deputados do Brasil é um local de grande importância para a população do país, uma vez que é o órgão responsável pela análise e criação de leis que regem o país. Tais deputados têm diversas atribuições, tais quais a criação e análise de projetos de leis, decretos legislativos, emendas à Constituição, dentre outros. Durante tais processos, uma parte muito importante são os discursos proferidos em plenário, uma vez que

os deputados apresentam, defendem ou criticam os demais projetos de seus pares. Portanto, faz-se necessário que a sociedade tenha conhecimento dos tópicos mais discutidos pela representatividade escolhida, assim entendendo o que está sendo efetivamente abordado no processo democrático em diferentes períodos políticos ou diferentes legislaturas. Sendo importante tal processo também para especialistas de diversas áreas, os quais procuram saber sobre o que tais deputados discutiram durante as sessões do plenário.

Este estudo debruça-se sobre dois períodos importantes para a política brasileira, o ano de campanha de Dilma Rousseff em sua segunda vitória e o ano de campanha de Luiz Inácio Lula da Silva em sua terceira vitória. Períodos os quais foram extremamente conturbados de maneira política Santos (2017), Seibt et al. (2023), sendo neste momento as duas últimas vitórias da esquerda institucional no Brasil ao cargo de Presidente da República.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho de conclusão de curso tem por objetivo apresentar um *pipeline* para a comparação de similaridade de tópicos entre discursos de deputados federais durante diferentes épocas da política brasileira. Além da realização de um estudo utilizando tal *pipeline*. A extração de tópicos será realizada utilizando o algoritmo *Propagation on Bipartite Graphs*, bem como técnicas de pré-processamento de dados.

1.1.2 Objetivos Específicos

Os objetivos específicos podem ser elencados em:

- Contribuir com a literatura multidisciplinar sobre métodos para extração de tópicos em discursos de deputados;
- Realizar a extração de tópicos em dois conjuntos de deputados, focando na separação por período histórico;
- Analisar os resultados obtidos, comparando os tópicos obtidos para cada um dos períodos analisados, de modo a conseguir concluir se os partidos mantiveram os seus tópicos ao longo do tempo ou se houve uma mudança significativa nos tópicos apresentados.

1.1.3 Organização do Trabalho

Este trabalho será dividido em cinco capítulos. No Capítulo 1 são apresentados a proposta de pesquisa e os objetivos que serão explorados. O Capítulo 2 abrange a fundamentação teórica necessária para cumprir com o objetivo proposto, passando pelos conceitos de Aprendizado de Máquinas, Aprendizado de Máquina com Grafos e Extração de Tópicos, além de justificativas além dos algoritmos utilizados e das métricas escolhidas para avaliar os resultados. No Capítulo 3 exibe-se os métodos utilizados para a execução dos experimentos. O Capítulo 4 destina-se à análise dos resultados obtidos e comparação dos tópicos obtidos para determinação da polarização ou não dos partidos. Por fim, o Capítulo 5 expõe as conclusões do trabalho.

2 Fundamentação Teórica

Neste capítulo, serão apresentados os principais conceitos utilizados para embasamento do trabalho e experimentos realizados.

2.1 Aprendizado de Máquina

Aprendizado de Máquina(AM) é um subcampo da Inteligência Artificial(IA) que busca desenvolver modelos que possam "aprender" por meio da experiência. AM estuda métodos computacionais para adquirir novos conhecimentos, novas habilidades e novos meios de organizar o conhecimento já existente (MITCHELL, 1997). O aprendizado se dá por meio de algoritmos que, baseados em estatística, identificam regras e padrões em grandes bases de dados. O estudo de técnicas de aprendizado baseado em computador também pode fornecer um melhor entendimento de nosso próprio processo de raciocínio (MITCHELL, 1997).

As técnicas de AM utilizam a abordagem de aprendizado indutivo, que consiste em obter conclusões genéricas, regras e padrões a partir de um conjunto particular de exemplos ou casos particulares previamente observados. O aprendizado indutivo pode ser dividido, de forma mais geral, em três tipos principais, descritos a seguir: supervisionado, semi-supervisionado e não-supervisionado.

Suprevisionado consiste na criação de um modelo capaz de aprender a partir de um conjunto de dados previamente rotulados e generalizar para instâncias não conhecidas (KOTSIANTIS et al., 2007). Neste tipo de AM, existe uma figura de um "professor externo"ou "especialista", o qual apresenta o conhecimento do ambiente por conjuntos de exemplos na forma: entrada e saída desejada (KUBAT, 1999). Desta forma o modelo possuirá um espaço definido e limitado de resultados que serão utilizados como referência durante o aprendizado e a classificação.

Não-Supervisionado neste caso não é fornecido conhecimento do ambiente ao modelo, ou seja, o modelo aprende sem a existência de um "professor externo". Em outras palavras, no aprendizado não supervisionado, o modelo desconhece a "resposta certa" e tenta encontrar uma estrutura topológica natural ou padrões naturais intrínseco nos dados (DAUMÉ, 2017). Nesse caso, o algoritmo de AM representa (ou agrupa) as entradas submetidas conforme uma estrutura natural presente aos dados.

Semi-supervisionado Uma limitação natural do aprendizado supervisionado é que ele necessita de um grande conjunto de instâncias rotuladas para alcançar um bom

desempenho. No aprendizado semi-supervisionado os dados de treinamento contêm poucos exemplos rotulados e um grande número de exemplos não rotulados. O objetivo de um modelo de aprendizado semi-supervisionado é fazer uso efetivo de todos os dados disponíveis, não apenas dados rotulados. Isto é, o modelo aprende tanto com os dados rotulados, quanto com as estruturas topológicas e os padrões naturais contidos intrinsecamente nos dados.

2.2 Grafos

Grafos são estruturas matemáticas usadas para modelar uma relação em par de objetos. Um grafo neste contexto é feito a partir de arestas e nós, também chamados de vértices. Os grafos emergiram como uma ferramenta de representação e análise de sistemas reais, devido à sua capacidade de descrever unidades e relações, bem como auxiliar a entender seu comportamento, organização, evolução e dinâmica (NEWMAN, 2001; WATTS; STROGATZ, 1998). Como exemplo conhecido pode-se citar as redes de relacionamento, sociais ou online, nas quais indivíduos são representados por vértices e seus relacionamentos, tais como amizade, interações ou hierarquia no trabalho ou laços familiares são representados por arestas.

Grafos podem ser divididos em 2 grandes grupos, os grafos orientados, também chamados de direcionados e não-orientados, também chamados de não-direcionado. A grande diferença entre os dois tipos existentes são suas arestas, onde em um grafo não-orientado a existência de uma aresta libera ao caminhar pelo grafo, sendo a operação de andar entre os nós de um grafo, a passagem em ambos os sentidos. Em outras palavras, na existência de uma aresta entre dois dados nós A e B, é possível fazer o uso da aresta partindo tanto de A quanto de B. Em grafos orientados, há uma mudança de figura, a operação de caminhar não é livre nos dois sentidos da aresta (MERRIS, 2011).

Um tipo de grafos muito importante para este trabalho são os grafos bipartidos, os quais são grafos constituídos de dois grupos de vértices, de modo que dois nós do mesmo grupo nunca são adjacentes (WEISSTEIN, 2002) podendo eles serem orientados ou não.

Tais grafos possuem grande importância uma vez que modelam diversas aplicações, tais quais documentos e termos em um conjunto de textos, compradores e itens para compra em um mercado e avaliadores e filmes em um sistema de avaliação de filmes (ZHA et al., 2001).

2.3 Aprendizado de Máquina com Grafos

Muitas tarefas definidas em grafos possuem um imenso uso de dados, como, por exemplo, a detecção de comunidades (VALEJO et al., 2018) e a classificação de vérti-

ces (ROSSI et al., 2014). Como pode ser visto, problemas endereçados por Aprendizado de Máquina (AM). Particularmente, a tarefa de detecção de comunidades é endereçada utilizando-se de algoritmos de aprendizado não-supervisionado. Em grafos, comunidades significam a organização de vértices em grupos, que podem ser identificados pela existência de um número diferenciado do entorno, ligando vértices de um mesmo grupo, que provavelmente possuem propriedades comuns ou desempenham funções similares no grafo, e poucas arestas ligando vértices de grupos diferentes. O processo de encontrar as estruturas naturais de comunidades em um grafo é chamado de Detecção de Comunidades e é realizado por meio de aprendizado não-supervisionado. Em outras palavras, algoritmos de AM visam encontrar estruturas topológicas naturais nos dados, sem auxílio de rotulação. Um exemplo real de aplicação de Detecção de Comunidades são os sistemas de compras online, onde agrupar os clientes por semelhança de interesses pode proporcionar a criação de sistemas de recomendações mais eficientes (FORTUNATO, 2010; ARRUDA; COSTA; RODRIGUES, 2012).

2.4 Extração de Tópicos

Faz-se necessário, inicialmente, entender o que é um tópico no contexto deste trabalho, e utilizando a definição dada por Blei (2012), a expressão tópico é usada considerando que o assunto do qual se trata em uma coleção de documentos é extraído automaticamente. Assim, tópico pode ser definido como um conjunto de palavras que ocorrem frequentemente em documentos semanticamente relacionados e essa é a definição de tópico tomada no decorrer deste trabalho.

A extração de tópicos, no que lhe concerne, é uma área que procura identificar e descobrir padrões latentes entre documentos e os termos neles presentes de modo que tais padrões sejam significativos para o entendimento das relações entre documentos e termos. A área tem avançado bastante mundialmente com trabalhos como Blei, Ng e Jordan (2003), Blei (2012). Estes trabalhos falam sobre modelos probabilísticos para a extração de tópicos, ou seja, modelos estatísticos que utilizam as palavras dos textos para descobrir os termos importantes para tal texto ou corpus de texto.

