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

Aplicação de Modelos de Tópicos em análises automatizadas de discursos de senadores brasileiros

Victor Landim Teixeirense Pinheiro

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. Thiago Faleiros

Brasília
2021

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

Aplicação de Modelos de Tópicos em análises automatizadas de discursos de senadores brasileiros

Victor Landim Teixeirense Pinheiro

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. Thiago Faleiros (Orientador)
CIC/UnB

Prof. Dr. Jan Mendonça Correa
CIC/UnB

Prof. Dr. Edison Ishikawa
CIC/UnB

Prof. Dr. João Gondim
Coordenador do Curso de Engenharia da Computação

Brasília, 12 de Agosto de 2021

Dedicatória

Dedico o presente trabalho à minha mãe Naice, a meu pai Paulo, a minhas irmãs Sofia e Isabel e à minha parceira Julia pelo valioso apoio incondicional.

Agradecimentos

Agradeço ao Professor Dr. Thiago de Paulo Faleiros e ao Professor Dr. Pedro Robson Pereira Neiva pela apresentação do problema e pelo atencioso acompanhamento dos trabalhos.

Agradeço, ainda, à Universidade de Brasília pela oportunidade de ingresso ao Ensino Superior e por propiciar o desenvolvimento desse estudo.

Resumo

O contexto atual da sociedade é marcado por um excesso de informações. Dessa forma, uma abordagem de análise automatizada pode facilmente explicitar padrões em grandes coleções de dados. Uma dessas abordagens é a Modelagem de Tópicos. Tal ferramenta consome grandes coleções de documentos e evidencia padrões na forma de *tópicos*, que são conjuntos de palavras que descrevem um campo semântico. Neste contexto, almeja-se obter padrões ao analisar os tópicos obtidos a partir da extensa base textual de discursos de senadores, disponibilizada pelo Senado Federal. Acredita-se que é possível correlacionar a evolução temporal dos tópicos dos discursos com eventos históricos, políticos e econômicos. Os resultados encontrados são comparados com projetos de leis, datas relevantes e artigos jornalísticos. Com isso, este trabalho pode promover transparência aos cidadãos em relação às informações obtidas dos discursos de seus parlamentares. Por fim, o trabalho pode ser estendido com a avaliação de outras propostas de implementação de Modelos de Tópicos mais modernas.

Palavras-chave: Discurso, Senador, LDA, Brasil, Tópico, Mineração de Texto

Abstract

It is understood that there is an overflow of information in our society. In this context, an automated approach to analysis can bring out patterns in large collections of data. One of those approaches is Topic Modeling. This tool typically outputs topics from collections of documents. Topics are a set of words that describe a clear semantic concept. In this regard, it is desired to extract patterns from the topics created from the large collection of Brazillian senator speeches, provided by the Federal Senate. The main hypothesis is that there is a correlation between the historical topic evolution and historical, political, and economic events. The results are matched with draft bills, relevant dates, and news articles. Thus, this work can contribute with transparency to Brazillian citizens regarding the patterns found in their politicians' speeches. Ultimately, this work can be extended with the evaluation of more modern implementations of Topic Models.

Keywords: Speech, Senator, Brazil, LDA, Topic Model, Text Mining

Sumário

1 Introdução **1**
1.1 Contextualização . . . 1
1.1.1 Mineração de texto . . . 1
1.1.2 Dados abertos . . . 2
1.2 Objetivo . . . 3
1.3 Organização do Trabalho . . . 4

2 Referencial Teórico **5**
2.1 Pré-processamento de Texto . . . 5
2.1.1 Limpeza de dados . . . 5
2.1.2 Tokenização . . . 6
2.1.3 *Stemming* . . . 6
2.1.4 Lematização . . . 7
2.1.5 Raspagem de dados Web . . . 7
2.2 Modelos de tópicos . . . 8
2.2.1 *Latent Dirichlet Allocation* . . . 8
2.2.2 Métricas de Avaliação . . . 11
2.2.3 Verossimilhança e perplexidade . . . 11
2.2.4 Avaliação semântica . . . 11
2.2.5 *Pointwise Mutual Information* (PMI) . . . 12
2.2.6 Escore de coerência . . . 13
2.2.7 *Normalized Pointwise Mutual Information* (NPMI) . . . 14

3 Metodologia **15**
3.1 Coleta de dados do Senado . . . 15
3.1.1 Raspagem de dados . . . 19
3.2 Pré-processamento . . . 20
3.3 Modelo LDA . . . 22
3.4 Análise anual . . . 26

4 Resultados **29**
4.1 Base de discursos 29
4.2 Análise temporal 30
4.3 Resultados positivos 32
4.4 Resultados negativos 35

5 Conclusão **37**
5.1 Trabalhos futuros 38

Referências **39**

Apêndice **40**

A Tópicos do modelo LDA para 65 tópicos **41**
A.0.1 Tópicos ricos: 41
A.0.2 Tópicos pobres: 44

B Gráficos de evolução temporal para tópicos ricos **46**

Lista de Figuras

3.1 Exemplo de resposta *JSON* para um discurso de senador.. 16
3.2 Página *web* de um discurso e seu código fonte *HTML*. 20
3.3 Etapas da coleta de dados. 21
3.4 Número de discursos após cada etapa de filtragem. 23
3.5 NPMI para k entre 5 e 650. 24
3.6 NPMI para k entre 40 e 100. 25
3.7 Exemplo de decomposição de tópicos antes e após filtragem de tópicos pobres. 26
3.8 Tópico 33 entre 1995 e 2019. 27

4.1 Contagem de discursos por ano entre 1995 e 2019. 30
4.2 Contribuição média de cada tópico. 31
4.3 Evolução do tópico 5 entre 1995 e 2019. 32
4.4 Evolução do tópico 27 entre 1995 e 2019. 33
4.5 Evolução do tópico 33 entre 1995 e 2019. 33
4.6 Evolução do tópico 41 entre 1995 e 2019. 34
4.7 Evolução do tópico 55 entre 1995 e 2019. 34
4.8 Evolução do tópico 63 entre 1995 e 2019. 35
4.9 Evolução do tópico 63 entre 1995 e 2019. 36

B.1 Evolução do tópico 0 entre 1995 e 2019. 46
B.2 Evolução do tópico 1 entre 1995 e 2019. 47
B.3 Evolução do tópico 2 entre 1995 e 2019. 47
B.4 Evolução do tópico 3 entre 1995 e 2019. 48
B.5 Evolução do tópico 4 entre 1995 e 2019. 48
B.6 Evolução do tópico 6 entre 1995 e 2019. 49
B.7 Evolução do tópico 8 entre 1995 e 2019. 49
B.8 Evolução do tópico 9 entre 1995 e 2019. 50
B.9 Evolução do tópico 12 entre 1995 e 2019. 50
B.10 Evolução do tópico 14 entre 1995 e 2019. 51

B.11 Evolução do tópico 16 entre 1995 e 2019. 51
B.12 Evolução do tópico 18 entre 1995 e 2019. 52
B.13 Evolução do tópico 19 entre 1995 e 2019. 52
B.14 Evolução do tópico 23 entre 1995 e 2019. 53
B.15 Evolução do tópico 25 entre 1995 e 2019. 53
B.16 Evolução do tópico 26 entre 1995 e 2019. 54
B.17 Evolução do tópico 29 entre 1995 e 2019. 54
B.18 Evolução do tópico 30 entre 1995 e 2019. 55
B.19 Evolução do tópico 32 entre 1995 e 2019. 55
B.20 Evolução do tópico 37 entre 1995 e 2019. 56
B.21 Evolução do tópico 39 entre 1995 e 2019. 56
B.22 Evolução do tópico 45 entre 1995 e 2019. 57
B.23 Evolução do tópico 46 entre 1995 e 2019. 57
B.24 Evolução do tópico 48 entre 1995 e 2019. 58
B.25 Evolução do tópico 50 entre 1995 e 2019. 58
B.26 Evolução do tópico 51 entre 1995 e 2019. 59
B.27 Evolução do tópico 53 entre 1995 e 2019. 59
B.28 Evolução do tópico 54 entre 1995 e 2019. 60
B.29 Evolução do tópico 56 entre 1995 e 2019. 60
B.30 Evolução do tópico 57 entre 1995 e 2019. 61
B.31 Evolução do tópico 58 entre 1995 e 2019. 61
B.32 Evolução do tópico 61 entre 1995 e 2019. 62
B.33 Evolução do tópico 62 entre 1995 e 2019. 62