A extração de tópicos possui grande aplicação na área de linguagem natural e descoberta de padrões. Isto pode ser utilizado como, por exemplo, faz Bun e Ishizuka (2002) ao extrair tópicos de um determinado texto para a análise de tópicos semanais de notícias ou ainda como feito, junto de outras técnicas, em Dong et al. (2013) para a classificação e recomendação de análises feitas na internet. Pode-se concluir, portanto, que a extração de tópicos é uma área de alto interesse e aplicabilidade. No estudo de Faleiros (2016), por exemplo, têm-se um estudo teórico de extração de tópicos, especificamente em fluxos de documentos textuais.

No meio do caminho tinha uma pedra
tinha uma pedra no meio do caminho
tinha uma pedra
no meio do caminho tinha uma pedra.

Nunca me esquecerei desse acontecimento
na vida de minhas retinas tão fatigadas.
Nunca me esquecerei que no meio do caminho
tinha uma pedra
tinha uma pedra no meio do caminho
no meio do caminho tinha uma pedra.

Figura 1 – Poema "No meio do caminho", de Carlos Drummond de Andrade, cuja modelagem está representada na Figura 2

Dado o aumento exponencial na quantidade de dados gerados de maneira virtual durante as últimas décadas, seria virtualmente impossível a realização de uma análise manual para a extração de tópicos em uma grande quantidade de dados. Instiga-se, portanto, o pensamento de que a extração de tópicos é melhor feita automaticamente por computadores, dado o volume de dados e a alta chance de uma falha humana durante a classificação ou a análise de correlação entre inúmeros textos.

Para demonstrar uma possível modelagem de textos em formato de grafos, será utilizado o exemplo dado em Antiqueira et al. (2005), no qual é modelado o texto da Figura 1 através de grafos. Na modelagem, primeiramente, as *stop words* (palavras que não necessitam ser indexadas, por possuir pouco significado, tais como preposições, artigos, conjunções e outros) são removidas. Em seguida, as palavras restantes são representadas por vértices no grafo. Por fim, pares de palavras adjacentes (que aparecem em sequência no texto) são conectadas de forma orientada. Além disso, ponderam-se as arestas, ou seja, cara aresta terá um número real associado, o qual representa a quantidade de vezes que o par de palavras apareceu de maneira conjunta no texto. Por exemplo, a sequência "meio caminho" apareceu seis vezes e a sequência "caminho meio" aparece uma única vez 2.

No escopo proposto por este trabalho, um algoritmo de AM em Grafos muito importante é o algoritmo proposto em Faleiros, Valejo e Lopes (2019), chamado de *Propagation on Bipartite Graph*(*PBG*). Este algoritmo considera a modelagem dos dados textuais a partir de um grafo bipartido, onde um conjunto de vértices é dividido em dois subconjuntos, representando os documentos e as palavras (vocabulário) existentes em tais textos, respectivamente. As arestas entre documentos e palavras representam a normalização da ocorrência de tais palavras no texto. O PBG é um algoritmo de AM não-supervisionado, ou seja, no caso da extração de tópicos, o algoritmo processa o grafo bipartido de documentos-palavras a fim de encontrar estruturas modulares locais, ou seja, grupos de vértices, naturalmente estabelecidos pelos padrões de conectividade existentes no grafo. Cada grupo de vértices irá representar conjuntos de palavras ou tópicos.

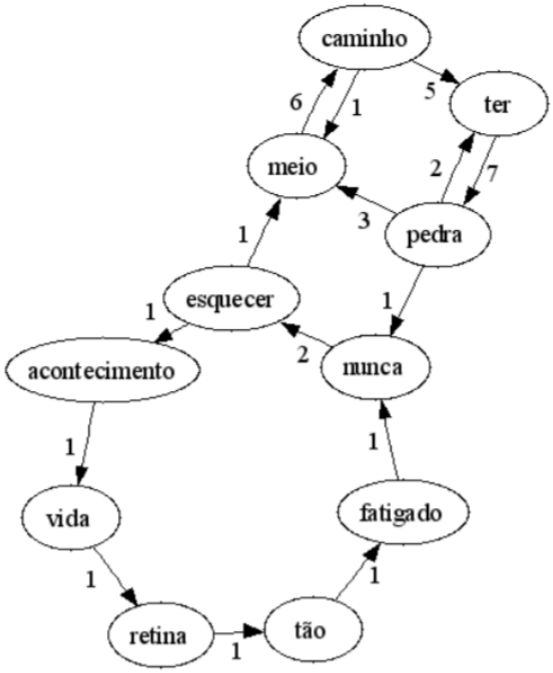


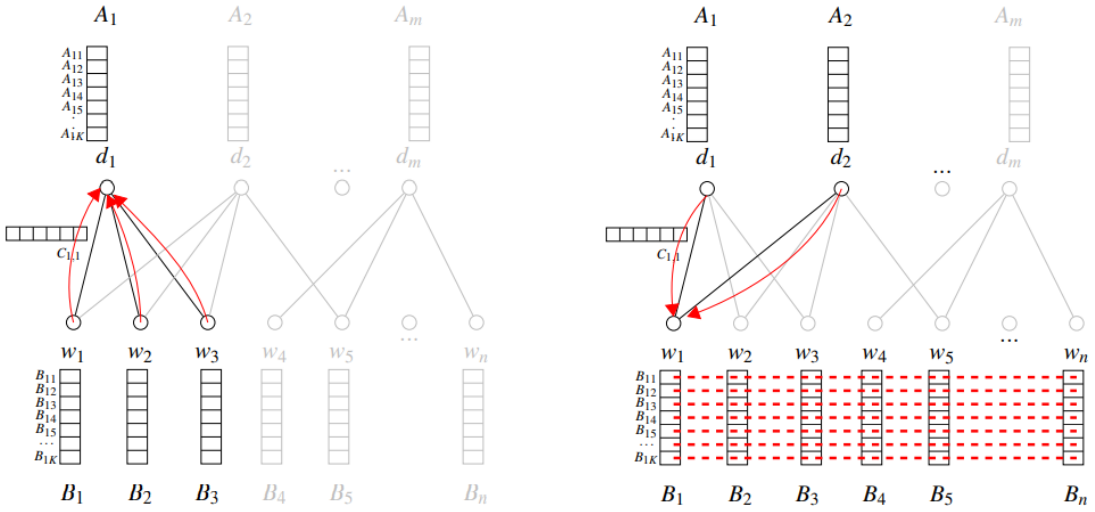
Figura 2 – Representação do texto da figura em formato de grafos complexos (ANTIQUEIRA et al., 2005)

O algoritmo PBG é um algoritmo iterativo que se baseia em duas principais operações, sendo elas chamadas de propagação local e propagação global. A propagação local concentra localmente as informações latentes de cada palavra nos vértices de documentos, como ilustrado na figura 3a. A propagação global, por sua vez, concentra a influência de todos os documentos nos vértices de palavras. Uma vez realizada a propagação local e global, cada vértice será representado por um número real que define a sua probabilidade de pertencer a um determinado tópico, como ilustrado na Figura 3b. A cada iteração o valor calculado anteriormente é usado como base e atualizado até a convergência do algoritmo.

2.5 Semelhança entre Cadeias de Caracteres

Uma necessidade neste trabalho é então medir a similaridade entre dois diferentes tópicos, ou, encontrar os tópicos mais similares. Tal tarefa pode ser definida então, dada a definição já estabelecida de tópicos, pela similaridade das palavras encontradas dentro de tais tópicos. Existem diversos meios, porém para realizar a medição de tal valor. Esta seção irá então abordar alguns dos métodos existentes para realizar tal análise, abordando seus pontos positivos e negativos.

Tal processo é extremamente importante para o processo de análise, uma vez que reduz o espaço a ser observado por um analista, reduzindo seu espaço de busca.



(a) Propagação Local (b) Propagação Global

Figura 3 – Propagação local (à esquerda) e global (à direita) do algoritmo PBG, onde d_m é um documento qualquer e w_n é uma palavra qualquer que existe em pelo menos um documento. Adaptado de Faleiros (2016).

Os métodos para o cálculo de tal similaridade podem ser divididos em três categorias (VIJAYMEENA; KAVITHA, 2016):

- **Similaridade Baseada em Cadeias de Caracteres**: A qual é uma similaridade baseada em operações em cima de sequências e composições de caracteres.
- **Similaridade Baseada em Corpus**: A qual se baseia na similaridade de palavras de um corpus de texto.
- **Similaridade Baseada em Conhecimento**: A qual se baseia na identificação de semelhança entre palavras por meio de informações derivadas de redes semânticas

Esta seção e suas subseções trarão um enfoque sobre algumas medidas de Similaridade Baseada em Cadeias de Caracteres, uma vez que este foi o método utilizado.

Tal método de busca por similaridade pode ser então subdividido em dois tipos, Medidas baseadas em similaridade de caracteres e medidas baseadas em similaridade de termos.

2.5.1 Medidas de Similaridade Baseadas em Caracteres

As Medidas baseadas em Caracteres são medidas que se baseiam no tamanho da cadeia de caracteres presentes em ambas as strings.

Algumas dessas medidas são:

- **Distância de Levenshtein**: É uma medida que se baseia na quantidade de operações (inserções, remoções e substituições) necessárias para transformar uma cadeia de caracteres na outra (LEVENSHTEIN, 1966).
- **Distância de Jaro**: Utiliza-se da quantidade de caracteres iguais e transposições. Uma extensão desta medida é Jaro-Wrinkler, que dá mais peso a cadeias que iniciam iguais (WINKLER, 1990).

2.5.2 Medidas de Similaridade baseadas em Termo

Diferentemente das medidas baseadas em caracteres, estas medidas estão mais preocupadas com a similaridade dos termos do que dos caracteres neles presentes. Tais métodos, em geral, dependem da representação dos termos a serem comparados no formato vetorial.

2.5.2.1 Distância de Manhattan

Utiliza-se do pensamento de qual a distância necessária percorrer em um espaço vetorial seguindo um caminho em formato de grade[1]. Ao final pode ser dada por:

$$d_1(P,Q) = ||P - Q||_1 = \sum_{i=1}^{n} |p_i - q_i|,\ (P,Q)\ vetores \tag{2.1}$$

2.5.2.2 Similaridade de Jaccard

Pode ser calculada de modo que o número de termos compartilhados é dividido pela quantidade total de termos únicos presentes nos dois conjuntos(JACCARD, 1908). A similaridade de Jaccard consegue operar em diversos tipos de conjuntos, sejam eles de caracteres ou vetores.

$$S_j(P,Q) = \frac{P \cdot Q}{P^2 + Q^2 - P \cdot Q},\ (P,Q)\ vetores \tag{2.2}$$

2.5.2.3 Similaridade de Cossenos

Uma das principais medidas utilizadas para a comparação de cadeias de caracteres, mede o cosseno do ângulo entre os termos.

$$S_c(P,Q) = \frac{P \cdot Q}{||P|| \cdot ||Q||},\ (P,Q)\ vetores \tag{2.3}$$

Neste trabalho esta foi a medida utilizada para medir a similaridade entre os tópicos, uma vez que é amplamente empregada na literatura e existem diversas implementações altamente otimizadas.

[1] <https://xlinux.nist.gov/dads/HTML/manhattanDistance.html>

3 Metodologia

Este capítulo será dedicado a explicar os métodos, algoritmos e ferramentas utilizados para a extração dos dados, extração dos tópicos e sua análise. Os algoritmos foram elaborados na linguagem Python[1] com auxílio de diversas bibliotecas para tal linguagem, que serão especificadas mais adiante.

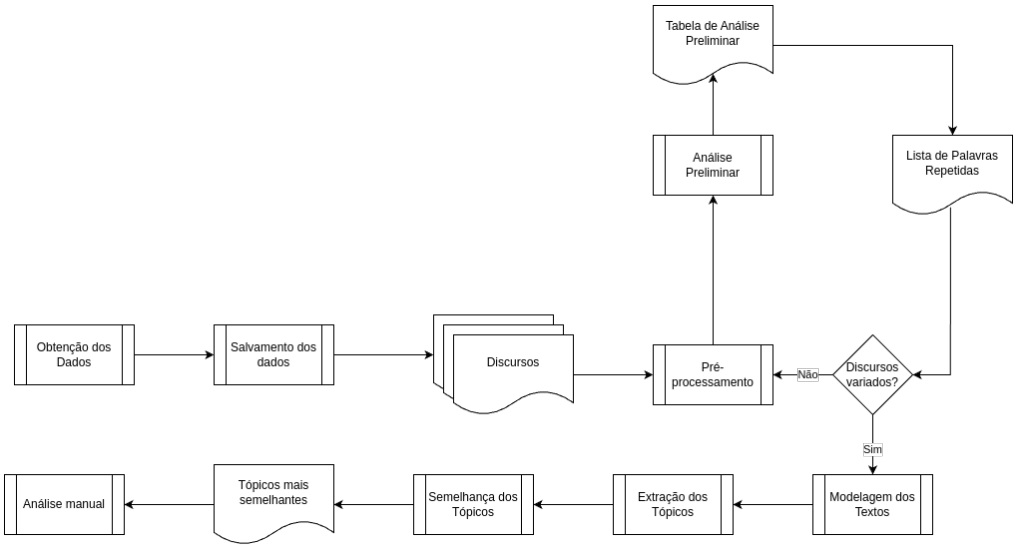
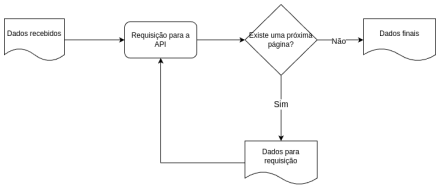
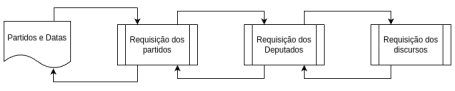


Figura 4 – Linha geral de processamento dos dados

O código em questão, para a extração dos discursos, sua análise preliminar, a extração dos tópicos e sua análise de similaridade pode ser obtido em <https://github.com/Interfaces-UFSCAR/speeches_analisys>. Para a instalação do projeto basta seguir os passos descritos no arquivo README.md do projeto.



(a) Fluxo usado como base para uma requisição de dados.



(b) Fluxo geral das requisições de dados

Figura 5 – Fluxo para a requisição dos dados dos discursos

```
{
  "dataHoraInicio": "2006-07-11T16:38",
  "dataHoraFim": null,
  "uriEvento": "https://dadosabertos.camara.leg.br/api/v2/eventos/13375",
  "faseEvento": {
    "titulo": "Ordem do Dia",
    "dataHoraInicio": null,
    "dataHoraFim": null
  },
  "tipoDiscurso": "PELA ORDEM",
  "urlTexto": "https://imagem.camara.gov.br/dc_20b.asp?largura=&altura=&tipoForm=diarios&selCodColecaoCsv=D&Datain=12%2F7%2F2006&txPagina=35197&txSuplemento=&enviar=Pesquisar",
  "urlAudio": null,
  "urlVideo": null,
  "keywords": "MEDIDA PROVISÓRIA, REAJUSTE, BENEFÍCIO PREVIDENCIÁRIO, PROVENTOS, APOSENTADORIA, PENSÃO PREVIDENCIÁRIA, SUPERIORIDADE, VALOR , SALÁRIO MÍNIMO, REQUERIMENTO, PREFERÊNCIA,
  VOTAÇÃO, EMENDA AGLUTINATIVA DE PLENÁRIO,  AUMENTO, ÍNDICE, REAJUSTE, APOSENTADO, PENSIONISTA, FALTA, QUÓRUM, APROVAÇÃO, AVALIAÇÃO. \r\nRECRUTA, FORÇAS ARMADAS, PROPOSTA, SALÁRIO MÍNIMO,
  PROPOSTA, DEFESA.",
  "sumario": "Inexistência de quorum regimental para votação da proposta concessiva do reajuste de 16,67% às aposentadorias e pensões mantidas pela Previdência Social. Apoio à proposição.
  Apresentação de proposta de pagamento de salário mínimo aos recrutas das Forças Armadas brasileiras e do benefício do auxílio-invalidez à classe.",
  "transcricao": "O SR. JAIR BOLSONARO (PP-RJ. Pela ordem. Sem revisão do orador.) - Sr. Presidente, lamento a falta do quorum necessário para a votação da proposta de aumento, em 16,67%, do
  benefício previdenciário de aposentados e pensionistas.\r\nNa verdade, havia quorum, já que os Deputados estavam na Casa. Contudo, não podiam vir ao plenário votar a favor do aposentado
  porque estariam contrariando o seu patrão, o Governo Federal. Isso é lamentável. São Parlamentares que não têm autonomia para votar. É o voto de cabresto.\r\n\tContrariando o meu partido,
  mais uma vez votei a favor dos aposentados.\r\n\tDiscutiríamos também a medida provisória que trata do reajuste de 10% para os militares. O Governo Federal não pode continuar pagando a 120
  mil recrutas brasileiros provento de 198 reais, praticamente a metade de 1 salário mínimo. Especialmente se esse Governo se diz social e democrático. No entanto, dá as costas para os
  recrutas do Exército, da Marinha e da Aeronáutica, que prestam o serviço militar obrigatório.\r\n\tPropomos seja pago o provento de 1 salário mínimo a cada recruta e concedido
  auxílio-invalidez aos militares, no mínimo equivalente ao soldo do cabo engajado.\r\n\tEra esse o meu recado."
},
```


Figura 6 – Exemplo de um discurso, neste caso do então Deputado Federal Jair Bolsonaro.

3.1 Obtenção dos Dados

Os dados obtidos para a realização do estudo são discursos de Deputados Federais em 2 períodos da política brasileira, sendo os períodos, os anos completos de 2014 e 2022, compreendendo, portanto, dois anos eleitorais brasileiros com uma diferença de duas legislaturas da Câmara dos Deputados ou oito anos. Tal distância temporal é de extrema relevância para a população brasileira pela alta da polarização neste período (FUKS; MARQUES, 2022) dentro do cenário brasileiro. Como pode ser visto nas imagens 5a e 5b o modelo para requisições é bastante simples e composto de alguns elementos utilizados de base para as diversas requisições.

Os dados obtidos foram adquiridos a partir da API de Dados Abertos da Câmara dos Deputados[2]. Os dados são disponibilizados de maneira gratuita e livre através de uma API pública. Foi utilizada para a requisição de tais dados a biblioteca *requests*[3] do Python, a qual traz diversas facilidades quando realizando requisições para sites e API, como era o caso deste estudo.

Após a coleta foi possível obter uma abundância de informações sobre os discursos e quem os proferiu. Tais informações incluem: Os discursos em si, um resumo feito pela Câmara dos Deputados sobre o discurso, palavras-chave, o nome do Deputado que proferiu o discurso, um identificador único para o Deputado, a sigla do partido do qual ele fazia parte ao momento do discurso, um identificador único do partido, além de outras informações, como pode ser visto na Figura 6.

Para a coleta foi então utilizada a classe Session da biblioteca requests, a qual possibilita grande agilidade em pedidos uma vez que há o salvamento da sessão existente, sendo então possível a realização de diversas requisições em uma única sessão, o que agiliza o processo de coleta de recursos.

Os dados então foram salvos em arquivos no formato textual para consulta posterior se necessário ou então realização de diversas análises em cima dos dados. Tal operação por

[1] <python.org>
[2] <dadosabertos.camara.leg.br>
[3] <https://requests.readthedocs.io/en/latest/>

sua vez utilizou a biblioteca Pandas[4], feita para lidar com grandes quantidades de dados e realizar o interfaceamento de tais dados para e de arquivos de texto. Isto possibilita a realização de diversas iterações de análises a serem executadas nos mesmos dados, o que foi de extrema importância neste estudo como mostrado mais a frente. Salvar os dados também permite que tais dados sofram diversas análises por outros especialistas, em especial das Ciências Sociais e Computação sobre os dados crus.