Lista de Tabelas

3.1 Tipos de pronunciamento e a distribuição de discursos. 18
3.2 Modelo de dado de um pronunciamento. 19
3.3 Principais tópicos do discurso 370187. 26
3.4 5 discursos que mais abordam o tópico 33 no ano 2005. 28

4.1 20 palavras mais frequentes nos discursos antes e após pré-processamento. . 31

Capítulo 1

Introdução

1.1 Contextualização

A sociedade contemporânea está imersa em um contexto de dados, seja na forma visual, seja auditiva, seja textual. É raro encontrar alguma área da sociedade que ainda não tenha adotado o uso de computadores, digitalizado documentos ou armazenado dados em nuvem. Sabe-se que, diariamente, é produzido pela humanidade uma quantidade astronômica de dados. Com o objetivo de realmente se *entender* grandes coleções de documentos, uma abordagem manual torna-se inviável. Diante disso, surge o promissor campo de estudo da Mineração de Texto [1]. Seu principal objetivo é desenvolver técnicas para extrair padrões, características e informações ocultas de grandes coleções de dados, de forma automática [2].

1.1.1 Mineração de texto

A Mineração de Texto é a área que está interessada em extrair informação e entendimento a partir de dados textuais [2]. São aplicações da Mineração de Texto: categorização textual, extração de palavras chave e modelagem de tópicos. Um processo genérico dessa área tipicamente envolve coletar dados, estruturá-los, pré-processá-los, extrair seus padrões e, por fim, avaliar os resultados obtidos [3]. É precisamente esta a metodologia que será adotada por este trabalho.

Neste sentido, uma poderosa ferramenta utilizada para auxiliar no desafio em questão são os Modelos de Tópicos. Modelos de Tópicos têm destaque por que evidenciam características gerais de uma grande coleção de documentos, de forma simples e sem a necessidade da atuação humana [4]. O *output* desse tipo de modelo é um conjunto de **tópicos**, ou temas, que indicam os assuntos que os documentos abordam. Um tópico nada mais é do que um conjunto de palavras semanticamente relacionadas que, quando

unidas, fazem alusão a um tema específico. Ao considerar o tópico composto pelas seguintes palavras, por exemplo: *corrupcao, denuncia, investigar, fatos, envolvendo, ministerio, etica, investigacao, inquerito, dinheiro*, é claro observar que aborda-se o tema "corrupção" em um contexto político. Gigantes da tecnologia como o Google, por exemplo, tem sucesso em aplicar modelos de tópicos para extrair características semânticas e agrupar documentos relacionados de forma eficiente em seus motores de busca. Dessa forma, o principal objetivo dos modelos de tópicos é aplicar métodos matemáticos e estatísticos que permitem a revelação de padrões semânticos ocultos em um grupo de documentos [2].

O modelo de tópicos mais popular é o LDA ou *Latent Dirichlet Allocation* [5]. Ele foi proposto por Blei em 2003 e marcou o início da área de estudos de Modelos de Tópicos Probabilísticos. Essa área tinha como principal objetivo extrair informações de forma eficiente de bases textuais de documentos, utilizando abordagens probabilísticas [5, 6]. Diferentemente dos processos aplicados desde então, o LDA se utiliza de um processo iterativo e generativo para criar tópicos. De forma simples, esse algoritmo primeiro assume que um documento é criado por meio de amostragens probabilísticas e, com isso, realiza um processo de inferência para realizar a operação inversa e decompor o documento em uma distribuição de tópicos.

1.1.2 Dados abertos

Nos últimos tempos, deu-se início à uma iniciativa promovida pelas principais democracias da atualidade a disponibilizar dados abertos de forma livre e indiscriminada aos cidadãos de seus países. Sabe-se que tornar informações orçamentárias, governamentais e legislativas acessíveis é essencial para promover transparência e aumentar a efetividade e a *accountability* [1] do setor público [7]. Essa iniciativa, no Brasil, ganhou força com a Lei de Acesso à Informação, de 2011 [2]. Tal lei, além de visar tornar públicos e disponíveis dados relacionados à gestão pública, também busca garantir sua disponibilidade. Hoje, no ano de 2021, existem publicados, a nível federal, estadual e municipal, diversos portais de acesso à informação na internet. Ainda a nível federal, nota-se um esforço maior na promoção da transparência. Pode-se acessar portais relacionados à órgãos públicos, ministérios, agências e, em especial, à casas legislativas [3] [4] [5] [7].

Diante deste contexto de dados abertos, o Senado Federal brasileiro disponibiliza para seus cidadãos o Portal da Transparência [6]. Em seu *website*, torna acessíveis tanto in-

[1] Accountability pode ser entedida como prestação de contas, transparência e responsabilização.
[2] http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm
[3] https://www.congressonacional.leg.br/dados/docs/index.htm
[4] https://dadosabertos.camara.leg.br
[5] https://www12.senado.leg.br/dados-abertos
[6] https://www12.senado.leg.br/transparencia

formações administrativas quanto legislativas de forma simples. Além de prover dados relacionados à orçamentos, gastos, leis e indicadores, o senado também disponibiliza dados referentes ao histórico da atividade dos parlamentares brasileiros, principalmente senadores e deputados. No contexto deste trabalho, o destaque desse conjunto de dados são as transcrições dos discursos proferidos por senadores. Especificamente para a casa legislativa do Senado Federal, é disponibilizada uma *API* (*Application Programming Interface*) que pode ser consultada pelos cidadão que desejem visualizar as agendas dos parlamentares, seus blocos políticos e também seus discursos proferidos. [7]

A priori, a base de discursos tem registrados centenas de milhares de pronunciamentos desde 1995. É visível como o processamento manual desse tipo de documento pode se tornar impraticável, tanto pelo fator custo, quanto pelo fator tempo. O processo de sumarização, por exemplo, é feito atualmente por profissionais do Senado, que devem realizar esse processo custoso de forma manual.

Acredita-se que, utilizando técnicas de mineração de texto, como a modelagem probabilística por meio do *LDA*, pode-se, de forma automatizada, extrair diversas formas de inteligência sobre a base de dados textuais dos discursos. Pode-se, por exemplo, com essas técnicas, sumarizar os discursos à medida que são proferidos, criar classificadores que explicitem a inclinação política de um parlamentar a respeito de determinado tema e, ainda, é possível determinar automaticamente os principais assuntos abordados em um determinado discurso.

1.2 Objetivo

O objetivo principal do presente trabalho é mostrar que, com o auxílio de Modelos de Tópicos Probabilísticos, é possível identificar eventos históricos em discursos de senadores por meio de tópicos latentes.

Sabe-se que, em se tratando de bases de dados textuais, pesquisadores da área de Mineração de Texto oriundos de países lusófonos encontram dificuldades em obter material de qualidade na língua portuguesa, pois infelizmente estes ainda são escassos. Portanto, este trabalho possui como objetivo secundário contribuir com a criação de uma base de discursos normalizada, tratada e pré-processada que contém discursos de senadores proferidos entre os anos 1995 e 2019.

Por fim, acredita-se que esse trabalho, ao avaliar tais dados, pode contribuir para a promoção da transparência do Senado Federal, reforçando a democracia no país.

[7] Dados Abertos Senado

1.3 Organização do Trabalho

O Capítulo 2 apresenta uma análise da literatura e expõe os principais conceitos necessários para o entendimento deste trabalho. O Capítulo 3 descreve a metodologia utilizada no processo de coleta de dados, pré-processamento e transformação dos dados. No Capítulo 4 discute-se os resultados obtidos bem como suas métricas de avaliação. Por fim, no capítulo 5, apresenta-se a conclusão do trabalho, revisa-se os resultados obtidos e discute-se o direcionamento de trabalhos futuros.

Capítulo 2

Referencial Teórico

2.1 Pré-processamento de Texto

Antes de aplicar as técnicas de mineração de texto, é necessário processar e transformar os dados textuais para um formato que seja mais apropriado para o processamento automatizado por computadores [2]. Esse tipo de algoritmo tem a expectativa de receber como *input* dados puramente numéricos, ao invés de textuais, que têm como unidades básicas letras e símbolos. Dessa forma, a ideia do processamento de texto é utilizar técnicas para converter dados textuais em uma sequência de componentes linguísticos que seguem uma estrutura pré-determinada e que atenda às expectativas dos algoritmos [2]. Esta seção irá explorar a base teórica relacionada ao pré-processamento, "tokenização", "lematização" e *stemming*.

2.1.1 Limpeza de dados

Tarefas comuns realizadas antes de processar quaisquer bases de dados textuais incluem: corrigir dados inválidos, tratar dados ausentes, realizar filtragens e minimizar ruídos [2]. Podem existir várias razões que justifiquem os tipos de incoerência de dados citados. Os dispositivos utilizados nos processos de coleta podem apresentar falhas técnicas, pode haver falha humana no processo, e, ainda, é possível que ocorra problemas na transmissão dos dados coletados. Então, é denominado "Limpeza de Dados" o conjunto de rotinas que remove essas imperfeições e fornece uma base de dados padronizada e, idealmente, livre de incoerências. Esse processo é essencial por que confere maior confiabilidade aos resultados obtidos nas etapas posteriores de processamento [1]. Em geral, na maioria das aplicações é comum aplicar as seguintes etapas de limpeza, que estão largamente disponíveis em bibliotecas de código [2]:

Limpeza de texto: De maneira abrangente, envolve a remoção e filtragem de ruído presente em *tags HTML*, *XML* e *JSON*. Este ruído pode ser, ainda, encontrado na forma de cabeçalhos ou rodapés que estão presentes em muitos dos documentos da coleção, conhecidos como *boilerplate*. Esse tipo de ruído é frequentemente tratado por meio do uso de expressões regulares.

Remoção de caracteres especiais: Quando o objetivo é se trabalhar apenas com palavras, então números, acentos, pontuação e caracteres especiais são descartados.

Padronização de caixas altas e caixas baixas: No contexto geral da Mineração de Texto, as palavras "Senado", "senado" e "SENADO" possuem o mesmo valor semântico e, portanto, devem ser avaliadas da mesma forma no contexto da análise. Então, todo o *corpus* (a base textual completa) é convertido para caixa alta ou caixa baixa para que um padrão possa ser estabelecido.

Remoção de *stopwords*: Define-se *stopwords* como palavras frequentes em um texto que não possuem valor semântico ou que devem ser desconsideradas por qualquer motivo, a depender da aplicação. Tipicamente são artigos, preposições e conectivos como "a", "de", "para". É comum encontrar em bibliotecas conjuntos de *stopwords* empacotados e prontos para uso em uma variada gama de idiomas, incluindo o português.

2.1.2 Tokenização

No contexto de Mineração de Texto, define-se *token* como a unidade textual mínima e independente que possui sua própria sintaxe e semântica [2]. Em um texto, pode-se optar por tokenizá-lo em frases que podem ser tokenizadas ainda em um conjunto de palavras. Assim, o processo de tokenização consistem em dividir um grande conjunto de dados textuais em componentes menores e contidos chamados de *token*. [2] Em geral, principalmente no contexto da língua portuguesa, constuma-se considerar o caractere "espaço" como separador que será considerado no critério de tokenização de palavras [2].

2.1.3 *Stemming*

O *stem* de uma palavra é o recorte da sua forma não flexionada, também chamada de forma base [2]. Por exemplo, as palavras "denúncia", "denunciar" e "denunciando" possuem o mesmo *stem*, que é "denunci". O entendimento é que palavras com o mesmo *stem* possuem o mesmo valor semântico e, portanto, devem ser consideradas da mesma forma. O principal objetivo do *stemming* é reduzir o tamanho do vocabulário.

2.1.4 Lematização

O processo de lematização é muito semelhante ao processo de *stemming*. Nos dois processos as palavras são reduzidas para sua forma raiz, mas a principal diferença é que, no caso da lematização, a palavra sempre é correta, isto é, é encontrada no dicionário, ao passo que o resultado do *stemming* frequentemente não é [2]. O elemento raiz deste processo é conhecido como *lemma* e, para o exemplo anterior, este pode ser "denúncia" ou "denunciar", a depender da implementação. O processo de lematização nem sempre é viável, pois seus algoritmos são mais complexos e dependem de contextos sintáticos e linguísticos específicos. Por isso, eles demandam um maior tempo de execução. Além disso, para idiomas pouco populares na literatura, como o português, suas implementações ainda não possuem desempenho satisfatório.

2.1.5 Raspagem de dados Web

A rede mundial de computadores é uma grande biblioteca digital onde pode se encontrar os mais variados conjuntos de informação, alcançando áreas como notícias, finanças, medicina, educação e entretenimento. Por ser diversa e extensa, a *web* é uma rica fonte de dados para a mineração de dados [1]. Contudo, existem desafios a serem superados para permitir uma correta extração desse conhecimento. Primeiro, páginas *web* são complexas por natureza, isto é, não possuem um padrão definido. Além disso, diferem de um banco de dados tradicional por que não é possível realizar buscas por índices como título, autor ou sumário [1]. Em geral, recorre-se ao auxílio dos motores de busca, que retornam como resultado páginas que contém palavras-chave da busca.

A estrutura básica de uma página *web* é conhecida como *DOM (Document Object Model)* ou Modelo de Objeto de Documento. O *DOM* é representado internamente como uma estrutura de árvore, onde cada elemento da página é um nó. Esses elementos são chamados de *tags HTML* e alguns exemplos são: `<p></p>` representam parágrafos, `<table></table>` representam tabelas e `<h1></h1>` representam títulos [1].

Uma vez que se conhece a estrutura *DOM* da página cujos dados tem-se interesse em extrair, a tarefa de coleta é trivial. É possível utilizar ferramentas automatizadas como a biblioteca BeautifulSoup4 [1], disponível para a linguagem Python. O desafio dessa tarefa está na correta seleção do elemento ou grupo de elementos do *DOM* que contém os dados textuais. Com a seleção feita, a biblioteca retorna o almejado dado textual puro.

[1] Referência BeautifulSoup4

2.2 Modelos de tópicos

No que tange o contexto textual dos dados, uma abordagem popular ao problema de processar eficientemente grandes coleções de documentos é a abordagem probabilística. Pesquisadores da área de aprendizado de máquina desenvolveram um conjunto de algoritmos capazes de anotar grandes bases de dados com informação temática [8]. Dessa forma, surge a área de modelos de tópicos probabilísticos. Essa área simplificou substancialmente o entendimento de grandes bases de dados textuais [6]. Os algoritmos dessa área identificam relações estatísticas entre as palavras dos documentos e, com isso, mostram que tema cada um aborda e quais as relações entre os documentos. O objetivo principal desse tipo de abordagem é relacionar os documentos por meio de **tópicos**, que em geral são palavras com valor semântico capaz de relacionar diferentes documentos dentro de um mesmo assunto. É relevante notar que os tópicos surgem apenas por relações estatísticas entre palavras, ou seja, não há a necessidade de rótulos prévios ou intervenção humana [8].