3.2 Análise Preliminar

A análise preliminar dos dados pode ser considerada um ciclo iterativo de análise rápida onde são realizadas 3 operações: pré-processamento dos dados, onde são retiradas palavras recorrentes tanto na língua portuguesa quanto recorrentes nos textos, além de pedaços de texto que não fazem parte do discurso em si e são enviadas pela API juntamente com a transcrição e pontuações, para tal são utilizadas as bibliotecas NLTK[5](MERROUNI; FRIKH; OUHBI, 2020) e SpaCy[6]. Extração de características gerais do corpus, como número de discursos, média de palavras por discurso, número de palavras diferentes no conjunto de textos e palavra que mais aparece dentro dos textos a serem analisados. Além da criação de uma lista de palavras que estão sendo repetidas.

Em relação ao pré-processamento, foram feitas as seguintes operações:

Primeiramente foram retiradas algumas estruturas existentes nos tópicos como a estrutura "(Nome do Deputado - Partido)", a qual não agrega nenhum valor ao tópico é somente uma estrutura de identificação colocada durante a transcrição dos discursos, uma vez que ela é praticamente idêntica à transcrição anexada ao diário oficial. Outra estrutura retirada são as abreviações "Sr(s)."e "Sra(s).", uma vez, que também não trazem valor semântico e estavam durante a análise atrapalhando a análise e não eram tratados pelas bibliotecas de retiradas de palavras de palavra. Por último, uma última estrutura retirada foram os números naturais e reais não escritos por extenso, isso também foi feito, pois tais valores, embora tenham um significado, sozinhos em um tópico não trazem informações relevantes e retiram espaço de palavras que trariam mais significado ao tópico.

Lemmatização dos tópicos, a qual consiste em obter a palavra sem inflexões, chegando ao que também é conhecido como *lemma*. Essa técnica leva em consideração a classe gramatical da palavra, no caso, é possível escolher as classes gramaticais escolhidas, sendo utilizadas como padrão as seguintes classes: substantivo, adjetivo, verbo e advérbio, pois são classes gramaticais extremamente importantes para a construção de uma frase, sendo possível adicionar ou retirar classes gramaticais. Esta técnica tem que ser usada no texto de forma pouco tratada, pois ela utiliza o texto transcrito para, por meio de um

[4] <https://pandas.pydata.org/>
[5] <https://www.nltk.org/>
[6] <https://spacy.io/>

modelo pré-treinado de aprendizado de máquinas, predizer o *lemma* de uma determinada palavra e sua classe gramatical, sendo a estrutura completa de uma frase muito importante para a predição.

Após isso o texto é então tokenizado, o qual é o processo em que um texto é dividido em seus tokens, no caso a separação do texto nas palavras existentes, isso é um passo importante, pois prepara o texto para a realização de alguns passos posteriores.

Por último é realizada a remoção de *stopwords* e pontuação, já que tais valores não são importantes pensando nos tópicos de um determinado discurso.

Após isto é feita uma análise manual para avaliar se as palavras recorrentes nos textos possuem uma mínima variação, em caso afirmativo, segue-se o fluxo para a extração de tópicos, em caso negativo é realizada uma nova iteração dos passos supracitados.

As características de número de discursos, média de palavras por discurso, número de palavras diferentes no conjunto de textos e palavra que mais aparece dentro dos textos a serem analisados trazem informações relevantes para a análise preliminar dos textos

A quantidade total de discursos traz uma ideia geral de se tal corpus de texto é grande o suficiente para a realização de um trabalho de aprendizado de máquinas, uma vez que um corpus muito pequeno de dados não consegue gerar uma generalização dos dados encontrados.

A quantidade de palavras diferentes no corpus, por sua vez, traz ao pesquisador uma ideia geral do tamanho do grafo que será gerado, quando se tratando da quantidade de nós.

A média de palavras presentes por discurso traz um valor numérico para a densidade do grafo que será estudado, ou seja, traz uma ideia tratando-se da quantidade de arestas que podem ser encontradas em um determinado grafo.

Por último, a palavra mais utilizada é utilizada para a verificação visual da existência de palavras que acontecem de maneira repetida dentro dos discursos, servindo como uma palavra recorrente para aquele conjunto de textos em específico. Com isso é então feita uma lista de palavras com alta repetição que atrapalham a visualização de tópicos, pois estão presentes em todos. Palavras as quais serão adicionadas as palavras consideradas *stopwords* e retiradas do texto no momento da realização do processamento. A lista final neste caso ficou com 19 palavras, sendo contida por termos como: senhor, fazer, dizer, deputado, querer, aqui, nº, per, art, presidente, brasil, ter, ano, poder, “ e ”. Como é possível ver, existentes diversas palavras, em geral, palavras muito comuns da língua portuguesa e alguns termos com pouca relevância, mas que pelo ambiente em que os discursos são proferidos são extremamente comuns, como o termo "art", que se refere a um artigo que está sendo discutido.

```
[[0.54935123 0.83559154 0.        ]
 [0.         1.         0.        ]
 [0.         0.         1.        ]]
```

Figura 7 – Exemplo de uma matriz TF-IDF para os textos: "Geeks for Geeks", "Geeks"e "r2j". Obtido de <https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/> em 27/02/2025.

Este tratamento além dos tópicos supracitados possui uma capacidade de realizar uma redução de dimensionalidade do problema, uma vez que reduz a quantidade de palavras utilizadas para a realização da extração de tópicos.

3.3 Extração de Tópicos

Durante a etapa de extração de tópicos é necessária a realização de dois passos principais, a modelagem dos textos em um formato de TF-IDF, o qual transforma os dados em um formato estruturado e que seja aceito pela implementação escolhida do algoritmo PBG. A modelagem do texto utiliza-se da biblioteca Scikit Learn para vetorização dos dados no formato TF-IDF. Os dados vetorizados são então passados para o modelo PBG[7], que realiza os diversos cálculos necessários para encontrar a estrutura intrínseca aos dados e retornar os tópicos encontrados. Foram então extraídos 10 tópicos para cada um dos partidos com 10 palavras em cada tópico.

	Word 0	Word 1	Word 2
Topic 0	boa	acreditar	tarde
Topic 1	entidade	filantrópico	parabenizar
Topic 2	representação	ver	deputados

Tabela 1 – Exemplo de uma tabela de tópicos, onde cada linha é um tópico. Neste caso, 3 tópicos, cada um com 3 palavras.

Os tópicos encontrados são então também salvos, com a utilização da biblioteca Pandas em um arquivo de texto para realização de análises de similaridade entre os tópicos através de métodos computacionais e uma posterior análise manual e encontro de diferenças entre os tópicos. Possibilitando também a um especialista a realização de diversas análises em cima dos tópicos extraídos. Os tópicos extraídos são salvos como matrizes de palavras.

[7] <https://github.com/matteusgui/PyPBG>

```
[[ 9.99999881e-01,  8.81888419e-02,  1.80074498e-02],
[-3.51191908e-02,  1.29839227e-01,  1.96917117e-01],
[ 1.80074498e-02,  3.28570455e-02,  1.00000024e+00]]
```

Figura 8 – Exemplo de uma matriz de semelhanças gerada nesta etapa da análise.

3.4 Análise dos Tópicos

Os tópicos extraídos são então obtidos de seus arquivos de texto, se necessário e é realizada uma conversão das palavras em vetores. Para isso é utilizado um modelo pré-treinado do algoritmo Word2Vec, que transforma palavras em vetores. O modelo utilizado é uma resultante do artigo de Hartmann et al. (2017)[8]. O modelo é então inserido em uma classe da biblioteca GenSim[9] para a representação do modelo em código e sua fácil consulta.

As palavras são então consultadas no modelo, gerando vetores que representam tais palavras de cada tópico, onde cada palavra de uma matriz como representada em 1 é representada por um vetor de 300 dimensões.

É realizada a extração da similaridade de cossenos entre os diferentes termos para cada um dos tópicos. A similaridade de cossenos é dada por um valor entre 0 e 1, onde 0 é a completa diferença entre os vetores e 1 os vetores são iguais.

Para isso são realizadas diversas etapas. Primeiramente, é realizada a semelhança de cossenos entre dois termos de tópicos, isso retorna um valor entre 0 e 1, onde 1 diz que as palavras são completamente iguais e 0 representa a completa não similaridade entre as palavras. São então comparados todos os termos de um determinado tópico com os termos de um segundo tópico do qual se desconfia haver a semelhança. Isso gera uma matriz de similaridade entre dois dados tópicos, como pode ser vista, em exemplo, na Figura 8, sendo necessária a criação de uma matriz em tal formato para cada uma das combinatórias de pares de tópicos a serem comparados.

Em posse das diversas matrizes de semelhança entre dois conjuntos de tópicos, é possível então realizar a média de cada uma das matrizes e, tendo tomado um conjunto de tópicos como referência, é possível obter, através da maior média entre as matrizes de similaridade feitas com o tópico em questão, o tópico do segundo conjunto com maior similaridade com o tópico em questão do primeiro conjunto.

Com tal medida é possível então estabelecer uma relação de similaridade entre 2 conjuntos de tópicos, dada pela média da matriz de similaridades. Neste caso, quanto maior a média da matriz de similaridades, maior é a semelhança entre dois dados tópicos.

Após isto é necessário então a realização de uma análise manual dos tópicos

[8] <http://nilc.icmc.usp.br/embeddings>
[9] <https://radimrehurek.com/gensim/>

semelhantes, de modo a encontrar as semelhanças e diferenças entre os tópicos. Utilizando tal processo é possível então encontrar as diferenças e semelhanças entre os tópicos de um partido em dois períodos distintos. Além de conseguir obter uma ideia aproximada do teor sobre qual tópico é tratado utilizando-se de uma breve análise dos discursos coletados.

4 Resultados

Os resultados obtidos podem ser divididos em partes semelhantes as do Capítulo 3, sendo elas:

- A coleta dos dados;
- A análise preliminar dos tópicos obtidos para os anos de 2014 e 2022;
- Os tópicos extraídos dos discursos;
- A análise dos tópicos existentes tendo em vista a polarização do discurso.

4.1 A coleta dos dados

Os dados obtidos são duas coleções de discursos, sendo elas divididas entre os anos de 2014 e 2022. Em ambos os casos há a coleta de dados sobre três estruturas da Câmara dos Deputados, sendo elas, os partidos, os deputados e os discursos em si.