2.2.1 *Latent Dirichlet Allocation*

Na área de modelos probabilísticos de tópicos, o algorítmo LDA (Latent Dirichlet Allocation) proposto em 2003 por Blei é tido como padrão e é a base das soluções seguintes desenvolvidas na área. LDA é um algoritmo generativo que assume que um documento é formado por uma mistura oculta de tópicos. Os tópicos, por sua vez, são descritos como distribuições de palavras de um vocabulário fixo. Neste contexto, o algoritmo entende que os tópicos são criados antes mesmo de qualquer documento, não o contrário. Assim, o LDA assume que os documentos são criados por meio de um processo generativo e imaginário que descreve etapas em que são selecionados os tópicos que compõem os documentos e as palavras que compõem cada tópico. Algoritmos de modelos de tópicos probabilísticos realizam análises que avaliam a distribuição probabilística combinada das variáveis observáveis e das variáveis ocultas envolvidas. Em geral, o objetivo é computar a distribuição condicional das variáveis ocultas *dadas* as variáveis observáveis. Neste contexto, as variáveis observáveis do modelo são o conjunto de palavras de um documento, ao passo que as variáveis ocultas são as efetivas distribuições dos tópicos. Ademais, o modelo é alimentado com variáveis que influenciam na granularidade das distribuições. Essas variáveis são conhecidas como hiper-parâmetros. O hiper-parâmetro α ajusta a granularidade da distribuição de tópicos dos documentos, de forma a definir se um documento é formado por um conjunto maior ou menor de tópicos. De forma análoga, o hyper-parâmetro β é responsável pela granularidade da distribuição de palavras em cada tópico [8, 6].

Faz-se necessárias as seguintes definições:

Palavra: Também chamado de *token*. É uma unidade discreta de dado, um item do vocabulário.

Documento: Sequência de N palavras denotada por $w = (w_1, w_2, ..., w_N)$.

Corpus: Coleção de um total de documentos de interesse, denotada por $D = \{d_1, d_2, ..., d_M\}$.

Vocabulário: Também chamado de dicionário. É o conjunto das palavras únicas do corpus.

O processo generativo relacionado a um documento d_j do corpus D é detalhado da seguinte forma:

1. Para cada um dos K tópicos (definido pelo modelador), crie as **distribuições de palavras**, definidas como distribuições de dirichlet $\phi_k \sim Dir(\phi_k, \beta)$, que seleciona as palavras que compõe cada um dos K tópicos.
2. Crie a **distribuição de tópicos**, definida como uma distribuição de dirichlet $\theta_j \sim Dir(\theta, \alpha)$ referente ao documento d_j. Define a composição de tópicos do documento.
3. Em seguida, para cada palavra w_i do documento d_j:
   (a) Amostre um tópico aleatório $z_{i,j}$ da distribuição de tópicos θ_j do documento d_j.
   (b) Selecione uma palavra aleatória $w_{j,i}$ dada a probabilidade $p(w_{j,i}|\phi_{z,j,i})$.

Como foi apresentado, o processo gerador é guiado por duas variáveis que descrevem as distribuições do modelo, θ e ϕ, com os respectivos hiper-parâmetros α e β . θ é a distribuição de tópicos que define que tópicos fazem parte da composição de um documento d_j. Dessa forma, possui dimensionalidade K, o número total de tópicos fixo e pré-definido. Analogamente, ϕ define a distribuição de palavras de um tópico k. Portanto, possui dimensionalidade n, o tamanho do vocabulário [6].

Ademais, uma das características diferenciais do LDA é que, apesar de todos os documentos do *corpus* compartilharem os mesmos K tópicos, cada documento os apresenta com diferentes proporções, seguindo sua própria distribuição de tópicos θ_j [8].

Com base no entendimento matemático exposto, retorna-se a atenção ao objetivo básico da modelagem de tópicos, que é extrair tópicos de uma coleção de documentos de forma automatizada. Neste problema, apenas as palavras dos documentos são entidades conhecidas. As palavras que compõem os tópicos e a composição de tópicos de cada

documento são elementos ocultos. O desafio computacional dos modelos de tópicos probabilísticos é realizar a *operação inversa* da etapa generativa. Ou seja, dados os documentos, almeja-se *inferir* os elementos latentes supracitados. Blei propõe a pergunta: qual é a estrutura oculta que provavelmente gerou a coleção de documentos observada [8]? Ainda, modelos de tópicos exibem a curiosa propriedade de que a estrutura latente que gera uma coleção de documentos se relaciona com a estrutura temática e semântica da coleção.

A notação matemática acima nos permite expressar a probabilidade conjunta entre as variáveis não observadas e as variáveis observadas da seguinte maneira:

$$p(z, w, \phi, \theta|\alpha, \beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \left(\prod_{i=1}^{V} p(z_{j,i}|\theta_j) p(w_{i,j}|z_{i,j}, \phi_{z,j,i}) \right) \tag{2.1}$$

A equação acima expressa diversas dependências complexas que, em si, definem o LDA. É desejado saber as distribuições de tópicos para os documentos e as distribuições de palavras para os tópicos. Porém, na prática, além dos hiper-parâmetros α e β, apenas as palavras w dos documentos são conhecidas. Ajustando a equação para isolar as variáveis à priori, tem-se que:

$$p(z, \phi, \theta|w, \alpha, \beta) = \frac{p(z, w, \phi, \theta|\alpha, \beta)}{p(w)} \tag{2.2}$$

A equação acima, chamada de cálculo da *posteriori* do LDA, computa a estrutura de tópicos dados os documentos observados. A princípio, esse problema computacional pode ser resolvido pela soma das distribuições conjuntas sobre todas as variáveis não observáveis [8]. Contudo, o número de atribuições cresce exponencialmente, de forma a tornar o problema intratável computacionalmente. Então, a área de tópicos probabilísticos propõe uma série de métodos eficientes para aproximar distribuições à posteriori como esta. Em geral, considera-se duas categorias de algoritmos de aproximação: algoritmos baseados em amostragem e algoritmos variacionais [8].

Algoritmos baseados em amostragem, como sugere o nome, obtém amostras da posteriori e tentam aproximá-la por meio de uma distribuição *empírica*. O principal exemplo é o Amostrador de Gibbs, que tem como base uma cadeia de Markov (longa cadeia de variáveis aleatórias e dependentes da variável anterior, cuja distribuição limitante é a própria posteriori). A cadeia de Markov, definida por meio das variáveis não observáveis, roda de forma iterativa, coletando amostras da posteriori e aproximando a distribuição por meio das amostras coletadas [8].

Métodos variacionais, em contraste aos métodos baseados em amostragem, propõem uma abordagem determinística, ao invés de uma abordagem empírica. Métodos variacionais, com base na estrutura de tópicos oculta, se utilizam de um conjunto de distribuições

parametrizadas que indicam o membro desse conjunto que mais se aproxima da posteriori em questão. Com isso, o problema computacional é visto como um desafio de otimização [8].

2.2.2 Métricas de Avaliação

2.2.3 Verossimilhança e perplexidade

A principal métrica utilizada para avaliar a qualidade de um modelo probabilístico de tópicos é a verossimilhança *held-out* do modelo [9]. Neste contexto, a questão principal é estimar o quão bem o modelo pode descrever corretamente documentos nunca vistos ou *held-out* (retidos). A metodologia adotada, em geral, é: organizar os documentos em dois grupos, documentos de *treino* e documentos de *teste*, tipicamente correspondendo à proporções de 90 e 10 por cento, respectivamente [5]. Então, prossegue-se criando o modelo com os documentos de treino. O objetivo, então, é encontrar a *probabilidade de documentos retidos*, que indica a competência do modelo em caracterizar documentos inéditos. Em geral, o valor da métrica que costuma ser avaliado é o logaritmo dessa verossimilhança de documentos retidos.

Neste contexto, criou-se o conceito de *perplexidade*. A métrica da perplexidade, matematicamente relacionada à verossimilhança *held-out*, indica o quanto um modelo se desvia do conjunto de treino em relação ao conjunto de teste [6] [5]. Um valor de perplexidade inferior sugere um modelo capaz criar generalizações de forma mais eficiente. Em um cenário de modelagem, a perplexidade se mostra muito útil pois permite que se compare modelos semelhantes, mas com número tópicos K distintos. Dessa forma, pode-se usar a perplexidade como critério de seleção para o K ideal [10]. A perplexidade é matematicamente descrita como:

$$perplexidade(x) = \exp\left(-\frac{\log p(w|\alpha,\beta)}{\log \sum_{j=1}^{m} n_{d_j}}\right) \tag{2.3}$$

Em que $\log p(w|\alpha,\beta)$ é o logaritmo da verossimilhança de documentos retidos do modelo.

2.2.4 Avaliação semântica

Porém, a avaliação de modelos de tópicos por meio de métricas como logaritmo da verossimilhança ou perplexidade ignora a representação interna dos modelos [11]. Ou seja, estas métricas puramente estatísticas mostram-se pouco úteis para identificar a real coerência semântica nos tópicos gerados pelos modelos. Com isso, Chang et al. [11] rompem com a forma com que modelos de tópicos eram avaliados e propõe um método novo, que

busca medir o quão interpretáveis são os tópicos, utilizando o julgamento humano. Neste trabalho, são propostos dois métodos para se avaliar tanto a qualidade da distribuição de palavras dos tópicos ϕ_k quanto a distribuição de tópicos dos documentos θ_j do modelo.

Chama-se de Intrusão de Palavras o método que busca avaliar se um humano é capaz de identificar coerência semântica em um determinado tópico. Ao voluntário é apresentado um conjunto de palavras que compõem um suposto tópico, porém uma dessas palavras é uma "intrusa". Ou seja, é uma palavra alheia ao tópico que foi adicionada com o objetivo de quebrar a relação de coerência do tópico. Se um tópico é coerente, então a palavra intrusa é facilmente evidenciada [11].

Por outro lado, o método da Intrusão de Tópicos tem como objetivo verificar se a decomposição dos documentos em tópicos pelo modelo corresponde às expectativas humanas. É apresentado aos avaliadores um trecho de um documento bem como seus quatro tópicos de maior probabilidade. Além dos quatro tópicos, é introduzido, novamente, um tópico "intruso", selecionado aleatoriamente de um dos tópicos de baixa probabilidade do modelo. Se a atribuição de tópicos do modelo é coerente e relevante, espera-se que os voluntários humanos sejam capazes de indicar facilmente os tópicos intrusos [11].

A relevância deste trabalho foi mostrar que as métricas de avaliação tradicionais são pouco eficientes em determinar a real coerência de tópicos criados de modelos probabilísticos. Além disso, este trabalho expôs a conclusão contra-intuitiva de que os propostos métodos de avaliação de coerência semântica apresentam uma correlação **negativa** em relação à métricas probabilísticas como a perplexidade [11]. Isso significa que tentativas de otimização dos valores de perplexidade podem não gerar tópicos coerentes ou interpretáveis por humanos.

2.2.5 *Pointwise Mutual Information* (PMI)

Newman et al. [12] apresentou em seu trabalho a métrica PMI (*Pointwise mutual information*), que calcula um escore baseado na presença de pares de palavras dos tópicos em grandes bases de dados externas como Wikipedia, Google e WorldNet. A metodologia proposta foi capaz de identificar tópicos pobres facilmente e, em geral, apresentou concordância com o julgamento humano.

O algoritmo do PMI considera as dez palavras mais significativas de cada tópico e, selecionando as palavras aos pares, calcula as co-ocorrências de um par $\{w_i, w_j\}$ de palavras na base da Wikipedia e do Google n-grams. O PMI define uma métrica de associação entre palavras. Para um dado tópico $w = (w_i, ..., w_{10})$, o escore é calculado como a mediana dos valores de PMI para cada par do tópico:

$$PMI - Score(w) = mediana\{PMI(w_i, w_j), ij \in 1...10\} \tag{2.4}$$

Em que o PMI de um par é calculado como:

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{2.5}$$

Então, um par de palavras semanticamente próximas apresenta um alto valor de PMI. Por exemplo: $PMI(rock, banda) = 4.5$ [12]. O autor expõe que foi observada um correlação entre a análise humana e o PMI em um intevalo de $\rho = 0.70...0.78$, considerada ótima.

2.2.6 Escore de coerência

Mimmo et al. [13] usa um princípio semelhante ao PMI, mas substitui a métrica anterior pelo cálculo do logaritmo da probabilidade condicional de pares de palavras. Este algoritmo propõe-se a criar uma métrica de coerência capaz de diagnosticar problemas em modelos de tópicos, sem a necessidade de atuação humana ou uso de bases de dados externas [13]. Com o auxílio da análise de especialistas da área, identificou-se os principais tipos de deficiência existentes em conjunto de tópicos gerados. Definiu-se 4 categorias para um tópico pobre:

Encadeado: Neste tipo de tópico, as palavras se conectam por uma estrutura de cadeia, em que as palavras fazem sentido apenas aos pares, mas não em conjunto.

Intruso: Tópicos intrusos podem ser formados por dois conjuntos de palavras relacionados separadamente ou por um tópico considerado bom, mas que contém algumas palavras "intrusas".

Aleatório: Há ausência total de sentido ou conexão semântica entre mais que alguns pares de palavras do tópico.

Desbalanceado: Um tópico desbalanceado apresenta um conjunto de palavras que fazem sentido juntas, mas combinam termos específicos e genéricos.

Então, observou-se que é possível identificar tópicos de baixa qualidade que pertencem a três das quatro categorias apresentadas de forma automatizada baseada em coocorrência de pares de palavras no corpus de interesse.

A *coerência de tópico* é calculada como:

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)})}{D(v_l^{(t)})} \tag{2.6}$$

Em que $D(v)$ é a frequência de palavras v no documento, $D(v, v')$ é a frequência de co-documentos dos pares de palavras v e v'. $V^{(t)} = (v_1^{(t)}, ..., v_M^{(t)})$ é a lista de palavras mais prováveis de um tópico t.

Os resultados apresentados foram muito promissores, frequentemente alinhados ao julgamento de especialistas e superando a performance do PMI. Dessa forma, Mimno et al. comprovou que toda a informação de co-ocorrência já está no corpus (não depende de bases externas) e que um corpus *held-out* também não é necessário [13].

2.2.7 *Normalized Pointwise Mutual Information* (NPMI)

Bouma et al., por sua vez, dá continuidade aos esforços empregados na criação do PMI e propõe uma alternativa normalizada [14]. A utilidade da normalização é que a escala de valores possui interpretação fixa ($NPMI \in [-1, 1]$), de forma que pode ser um bom candidato a substituir o PMI. Além disso, a normalização acaba com a tendência do PMI em favorecer palavras de baixa frequência [15].

Considerando as N palavras mais prováveis de um tópico t, define-se NPMI como:

$$NPMI(t) = \sum_t j = 2^N \sum_{i=1}^{j-1} \frac{\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-\log p(w_i, w_j)} \tag{2.7}$$

O NPMI possui interpretação simples. Quando duas palavras aparecem *apenas* juntas, então $NPMI(w_i, w_j) = 1$. Se as palavras são distribuídas de forma independente, então $NPMI(w_i, w_j) = 0$. Por fim, quando as palavras são observadas separadamente, mas nunca em conjunto, tem-se que $NPMI(w_i, w_j) = -1$ [14].

De fato, Lau et al. [15], em seu trabalho *Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality*, compara diversas métricas de avaliaçao de modelos de tópicos e indica NPMI como a mais performática em todos os cenários. Em termos de desempenho, o NPMI esteve levemente abaixo do resultado da avaliação humana, e se igualou a ela em alguns do testes.

Capítulo 3

Metodologia

3.1 Coleta de dados do Senado

Para realizar a coleta dos discursos textuais, fez-se uso do portal de Dados Abertos do Senado, uma ferramenta disponibilizada como parte da iniciativa de dados abertos do Governo Brasileiro [1]. De forma a promover a transparência, o Senado Federal disponibiliza para os cidadãos uma *API* que provém diversos serviços que disponibilizam informações relativas a essa casa legislativa. Alguns dos serviços são: [16]

- ***AgendaReuniaoService***: Este serviço provê informações sobre agenda de reuniões das Comissões, com pauta, finalidade etc.
- ***LegislacaoService***: Provê serviços para recuperação de Normas Jurídicas Federais.
- ***GlossarioService***: É o Serviços de glossário. Exibe informações dos termos utilizados no Processo Legislativo.
- ***ListaSenadorService***: Este serviço provê diversos métodos para recuperação de informações de Parlamentares e suas bancadas no Senado Federal, bem como sua atuação parlamentar, incluindo discursos proferidos.
- ***PlenarioService***: Este serviço provê métodos para recuperação de informações de Sessões Plenárias, como agenda, pauta, resultados e discursos.