Os dados da estrutura que representa um partido é constituída por: um identificador único do partido dentro da API, a sigla do partido, o nome registrado do partido e uma URL que apontará para os dados coletados. Tal informação permite fazer separações ainda menores dos dados pelos partidos, o que é muito útil para este estudo.

Para a representação dos deputados de cada partido são coletados os seguintes dados: um identificador, similar ao dos partidos, mas referente a um deputado, o e-mail do parlamentar, um identificador da legislatura do parlamentar, correspondendo a uma legislatura da Câmara dos Deputados, o nome do deputado, a sigla do Partido, a sigla do estado o qual tal parlamentar representa, uma URL que aponta para os dados coletados, uma URL que aponta para os dados do partido do parlamentar, como descrito no parágrafo anterior e, se houver, uma URL que retorna uma imagem do parlamentar em questão.

Por último, têm-se os discursos dos deputados, dos quais são armazenados os seguintes dados na coleta: data e horário de início e fim dos discursos, dados da fase do evento em que está acontecendo o discurso, como data e hora de início e fim do evento, além do título do evento, além disso, são coletados dados de palavras-chave do discurso, as quais não possuem um método aberto pela Câmara dos Deputados, um sumário do discurso, o tipo do discurso, a transcrição, e URLs sobre o evento em que o discurso aconteceu, do texto no diário da Câmara dos Deputados, além de se houver apontadores para o Áudio e Vídeo do discurso.

Em relação à quantidade de dados coletados, foram coletadas, para 2014 16931 discursos em um conjunto de 23 partidos. Para o ano de 2022, por sua vez, foram coletados um total de 16788 discursos em um conjunto de 23 partidos, os quais possuem diferenças com os de 2014. Os dois conjuntos são os dados completos de discursos de seus respectivos anos.

4.2 Análise Preliminar

A análise preliminar, como ilustrado na Figura 4 constitui-se de um processo iterativo que procura tanto reduzir a dimensionalidade dos dados encontrados quanto encontrar palavras que sejam muito recorrentes nos textos e que atrapalhem o processamento dos tópicos. Além disso, ela traz percepções rápidas sobre o texto já levando a algumas conclusões.

Partidos	mediaPalavras	#PalavrasDiferentes	Mais usada	Quantidade
DEM	168.97	11228	presidente	813
PCdoB	168.05	9203	presidente	609
PDT	204.44	8775	ter	466
PEN	164.75	311	mulher	4
PMDB	175.15	18298	presidente	2320
PMN	100.75	1340	presidente	32
PP**	170.65	10798	ter	747
PPS	156.32	8352	presidente	504
PR	137.83	9214	presidente	755
PRB	239.60	6549	ter	206
PROS	187.96	7309	ter	325
PRP	149.80	714	mulher	10
PSB	294.47	13597	ter	681
PSC	158.09	6192	ter	298
PSD	172.56	12155	presidente	1053
PSDB	208.74	14628	ter	1508
PSDC	109.34	528	presidente	9
PSOL	236.34	10804	presidente	468
PT	219.65	24298	ter	4017
PTB	191.68	10826	presidente	641
PTdoB	177.07	2120	ter	44
PV	163.05	5531	presidente	226
SD	186.96	9351	ter	593

Tabela 2 – Resultado da primeira iteração realizada de análise preliminar nos dados de 2014

Observando a Tabela 2 é possível ter algumas análises já, por exemplo, já é possível perceber que partidos como PEN, PRP e PSDC não são muito viáveis para uma análise

dos tópicos, uma vez que possuem uma quantidade muito baixa de discursos, o que não consegue generalizar o conhecimento existente dentro dos discursos.

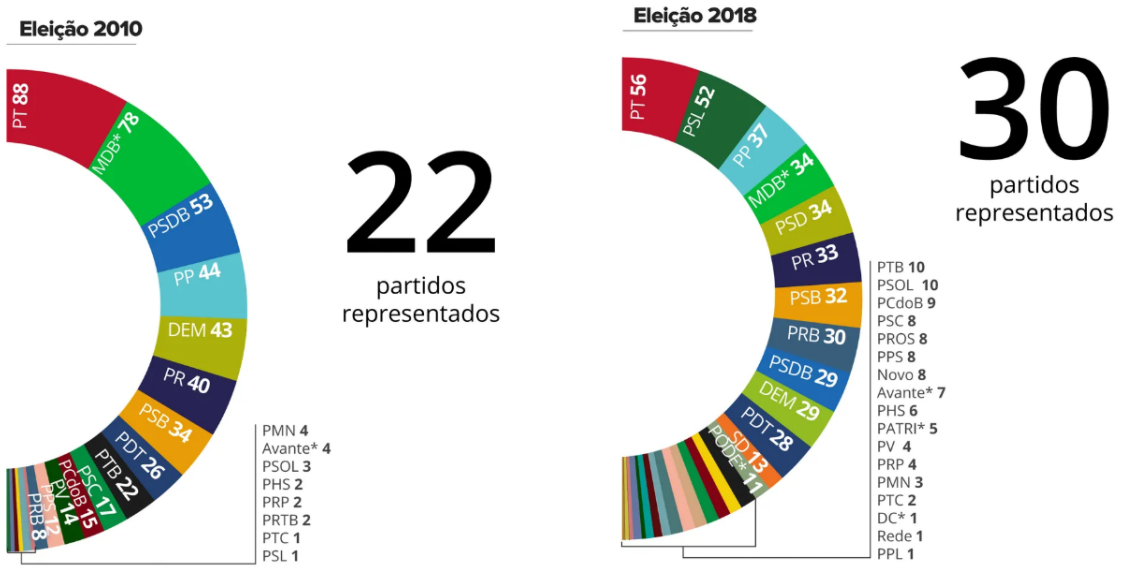
É possível também perceber a presença incessante de duas palavras, sendo elas presidente e ter. Por uma breve análise manual nos dados, é possível perceber que tais palavras aparecem por dois principais motivos. Presidente é uma palavra com altíssima recorrência muito pelo fato de que na maioria dos discursos a palavra é concedida ou dirigida ao presidente da sessão, ou mesa, ou seja, é usado como um interlocutor do discurso. Isto não traz, porém, um valor em sentido de tópicos para o texto, tal palavra então pode ser retirada. Já a palavra ter acontece diversas vezes, pois ela e suas inflexões são extremamente recorrentes na fala brasileira e escrita brasileira e que não trará significado real aos tópicos existentes dentro do discurso.

Tal análise demonstrou durante o trabalho a necessidade da criação de uma lista de palavras altamente recorrentes.

Partidos	mediaPalavras	#PalavrasDiferentes	Mais usada	Quantidade
DEM	153.55	11211	brasileiro	813
PCdoB	154.29	9187	brasileiro	609
PDT	187.34	8759	estado	466
PEN	149.25	303	mulher	4
PMDB	162.14	18282	estado	2320
PMN	91.47	1327	votar	32
PP**	157.02	10782	grande	747
PPS	143.76	8336	brasileiro	504
PR	126.47	9199	estado	755
PRB	220.54	6534	estado	206
PROS	170.74	7293	grande	325
PRP	136.6	704	mulher	10
PSB	274.13	13580	trabalho	681
PSC	142.22	6177	brasileiro	298
PSD	157.62	12139	estado	1053
PSDB	188.73	14611	brasileiro	1508
PSDC	97.78	519	estado	9
PSOL	220.76	10788	público	468
PT	202.16	24280	grande	4017
PTB	176.47	10810	grande	641
PTdoB	162.89	2107	deficiência	44
PV	149.85	5516	grande	226
SD	169.54	9335	estado	593

Tabela 3 – Resultado da última iteração realizada de análise preliminar dos dados de 2014.

Após algumas iterações da Análise Preliminar, obteve-se a Tabela 3, a qual, como é possível ver, já possui diversas palavras como palavra mais presente. Outras análises já podem ser observadas neste caso.



(a) Resultado da eleição de 2010 na Câmara dos Deputados.

(b) Resultado da eleição de 2018 na Câmara dos Deputados

Figura 9 – Resultados das eleições de cada uma das respectivas legislaturas. Obtidos de <https://tinyurl.com/mr2hew25> em 02/01/2024.

Pode-se perceber, por exemplo, que há na maior parte dos textos um alto nível de conectividade, com poucos partidos possuindo menos de 100 palavras consideradas relevantes para a extração de tópicos, sendo que os partidos que possuem uma quantidade menor de palavras médias menor do que 100, são partidos que foram considerados inaptos para a extração de tópicos, em sua grande maioria, porém os partidos demonstram uma quantidade média de palavras por discurso acima de 150 palavras.

Como é possível observar a partir da figura 9a, os partidos com grandes quantidades de discursos são os partidos que possuem uma grande representatividade na Câmara, o que é esperado, havendo ainda uma grande diferença entre os partidos com maior representatividade, no caso o PT e o PMDB[1]. Sendo possível observar a partir destes dados que os parlamentares do PT fizeram quase duas vezes a quantidade de discursos feitos pelos parlamentares do PMDB, os partidos tendo menos de 15% de diferença entre a quantidade de parlamentares.

Outro fator que corrobora tal fato é a questão de a média de palavras por discurso ser maior em aproximadamente 40 palavras por discurso. Isto mostra que eles não só falam mais vezes quanto que os discursos são mais longos e conexos, enquanto grafos.

Observando a Tabela 4 é possível perceber que mesmo sem a retirada de todos os termos repetidos, existe uma redução significativa nos discursos dos deputados. É possível perceber também que a quantidade de discursos realizados não sofreu uma mudança muito significativa, houve, porém, uma difusão dos discursos nos partidos.

[1] Mudança posterior a 2014 para a sigla MDB.

Partidos	mediaPalavras	#PalavrasDiferentes	Mais usada	Quantidade
AVANTE	93.5	2454	presidente	114
CIDADANIA	156.18	5197	ter	294
DEM	194.18	6675	ter	287
MDB	117.46	7323	ter	610
NOVO	124.72	8265	ter	1000
PATRIOTA	93.3	1606	presidente	50
PCdoB	111.85	8248	ter	854
PDT	114.39	7824	ter	751
PL	153.38	13507	ter	1575
PODE	95.84	4291	presidente	269
PP	152.33	7892	ter	543
PROS	134.29	3386	ter	136
PSB	125.57	8680	ter	756
PSC	101.57	1888	presidente	58
PSD	136.52	7207	presidente	570
PSDB	143.69	5205	ter	265
PSL	109.65	5945	presidente	426
PSOL	105.31	7624	ter	916
PT	139.95	15168	ter	3272
PTB	252.84	2809	presidente	38
REDE	108.73	2691	indígena	173
REPUBLICANOS	141.84	5988	ter	326
SOLIDARIEDADE	158.94	1606	ter	33

Tabela 4 – Resultado da primeira iteração realizada de análise preliminar nos dados de 2022.

Outro fato analisado neste momento foi o surgimento de partidos com nomes diferenciados. Os quais com uma pesquisa curta, descobriu-se que são somente renomeações de partidos já existentes anteriormente ou junções de diferentes partidos já existentes. Acredita-se que tais mudanças ocorreram por dois principais motivos, sendo o primeiro a inclusão da cláusula de barreira[2], a qual forçou legendas menores a se unirem ou a legendas de maior porte em uma legenda conjunta. O segundo motivo é a troca de nomes por envolvimentos em escândalos que levariam a sigla do partido e seus deputados a problemas para eleição de representantes.

Observando então a Tabela 5 é possível perceber uma quantidade significativa de mudança de quantidade de discursos dos partidos mais representados da Figura 9b, na qual houve uma dissipação dos discursos em outros partidos, em especial, partidos que sofreram rápida ascensão, como o partido Patriota, que é uma troca de nome do Partido Ecológico Nacional (PEN), o qual saiu de 4 discursos realizados no ano de 2014 para 1000 no ano de 2022. Outro partido com uma ascensão que pode ser considerada meteórica é o

[2] <https://www.tse.jus.br/comunicacao/noticias/2023/Setembro/glossario-eleitoral-explica-o-que-e-clausula-de-barre

Partidos	mediaPalavras	#PalavrasDiferentes	Mais usada	Quantidade
AVANTE	84.95	2440	povo	114
CIDADANIA	142.08	5182	projeto	294
DEM	177.6	6659	hoje	287
MDB	106.5	7308	estado	610
NOVO	112.4	8249	projeto	1000
PATRIOTA	83.88	1592	momento	50
PCdoB	102.42	8232	brasileiro	854
PDT	102.86	7807	projeto	751
PL	138.49	13491	brasileiro	1575
PODE	86.77	4276	estado	269
PP	138.19	7876	brasileiro	543
PROS	121.23	3372	câncer	136
PSB	114.43	8664	brasileiro	756
PSC	91.14	1874	hoje	58
PSD	123.66	7191	estado	570
PSDB	131.70	5190	energia	265
PSL	99.42	5929	brasileiro	426
PSOL	95.94	7608	brasileiro	916
PT	127.37	15151	brasileiro	3272
PTB	236.42	2795	grande	38
REDE	100.77	2677	indígena	173
REPUBLICANOS	129.51	5972	pessoa	326
SOLIDARIEDADE	146.03	1591	lei	33

Tabela 5 – Resultado da última iteração realizada de análise preliminar nos dados de 2022.

PL, o qual é a troca de nome do Partido Republicano (PR), o qual dobrou seus discursos no período analisado.

Uma diferença significativa em comparação com os dados da Tabela 3 são os tamanhos de cada um dos discursos que mesmo para partidos com grande representatividade sofreram uma drástica redução. Isso mostra uma indicação de que como uma análise geral, os partidos estão preferindo discursos mais curtos em relação aos dados anteriormente levantados. Uma possível causa de tal mudança é o aumento da presença de redes sociais perante o público brasileiro e a mudança de estratégia dos partidos para divulgação de seus conteúdos e ideias.

É possível perceber comparando tanto a Figura 9 quanto as Tabelas 3 e 5 um derretimento em quantidade de deputados e discursos do partido PMDB, o qual teve seus discursos reduzidos para aproximadamente 26% de sua participação no ano de 2014.

Comparando também as Figuras 9a e 9b, é possível perceber um aumento na quantidade de partidos representados. Havendo também um derretimento dos partidos mais representados dentro da Câmara dos Deputados, o que pode ser mostrado por uma redução significativa do Partido dos Trabalhadores (PT) em um número total de

32 deputados, o qual continua o partido mais representado na Câmara dos Deputados. Havendo, por outro lado, um surgimento de partidos antes poucos representados, como já mostrado anteriormente pelos dados coletados dos discursos.

Com tamanha mudança, este estudo levanta a questão de se tais partidos, os quais perderam representação ou ganharam representação, mantiveram seus tópicos de discussão, ou se houve uma mudança significativa nos tópicos existentes.

4.3 Tópicos extraídos dos discursos

Para a extração de tópicos, foi utilizado o algoritmo PBG e para a realização da extração de tópicos foram escolhidos alguns partidos muito importantes neste período da política brasileira, o Partido dos Trabalhadores, pelo fato de que ganhou a eleição para presidente nos dois anos analisados.

Outro partido analisado foi o PMDB ou MDB, o qual foi um partido que sofreu uma grande derrocada no número de deputados representados entre os dois períodos analisados. Tendo um de seus principais representantes, Michel Temer, eleito ao cargo de vice-presidente da república no ano de 2014 e mais tarde tornado-se presidente interino do Brasil devido ao impeachment da ex-presidente Dilma Rousseff.

Um terceiro partido analisado foi o Partido da República (PR), o qual mudou de nome para Partido Liberal(PL), pois durante o ano de 2022 o partido era a filiação do então Presidente da República Jair Bolsonaro. O que levou diversos apoiadores do presidente dentro da Câmara a migrarem para tal partido. Outro fator que traz um apelo para a análise do partido, ele já foi considerado um partido de base para os governos Dilma Rousseff, o que demonstra uma mudança significativa de alinhamento entre as legislaturas e uma adaptação contínua ao cenário político brasileiro. Tendo sofrido, portanto uma grande mudança em sua base na Câmara dos Deputados no decorrer tempo.

Realizou-se então a extração com uma quantidade total de 10 tópicos e com 10 termos por tópico, pois com a realização de alguns experimentos, percebeu-se que se utilizados mais tópicos ou mais palavras, havia para o primeiro caso, uma separação de tópicos, para o segundo caso havia a possibilidade de palavras mais comuns e que não agregariam aos tópicos existentes.

Durante a extração dos discursos foi possível analisar uma grande diferença de tempo necessário para a computação dos tópicos de 2014 e 2022, sendo os primeiros muito mais demorados. O que nos indicia ainda mais que tais dados são mais densos do que os dados presentes em 2022.

Durante os experimentos de extração, foi percebido que há a existência de pelo menos um tópico que concentra palavras gerais dos discursos, como palavra, abrir, dispensar,

conceder, corretor, destinar. As quais são muito ligadas a estrutura dos discursos utilizados pelos diferentes partidos, como o fato de se conceder a palavra, abrir os trabalhos ou o discurso, entre outras coisas.

4.4 Análise dos tópicos existentes

Partindo então para a análise dos tópicos extraídos, é primeiramente necessário recapitular alguns eventos que devem aparecer com uma constância dentro dos tópicos em cada um dos anos. Para 2014 é necessário lembrar que neste foram lembrados os 50 anos do início da ditadura militar no Brasil, a qual começou no dia 1º de abril de 1964. Outro fato importante é que no ano de 2014 o Brasil sediou a Copa do Mundo de Futebol da FIFA, um evento que trouxe para o Brasil uma visibilidade enorme em âmbito nacional. Um último fator importante para ser lembrado é que no ano de 2014 estava acontecendo na África Ocidental, um surto de Ebola.

Observando o ano de 2022, por sua vez, alguns fatos podem ser levantados como o fim do estado de pandemia de COVID-19 e retorno das atividades presenciais ao redor do país. Outro fator importante para a análise do ano de 2022 foi o levantamento de dúvidas sobre o sistema eleitoral brasileiro por parte da população de modo a questionar a sua segurança. Ademais, houve a privatização da Eletrobras e diversas polêmicas envolvendo o Supremo Tribunal Federal.

4.4.1 Aplicação de métodos de semelhança

Os métodos de semelhança aplicados retornam para o desenvolvedor duas informações, os tópicos mais semelhantes e o valor da similaridade entre eles. O método utilizado utiliza os tópicos de 2014 como base para extração da similaridade, ou seja, revela se há no ano de 2022 algum tópico semelhante aos obtidos em 2014. Em outras palavras, ela diz quais foram os tópicos mantidos entre os dois períodos.

Durante os experimentos percebeu-se que valores acima de 0.085 de similaridade possuem uma maior probabilidade de retornar tópicos semelhantes, embora ainda seja necessária certa análise manual. Valores abaixo disso tendem a retornar qualquer tópico que possua uma quantidade maior de palavras de uso geral.

Os resultados obtidos podem então sofrer uma análise manual para encontrar se os partidos realizaram ou não uma mudança grande em seus discursos.

4.4.2 Análise Manual

A análise manual, por sua vez, é o processo necessário e manual para extração de significado dos tópicos existentes, uma vez que é necessária a interpretação humana sobre

os conjuntos de palavras.

Um tópico que permeia os 3 partidos em ambas as legislaturas é um tópico de agradecimento, o qual é, em geral, composto por palavras como obrigado, agradecimento, obra, região, amigo, dentre outras. Tal tópico permeia grande parte dos discursos e fala sobre o agradecimento dos deputados perante a destinação de verbas ou realização de obras por outras pessoas da vida pública, como prefeitos e governadores. Tal tópico permeia grande parte dos discursos pelo fato de que, em geral, os deputados agradecem alguma pessoa e então encaminham para o tema principal do discurso.

4.4.3 Partido dos Trabalhadores

Há nos tópicos dos Partidos dos Trabalhadores alguns tópicos bastante interessantes de serem analisados.