Contudo, o Senado não fornece os dados completos para *download* direto. É necessário, então, realizar uma coleta iterativa e programática dos dados. Para realizar esta tarefa, empregou-se a linguagem de programação *Python*. Inicialmente, seguiu-se a seguinte abordagem: consultou-se o serviço *ListaSenadorService* para obter a lista de senadores com mandatos entre as legislaturas 45 e 56, que corresponde ao intervalo entre os anos de 1995

[1] https://legis.senado.leg.br/dadosabertos/docs

e 2019, aproximadamente. Na sequência, usou-se outro *endpoint* do mesmo serviço para obter o conjunto de discursos proferidos por cada um dos senadores. Preliminarmente, o resultado da coleta consistiu em um conjunto de **89,098** discursos únicos. A Figura 3.1 apresenta, como exemplo, um item que compõe a resposta da *API* para a busca dos discursos do senador Garibaldi Alves. [2].


```
{
"CodigoPronunciamento": "87873",
"TipoUsoPalavra": {
  "Codigo": "4819",
  "Sigla": "DIS",
  "Descricao": "Discurso",
  "IndicadorAtivo": "Sim"
},
"DataPronunciamento": "1991-10-10",
"SiglaPartidoParlamentarNaData": "PMDB",
"UfParlamentarNaData": "RN",
"SiglaCasaPronunciamento": "SF",
"NomeCasaPronunciamento": "Senado Federal",
"TextoResumo": "\n      VINDA DO PAPA JOÃO PAULO II AO BRASIL.",
"Indexacao": "ELOGIO, VISITA OFICIAL, JOÃO PAULO II, PAPA, BRASIL.",
"UrlTexto":
    "http://www25.senado.leg.br/web/atividade/pronunciamentos/-/p/texto/87873",
"UrlTextoBinario":
    "http://www.senado.leg.br/atividade/rotinas/pronunciamento/getDocumento.asp?
t=87873",
"Publicacoes": {
  "Publicacao": {
    "DescricaoVeiculoPublicacao": "DCN2",
    "DataPublicacao": "1991-11-11",
    "NumeroPagInicioPublicacao": "6944",
    "IndicadorRepublicacao": "Não",
    "UrlDiario":

"http://legis.senado.leg.br/diarios/BuscaDiariotipDiario=1&dataDiario=11/11/1991&pag
inaDireta=6944"
  }
}
}
```


Figura 3.1: Exemplo de resposta *JSON* para um discurso de senador. (Fonte: [16]).

Curiosamente, além do serviço *ListaSenadorService*, o Senado possui outro serviço que disponibiliza dados dos discursos, o *PlenarioService*. Este serviço permite a busca de discursos em um intervalo de tempo especificado. Então, como os objetivos de coletar o maior número de discursos possível e solucionar quaisquer dúvidas a respeito da consistência da base do Senado, repetiu-se a mesma busca realizada anteriormente. Codificou-se, então, um *script* que coleta os discursos proferidos ano a ano, entre 1995 e 2019. Como

[2] *https://legis.senado.leg.br/dadosabertos/senador/87/discursos?casa=SF*

esperado, obteve-se números discrepantes em relação à busca anterior. A coleta baseada em data retornou um total de **100,925** discursos únicos. Dessa maneira, apesar de ser extensa, a base de discursos do Senado mostrou-se inconsistente. Após realizar um processo de análise exploratória dos dados, pode-se constatar, também, que o agrupamento de discursos por ano é distinta para cada um dos serviços utilizados. Contudo, apesar das discrepâncias observadas, os discursos possuem um identificador único. Assume-se que este identificador é, de fato, confiável, pelo fato de que cada discurso possui uma página *web* única que contém sua transcrição textual.

Após as descobertas supracitadas, foi codificada uma rotina que *mescla* as duas bases de discursos coletadas pelos serviços *ListaSenadorService* (89,098) e *PlenarioService* (100,925), levando em conta os elementos comuns, com base nos identificadores únicos. Revelou-se uma co-ocorrência de **76,235** discursos. Isto é, os discursos do *ListaSenadorService* contém **12,863** discursos adicionais e os discursos do *PlenarioService* contém **24,690** discursos adicionais. A adição de 37,553 discursos à base de dados é muito valiosa pois os algoritmos da área de Mineração de Texto beneficiam-se de bases textuais extensas. Além disso, a combinação de discursos foi útil pois discursos co-ocorrentes puderam ter suas informações combinadas, de forma a solucionar outros problemas de inconsistência e dados ausentes.

O conjunto de **113,788** discursos obtidos pela união das respostas dos serviços *ListaSenadorService* e *PlenarioService* compõem a base de dados de discursos usado neste trabalho.

Após criada a base de dados final, retorna-se a atenção para a natureza de cada discurso. O campo `TipoUsoDaPalavra` presente nas respostas *JSON* tem especial relevância para esse trabalho. Ele explicita o tipo de pronunciamento relacionado a um determinado discurso. No contexto do Senado Federal, é sabido que existem reuniões e encontros diversos, cada um com uma finalidade específica. A Tabela 3.1 apresenta os possíveis tipos de pronunciamento e a distribuição percentual dos discursos da base nessas categorias.

O tipo de pronunciamento de interesse para esse trabalho é o tipo Discurso, que compõe a maior parte dos pronunciamentos. Esta categoria de pronunciamento descreve discursos proferidos por qualquer senador, de forma livre. Tipicamente, estes pronunciamentos são direcionados à base apoiadora dos parlamentares e não dependem de nenhum direcionamento ou temática prévia, ao contrário dos outros tipos. Assim, os senadores têm a oportunidade de transmitir sua mensagem com bastante autonomia, bem maior do que a que eles encontram quando se posicionam nas votações nominais, sujeitas aos constrangimentos partidários e de relação com o Poder Executivo. Essa característica é fundamental, pois tem-se como objetivo o descobrimento de temas independentes e tão diversos quanto possível. Além disso, sabe-se também que os senadores utilizam deste

Tabela 3.1: Tipos de pronunciamento e a distribuição de discursos.

Contagem	**Sigla**	**Descrição**
81858 (71.94%)	DIS	Discurso
7013 (6.16%)	FP	Fala da Presidência
6938 (6.10%)	POR	Pela ordem
4775 (4.20%)	PDI	Discussão
4398 (3.87%)	PEN	Encaminhamento
3732 (3.28%)	PL	Pela Liderança
1786 (1.57%)	POB	Orientação à bancada
1450 (1.27%)	CIN	Comunicação inadiável
899 (0.79%)	QO	Questão de Ordem
241 (0.21%)	CRE	Como Relator
187 (0.16%)	PPP	Como Relator - Para proferir parecer
153 (0.13%)	EXP	Explicação pessoal
115 (0.10%)	NID	Não classificado
75 (0.066%)	PIC	Interpelação a convidado
66 (0.058%)	PEC	Exposição de convidado
52 (0.046%)	PIM	Interpelação a Ministro de Estado
21 (0.018%)	PEM	Exposição de Ministro de Estado
16 (0.014%)	RQO	Resposta à Questão de Ordem
9 (0.0070%)	DISPUB	Discurso encaminhado à publicação
3 (0.0026%)	PCO	Contradita a Questão de Ordem
1 (0.0008%)	DISPRES	Discurso proferido da Presidência

Tabela 3.2: Modelo de dado de um pronunciamento.

Campo	**Descrição**
id	Identificador único do pronunciamento
text	Pronunciamento textual íntegro
date	Data do pronunciamento
keywords	Indexação manual feita pelo Senado
party	Partido político do senador na data
state	Estado do senador na data
senator name	Nome do senador
senator id	Identificador único do senador
speech type	Tipo do pronunciamento
session type	Tipo da sessão plenária em que o discurso foi feito

espaço para discorrer sobre assuntos diversos como eventos recentes ao país e projetos de lei em debate, por exemplo. São essas as características que se deseja explicitar por meio dos tópicos.

Com este entendimento, realizou-se uma filtragem dos discursos usando como critério o tipo de pronunciamento. O conjunto resultante contém **81,858** discursos únicos.

Ademais, os dois serviços consumidos retornam esquemas de dado distintos. A Tabela 3.2 explicita o novo esquema de dado proposto, que contém apenas campos relevantes.