Um dos tópicos interessantes é o tópico relacionado à reforma agrária, a qual pode ser vista em um tópico contendo as palavras: terra, familiar, agrário social, agricultura, reforma, agricultor, decreto, movimento e participação. Tais palavras podem ser vistas como um pedido pela reforma agrária, a qual é um anseio da esquerda brasileira há diversas décadas. Percebe-se, portanto, uma preocupação com os agricultores brasileiros e seus anseios, em especial, os pequenos agricultores. Os quais, segundo uma breve análise, seriam bastante beneficiados com uma reforma agrária. Pode-se dizer que tal tópico trata, em certa medida, sobre a questão socioambiental. O método utilizado encontrou uma similaridade para tal tópico, tendo, porém, o tópico sofrido uma mudança significativa em relação ao subtema tratado, ainda se fala da questão socioambiental, porém, o foco em 2022 é muito mais voltado aos indígenas. Isto se dá, pois em 2022 houve a tentativa de dificultar a demarcação de terras indígenas entre 2021 e 2022[3] o que acabou reverberando para dentro dos tópicos petistas de 2022. Outro fator que aparece em tal tópico é a liberação de agrotóxicos durante o período da pandemia de COVID-19, anos nos quais mais foram liberados agrotóxicos na história do Brasil[4].

Uma possível conclusão é que houve uma perda de direitos antes considerados conquistados, como as terras indígenas, os quais precisaram ser novamente defendidos. Tirando de circulação discussões como a reforma agrária, a qual beneficiaria uma quantidade enorme de pessoas.

Outro tópico interessante de ser analisado é um tópico sobre privatização de empresas públicas, o qual é mantido por palavras como: empresa, imprensa, petróleo, refinaria, negócio e país, além da palavra tucano, a qual neste contexto representa as

[3] <https://g1.globo.com/politica/noticia/2021/05/29/reabilitado-projeto-que-dificulta-demarcacao-de-terra-indigena-
ghtml>

[4] <https://g1.globo.com/jornal-nacional/noticia/2021/01/27/ano-de-2020-bate-recorde-de-liberacao-de-agrotoxicos-e
ghtml>

pessoas associadas com o Partido da Social Democracia Brasileira (PSDB), os quais a época eram chamados de tucanos e opositores diretos do governo petista. Muito provavelmente ligado à privatização de subsidiárias da Petrobras, como a BR Distribuidora, que abriu ações ao mercado alguns anos depois, com uma breve análise é possível analisar que neste momento os petistas são contra a privatização de empresas estatais. Foi encontrada através da medida de semelhança uma semelhança com um tópico de 2022 que fala também sobre privatização, em especial de dois setores, o setor de energia e o setor de fármacos e saúde. É possível ver então claramente uma influência da discussão da privatização da Eletrobras nos discursos. Acredita-se que a existência da área da saúde se dê pelo fato da pandemia, então há uma defesa ainda maior da não privatização de tal área.

A conclusão a qual é possível obter de que tal análise é que houve uma manutenção dos tópicos contrários a privatização de empresas estatais. Havendo, porém, uma mudança de foco no setor a ser defendido da privatização.

Um terceiro tópico interessante de ser tratado neste trabalho é o tópico que vai tratar do processo eleitoral democrático e as necessárias reformas, ideia a qual pode ser sustentada por diversas palavras como político, eleição, povo, democracia, reforma, campanha, voto e eleitoral. Tal tópico aparece nos textos sobre o pretexto do sistema político brasileiro é antiquado em relação a seu modelo e são necessários novos meios de democracia direta. Os métodos de semelhança apontam que este tópico é parecido com um tópico em 2022 o qual fala sobre a ideia de que haveria uma ameaça a democracia através das urnas, a qual se mostrou infundada[5], descredibilizando tal informação, ao longo dos discursos, referindo-se ao tema inclusive como uma pauta utilizada para o discurso de ódio, tal informação é sustentada por palavras como: brasileiro, democracia, povo, eleitoral, crime, processo democrático e ódio.

Um possível desfecho para tal tópico e sua similaridade é que houve um grande retrocesso na discussão do processo democrático, uma vez que se estava discutindo temas como democracia direta para uma maior participação popular na democracia, para uma defesa do processo democrático de direito em um período de 2 legislaturas.

Como é possível ver na Figura 10 existe uma manutenção de diversas palavras nos tópicos, representadas em roxo. Tais termos acompanham, porém, palavras que possuem maior presença dentro dos tópicos e que aparecem mais em um ou outro ano de análise. Isso, como mostrado na análise anterior, indica que houve uma manutenção de tópicos semelhantes, porém com uma mudança significativa dentro do escopo abordado dentro do tópico, o que reforça o ponto levantado até o momento.

Pode-se concluir, portanto, que dentro dos tópicos do PT houve uma redução significativa em relação à qualidade das discussões, voltando-se a discussão sobre questões

[5] <https://www.cnnbrasil.com.br/politica/auditoria-do-tcu-diz-que-possibilidade-de-fraude-nas-eleicoes-de-2022-e-proxima->

Figura 10 – Nuvem de palavras combinada dos tópicos do Partido dos Trabalhadores. Termos em vermelho apareceram em maior quantidade nos tópicos de 2014. Termos em azul tiveram mais presença em 2022. Termos em roxo tiveram a mesma representatividade nos tópicos em ambas as legislaturas.

anteriormente consideradas básicas e que voltaram a ser questionadas dentro da discussão política.

4.4.4 Movimento Democrático Brasileiro

A análise para o Movimento Democrático Brasileiro (MDB) levanta algumas questões bastante interessantes, com análises muito diferentes da do Partido dos Trabalhadores. Tal partido, como dito, possui relevância, pois é o partido do então vice-presidente Michel Temer, o qual foi levantado após o impeachment de Dilma Rousseff ao cargo de presidente da República.

O primeiro tópico a ser analisado é um tópico do ano de 2014 que aglomera palavras de direcionamento dos discursos ao público, em especial a quem assiste à TV Câmara, o que é sustentado por palavras como nobre, telespectador, tv, instante, brasileiro, presente. Tal discurso é interessante, pois não há uma correspondência clara aos dados desse discurso, o que mostra uma clara mudança no modo de discursar na tribuna, voltando-se para o público geral. Acredita-se que tal mudança se dê muito pelo aumento da presença das redes sociais dentre a população brasileira. O que influenciou o modo de discursar na tribuna. Algo que provavelmente relaciona-se com a redução do tamanho dos discursos, os quais são em geral editados e dirigidos para redes sociais de vídeos rápidos.

Outro conjunto de tópicos interessantes de serem analisados são os tópicos do MDB que aglutinam palavras sobre distribuição de recursos e dívidas, o qual é composto por palavras como hospital, energia, dívida, setor, recurso e atendimento. Este tópico aglutina duas ideias, a distribuição de recursos para o setor de saúde ao redor do Brasil e uma crítica ao endividamento do setor de energia. O método de similaridade encontrou, porém,

uma possível correlação com um tópico sobre a disponibilização do tratamento de doenças raras, como a Atrofia Muscular Espinhal (AME), no Sistema Único de Saúde (SUS) o que é sustentado por palavras como doença, raro, tratamento, plano, público, cerca, saúde, doente. Tópico muito impulsionado pelo Deputado Valtenir Pereira, o qual possui caso de AME na família. Como é possível ver, existe uma certa correlação, porém, muito distante do tema tratado pelos parlamentares do MDB no ano de 2014.

Uma análise possível de ser feita neste caso é a questão de uma mudança de perspectiva dos deputados do partido, saindo de um lugar de discussão de tópicos referentes a uma grande parcela da população para tópicos mais anedóticos e pessoais dos deputados representantes do partido.

Algo curioso que acontece com os tópicos do MDB em 2022, porém, é a não existência de um tópico sobre privatizações, nem a favor, nem contra tais medidas, mesmo tendo sido discutida na Câmara dos Deputados a privatização da Eletrobras. Como visto com os tópicos e discursos do PT, não havendo uma citação sequer às palavras derivadas do verbo privatizar. Dado que o MDB é um partido de centro e que esteve por trás da maioria dos governos desde a redemocratização, e com a existência de um cenário incerto, preferiu manter-se neutro.

Figura 11 – Nuvem de palavras combinada dos tópicos do Movimento Democrático Brasileiro. Termos em vermelho apareceram em maior quantidade nos tópicos de 2014. Termos em azul tiveram mais presença em 2022. Termos em roxo tiveram a mesma representatividade nos tópicos em ambas as legislaturas.

Como é possível ver na Figura 11, existe uma separação de termos muito maior do que quando comparado a Figura 10, com muitos termos nas cores vermelho e azul, demonstrando uma mudança mais acentuada nos tópicos no período analisado neste trabalho. É possível perceber também uma presença de algumas palavras com enorme participação, como estado, grande e dia. Isso pode indicar a utilização desses termos como termos para alongar o discurso e gerar no público a percepção de uma fala longa e prolixa.

4.4.5 Partido Liberal

O Partido Liberal (PL) anteriormente denotado como Partido da República (PR) possui uma relevância por ser o partido no qual em 2022 estava filiado o então presidente da república Jair Bolsonaro e o partido pelo qual se candidatou naquele ano para reeleição ao cargo ocupado naquele momento. Acredita-se, portanto, que os tópicos de tal partido possuam certa significância ao cenário político do momento estudado.

Um conjunto de tópicos muito interessante é um par de tópicos os quais discorrem sobre os direitos da população, em 2014 tal tópico discorre sobre a necessidade de direitos por parte da população, o que é sustentado pelas palavras: direito, mulher, considerar, público, ensino, tratar, estratégico, educação, trabalho e por final a palavra ditadura. Esta última palavra aparece neste conjunto de tópicos por alguns motivos, sendo o principal deles a defesa dos direitos de liberdade de expressão, o qual foi suprimido da população brasileira durante o período da ditadura militar. O método de semelhança por sua vez encontrou uma similaridade com um tópico contendo palavras como: liberdade, democracia, constituição, colega, direito, ditadura e defender. Tal tópico por sua vez apresenta um olhar muito diferente do tópico anterior, defendendo neste caso a necessidade de uma intervenção contra o ministro Alexandre de Moraes, uma vez que este estaria rasgando a constituição e levando o Brasil a caminho de uma ditadura.