3.1.1 Raspagem de dados

Além de metadados como senador associado, resumo do pronunciamento e partido político, os serviços retornam também o endereço web em que se encontra a transcrição integral do discurso, que é, de fato, o dado mais relevante para este estudo. Esse dado é retornado pelo campo `"UrlTexto"`, que é o endereço de uma página *web* que contem o dado textual em formato *HTML*, conforme mostra a Figura 3.2. Por exemplo, esta é a página referente a um discurso proferido pelo senador Garibaldi Alves em outubro de 2007: *https://www25.senado.leg.br/web/atividade/pronunciamentos/-/p/texto/370609*. O valor numérico ao final do endereço web é o identificador único do pronunciamento.

Frequentemente em processos de coleta de dados, o material de interesse possui informação alheia que não contribui com conteúdo em si [17]. Dessa maneira, empregou-se técnicas de raspagem automatizada de dados para remover o ruído do *HTML* a fim de revelar o conteúdo textual puro dos discursos. A biblioteca Python BeautifulSoup4 foi utilizada para axiliar no processo de raspagem.

De forma análoga à etapa anterior, também houve desafios na etapa de raspagem. Os discursos inconsistentes, em geral, foram descartados e não foram contabilizados na

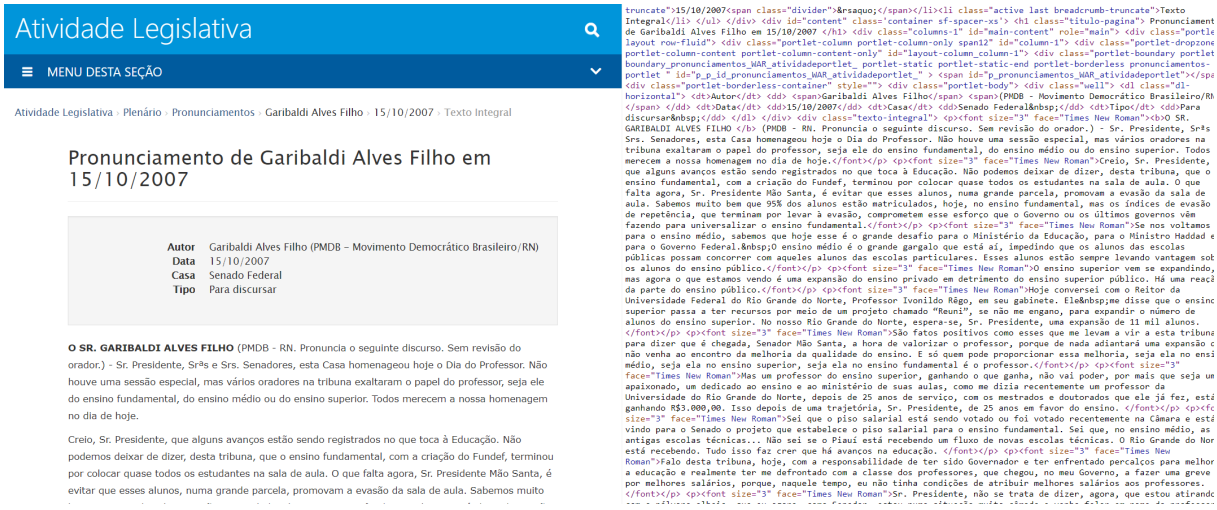

Figura 3.2: Página *web* de um discurso e seu código fonte *HTML* (Fonte: [16]).

contagem apresentada na seção anterior. Os principais tipos de inconsistências de dado foram as seguintes:

- **Discurso demasiadamente curto**: Esses discursos contém apenas cabeçalho e/ou introdução. O conteúdo em si não está presente.

  Exemplo: *Discurso 310787*

- **Discurso indisponível**: É o tipo de inconsistência mais frequente. A página *web* apresenta os metadados do pronunciamento, mas informa que o conteúdo, de fato, está indisponível.

  Exemplo: *Discurso 476436*

- **Discurso em vídeo**: Esses discursos não dispõem de transcrição taquigráfica. É disponibilizado um recorte do video da sessão plenária em questão (pouco útil para o contexto deste estudo).

  Exemplo: *Discurso 476436*

Em conclusão, o conjunto total de etapas seguidas para realizar o processo de extração de dados é explicitado na Figura 3.3.

3.2 Pré-processamento

A etapa de pré-processamento mostrou-se a mais sensível e a que mais impactou na qualidade dos resultados finais. Por isso, foi necessário adotar uma abordagem exploratória e

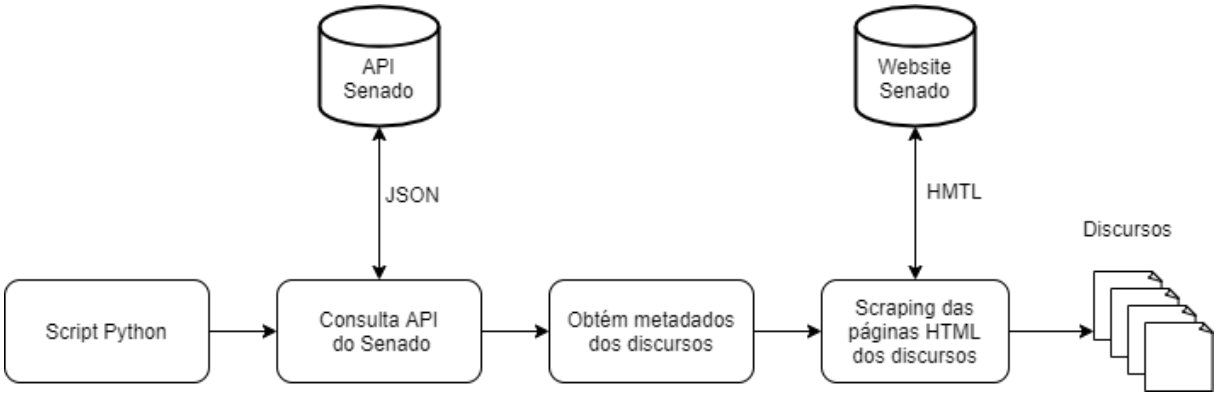

Figura 3.3: Etapas da coleta de dados.

iterativa. Além disso, essa etapa tem como objetivo produzir um modelo de *bag of words* ou "sacola de palavras" para cada discurso. Seguiu-se as seguintes etapas:

- Remoção de números, acentos, caracteres especiais, letras maiúsculas, múltiplos espaços em branco e quebras de linha;

  Ex: `São 50 mil reais.` → `sao mil reais`

- Nomes compostos de estados brasileiros foram concatenados para que sua informação semântica fosse preservada nos tópicos;

  Ex: `Rio de Janeiro` → `riodejaneiro`

- Removeu-se palavras com menos de 4 letras. Foi determinado experimentalmente que palavras com um número inferior de letras não carrega informação relevante.

  Ex: `O Senado existe para igualar a representatividade dos Estados.`
  → `senado existe para igualar representatividade estados`

- Excluiu-se *stopwords*, que é um conjunto de palavras rotuladas como "vazias" principalmente por não possuírem valor semântico. Tipicamente são artigos, preposições e outras palavras excessivamente frequentes. Inicialmente foram usadas as *stopwords* das bibliotecas Python Spacy [3] e NLTK[4]. Ademais, para o caso deste trabalho, criou-se modelos LDA e tomou-se nota, de forma manual e experimental, das palavras "irrelevantes" que compunham os tópicos gerados. Essas palavras foram, de forma iterativa, adicionadas à lista de *stopwords* total. Dessa forma, assegurou-se que o *corpus* continha o mínimo de ruído possível. Ainda, descartou-se discursos que, após o *stemming*, continham menos de 20 palavras.

[3] *Referência da biblioteca Spacy.*
[4] *Referência da biblioteca NLTK.*

Ex: `O Senado existe para igualar a representatividade dos Estados.`
→ `igualar representatividade`

- Aplicou-se a técnica de *stemming*, utilizando a biblioteca *Python NLTK*. Para garantir melhor legibilidade adiante, palavras com um mesmo *stem* foram substituídas pela ocorrência mais frequente em todo o *corpus*.