Isto demonstra, portanto, uma clara mudança de paradigma do PL em relação ao tópico democracia e direitos. Isso pode ser explicado pela entrada de diversos representantes da extrema-direita dentro do partido, como o próprio ex-presidente Jair Bolsonaro e seus filhos, alguns dos principais disseminadores das ideias de extrema-direita dentro da política brasileira.

Outro par de tópicos a ser analisado neste caso são, em 2014 um tópico que levanta questões sobre segurança em diversas áreas incluindo a violência do sistema prisional brasileiro e a alta taxa de retorno às prisões, algo que é sustentado por palavras como segurança, policial, saúde, público, falar, questão, sistema e violência. Os métodos de similaridade por sua vez, encontraram similaridade com um tema o qual possui palavras como: policial, falar, bandido, criminoso, segurança, ministro e crime. Há, porém, uma grande diferença entre os tópicos, pois com uma análise manual é possível perceber que o tema tratado seriam possíveis atos criminosos do ministro Alexandre de Moraes ou então de Lula, o qual é muitas vezes referido como ladrão e bandido.

Isto reforça o tópico supracitado sobre a democracia. Isto mostra, portanto, um afunilamento de diversos tópicos e redução das discussões públicas sobre temas relevantes para o bem comum da nação.

Outro fato interessante de se analisar é a repetição de um tópico para o direcionamento dos discursos a quem acompanha a TV Câmara, assim como realizado pelo MDB,

o qual some após 2 legislaturas. Isto reforça, portanto, a mudança no direcionamento dos discursos da televisão para um discurso voltado para as redes sociais.

Figura 12 – Nuvem de palavras combinada dos tópicos do Partido Liberal. Termos em vermelho apareceram em maior quantidade nos tópicos de 2014. Termos em azul tiveram mais presença em 2022. Termos em roxo tiveram a mesma representatividade nos tópicos em ambas as legislaturas.

Analisando, neste caso, a 12 é possível ver uma presença de alguns termos com grande aparição nos dois conjuntos de tópicos, com a presença de diversos termos menos frequentes em ambos os conjuntos de tópicos. É possível, porém, perceber uma abundância de termos bem pequenos de ambas as cores, isso pode indicar que os discursos do partido possuem uma abundância de tópicos sendo tratados em ambas as legislaturas. Isso indica que não há, dentro do Partido Liberal, um consenso de tópicos a serem discutidos. Indica também que houve, mesmo com grande quantidade de tópicos, uma mudança nos tópicos estabelecidos entre as legislaturas comparadas.

Pode-se concluir, portanto, que o PL sofreu uma drástica mudança durante o período analisado neste trabalho, o que resultou em uma mudança significativa de seus tópicos para a direita. Uma parte de tal mudança pode ser creditada ao ex-presidente Jair Bolsonaro e seus seguidores, os quais se filiaram ao PL em 2022.

5 Conclusão

Neste trabalho explorou-se um *pipeline* para a extração de tópicos de discursos de Deputados Federais. Através dos experimentos conduzidos, foi possível perceber diferentes mudanças entre os diferentes partidos analisados.

Os resultados obtidos indica que é possível a utilização do método PBG em conjunto com métodos de comparação de strings para realizar uma análise dos discursos dos deputados federais para a comparação na mudança ou manutenção dos tópicos discutidos por estes na Câmara dos Deputados.

A análise realizada levantou por sua vez, uma manutenção de tópicos com uma volta de tópicos anteriormente considerados conquistados para o Partido dos Trabalhadores, como, por exemplo, a volta da discussão sobre demarcação de terras indígenas e a não utilização de agrotóxicos, além da necessidade de uma volta à defesa da democracia brasileira.

As conclusões que se pode tirar por sua vez para o Partido Liberal é que tal partido sofreu uma grande mudança de perspectiva durante as 2 legislaturas, o que pode ser atribuído a entrada de diversos deputados ligados ao ex-Presidente Jair Bolsonaro.

Em relação a limitações do trabalho, percebeu-se que, devido ao uso de modelos treinados em dados genéricos, os vetores utilizados na comparação de tópicos não possuem necessariamente o significado esperado. Levando o *pipeline* a encontrar algumas relações que não foram frutíferas para a análise. Outra limitação do *pipeline* desenvolvido é a sua atual limitação de métodos, estando limitado ao momento para o uso dos algoritmos PBG para a extração de tópicos e o método de similaridade de cossenos para a comparação dos tópicos, o que traz um certo engessamento para o processo e suas possibilidades.

Em relação a trabalhos futuros, acredita-se que o *pipeline* aqui proposto possa ser ampliado em trabalhos futuros para diversos outros algoritmos de análise de textos, criando então um ambiente para a análise dos Dados Abertos da Câmara dos Deputados

Referências

ANTIQUEIRA, L. et al. Modelando textos como redes complexas. In: *Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana*. [S.l.: s.n.], 2005. p. 22–26. Citado 3 vezes nas páginas 13, 26 e 27.

ARRUDA, G. F. de; COSTA, L. da F.; RODRIGUES, F. A. A complex networks approach for data clustering. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 391, n. 23, p. 6174–6183, 2012. Citado na página 25.

BLEI, D. M. Probabilistic topic models. *Communications of the ACM*, ACM New York, NY, USA, v. 55, n. 4, p. 77–84, 2012. Citado 2 vezes nas páginas 19 e 25.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado 2 vezes nas páginas 19 e 25.

BUN, K. K.; ISHIZUKA, M. Topic extraction from news archive using tf* pdf algorithm. In: IEEE. *Proceedings of the Third International Conference on Web Information Systems Engineering, 2002. WISE 2002.* [S.l.], 2002. p. 73–82. Citado na página 25.

DAUMÉ, H. *A course in machine learning*. [S.l.]: Hal Daumé III, 2017. Citado na página 23.

DONG, R. et al. Topic extraction from online reviews for classification and recommendation. In: AAAI. *Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI 13), Beijing, China, 3-9 August 2013*. [S.l.], 2013. p. 1310–1316. Citado na página 25.

FALEIROS, T. d. P. *Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais*. Tese (Doutorado) — Universidade de São Paulo, 2016. Citado 3 vezes nas páginas 13, 25 e 28.

FALEIROS, T. d. P.; ROSSI, R. G.; LOPES, A. A. Optimizing the class information divergence for transductive classification of texts using propagation in bipartite graphs. *Pattern Recognition Letters*, v. 87, p. 127 – 138, 2017. Citado na página 19.

FALEIROS, T. D. P.; VALEJO, A.; LOPES, A. Unsupervised learning of textual pattern based on propagation in bipartite graph. *Intelligent data analysis*, 2019. Citado 2 vezes nas páginas 19 e 26.

FORTUNATO, S. Community detection in graphs. *Physics reports*, Elsevier, v. 486, n. 3-5, p. 75–174, 2010. Citado na página 25.

FUKS, M.; MARQUES, P. H. Polarização e contexto: medindo e explicando a polarização política no brasil. *Opinião Pública*, SciELO Brasil, v. 28, n. 3, p. 560–593, 2022. Citado na página 32.

HARTMANN, N. et al. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017. Citado na página 36.

JACCARD, P. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, v. 44, p. 223–270, 1908. Citado na página 29.

KOTSIANTIS, S. B. et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, Amsterdam, v. 160, n. 1, p. 3–24, 2007. Citado na página 23.

KUBAT, M. Neural networks: a comprehensive foundation by simon haykin, macmillan, 1994, isbn 0-02-352781-7. *The Knowledge Engineering Review*, Cambridge University Press, v. 13, n. 4, p. 409–412, 1999. Citado na página 23.

LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*, 1966. Citado na página 29.

MERRIS, R. *Graph theory*. [S.l.]: John Wiley & Sons, 2011. Citado na página 24.

MERROUNI, Z. A.; FRIKH, B.; OUHBI, B. Automatic keyphrase extraction: a survey and trends. *Journal of Intelligent Information Systems*, Springer, v. 54, n. 2, p. 391–424, 2020. Citado na página 33.

MITCHELL, T. *Machine Learning'. McGraw-Hill, New York*. [S.l.]: NY, 1997. Citado na página 23.

NEWMAN, M. E. Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, APS, v. 64, n. 1, p. 016131, 2001. Citado na página 24.

ROSSI, R. G. et al. Inductive model generation for text classification using a bipartite heterogeneous network. *Journal of Computer Science and Technology*, v. 29, n. 3, p. 361–375, 2014. Citado na página 25.

SANTOS, M. A. D. Campanha não oficial–a rede antipetista na eleição de 2014. *Revista Fronteiras–estudos midiáticos*, v. 19, n. 1, 2017. Citado na página 20.

SEIBT, T. et al. Verdade, mentira e fake news nos discursos de lula e bolsonaro: uma análise de sentidos a partir do twitter dos candidatos na campanha para presidente do brasil em 2022. *Mediapolis–Revista de Comunicação, Jornalismo e Espaço Público*, n. 17, p. 47–63, 2023. Citado na página 20.

VALEJO, A. et al. Community detection in bipartite network: A modified coarsening approach. In: *Communications in Computer and Information Science (CCIS)*. [S.l.]: Springer International Publishing, 2018. v. 795, p. 123–136. Citado na página 24.

VIJAYMEENA, M.; KAVITHA, K. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, v. 3, n. 2, p. 19–28, 2016. Citado na página 28.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citado na página 24.

WEISSTEIN, E. W. Bipartite graph. *https://mathworld. wolfram. com/*, Wolfram Research, Inc., 2002. Citado na página 24.

WINKLER, W. E. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. ERIC, 1990. Citado na página 29.

ZHA, H. et al. Bipartite graph partitioning and data clustering. In: *Proceedings of the tenth international conference on Information and knowledge management*. [S.l.: s.n.], 2001. p. 25–32. Citado na página 24.