  Ex: `Falamos fala falei falou falamos.`
  → `falamos falamos falamos falamos falamos`

- Finalmente, realizou-se o processo de filtragem de extremos no *corpus*. Adotou-se a metodologia observada na literatura [18] - removeu-se palavras muito frequentes e pouco frequentes, isto é, que aparecem em mais de 90% e em menos de 0.5% de discursos, respectivamente. Apenas esta filtragem reduziu o tamanho do vocabulário de **262,601** palavras para **5,811** palavras distintas.

A Figura 3.4 ilustra a contagem progressiva dos discursos após cada filtragem.

3.3 Modelo LDA

Para gerar as matrizes documento-tópico, utilizou-se a biblioteca Python *Gensim* com a implementação *MALLET* [5], por sua capacidade conhecida de gerar tópicos de qualidade, em comparação a outras implementações. Em termos de parâmetros, utilizou-se o recurso de α automático, que permite que o algoritmo encontre automaticamente um valor do hiper-parâmetro α ótimo para o corpus em questão. Além disso, foi utilizado o valor 10,000 para o parâmetro do número de iterações.

Um dos desafios do problema de modelagem probabilística de tópicos é a seleção do número de tópicos k ideal, que depende tanto da coleção de documentos quanto da aplicação de interesse. Esta escolha tomou como direcionamento inicial a métrica $NPMI$. Conforme exposto no referencial teórico deste trabalho, a literatura aponta a métrica $NPMI$ como um indicador confiável de coerência de modelos de tópicos [11]. Computou-se então, o valor de $NPMI$ para diversos modelos LDA com diferentes número de tópicos. O resultado é visto na Figura 3.5 e na Figura 3.6.

A Figura 3.5 expõe os valores de $NPMI$ para um grande intervalo entre $k = 5$ e $k = 650$. É visível um expressivo pico na vizinhança de $k = 100$ seguido de um rápido decrescimento de coerência para altos valores de k.

A Figura 3.6, por sua vez, apresenta os mesmos valores, mas ampliados em torno do pico mencionado. O gráfico indica como candidatos de k os valores 40, 50, 60, 65, 85 e

[5] *Referência da biblioteca LDA MALLET.*

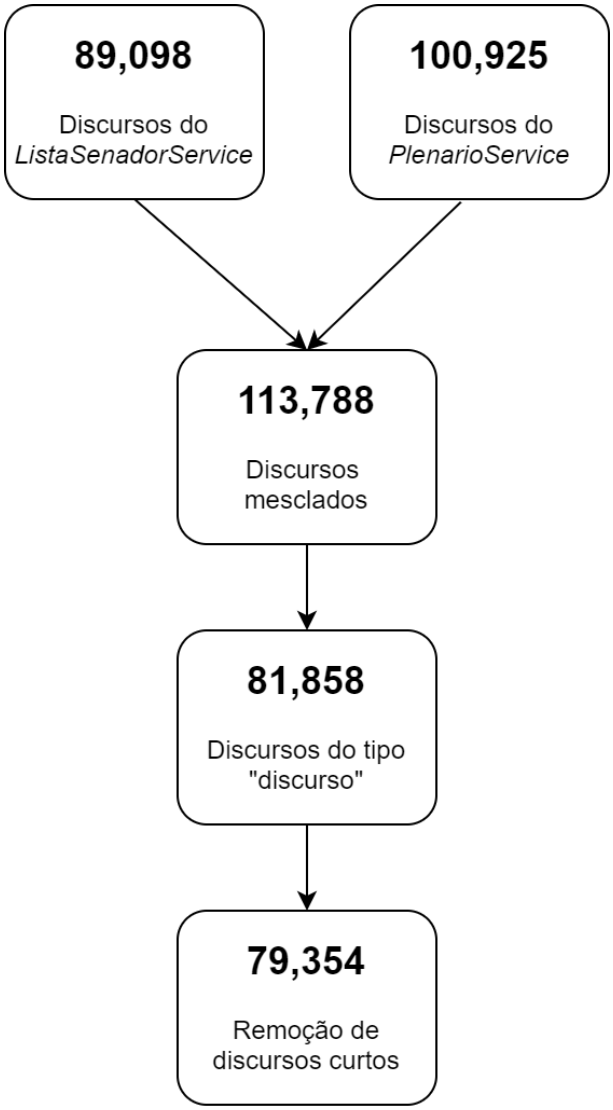

Figura 3.4: Número de discursos após cada etapa de filtragem.

95, que são os picos mais expressivos de coerência. Este trabalho optou por selecionar como valor final $k = 65$. Essa escolha se dá pelo fato do valor ser o primeiro grande pico expressivo, se diferenciando dos demais apenas por poucos milésimos. Além disso, o valor mediano oferece bom equilíbrio entre tópicos genéricos e tópicos específicos, que apresenta suficiente diversidade temática. Conforme exposto anteriormente, um k menor provoca tópicos mais genéricos e menos diversos, ao passo que um k maior resulta em tópicos mais específicos.

Contudo, o modelo criado para $k = 65$ certamente contém tópicos pobres. O próximo desafio, então, é manter apenas o conjunto de tópicos mais expressivos do modelo. Computou-se os valores de $NPMI$ para cada um dos 65 tópicos criados. Neste caso, um tópico t nada mais é que um conjunto de palavras $(w_1, w_2, ...w_n)$. Então, o NPMI

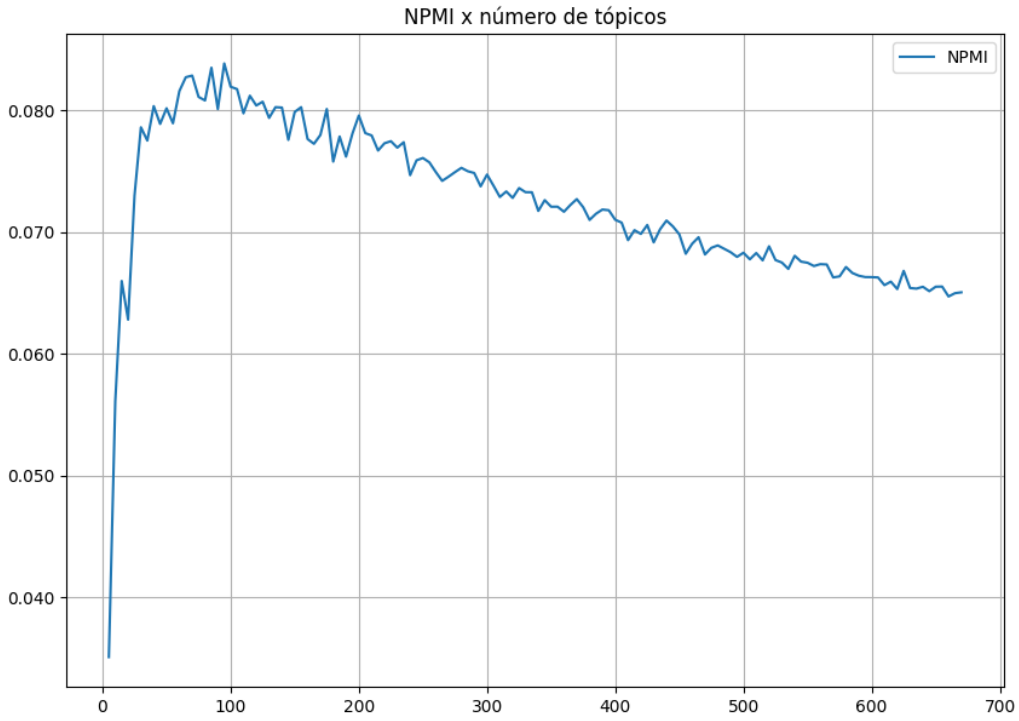

Figura 3.5: NPMI para k entre 5 e 650.

do tópico é calculado como a mediana dos NPMI para cada par de palavras do tópico, obedecendo a equação:

$$NPMI(t) = mediana\{NPMI(w_i, w_j), w_i, w_j \in t\} \tag{3.1}$$

Porém, verificou-se experimentalmente que os resultados dessa métrica foram pouco úteis e pouco consistentes. No geral, os valores encontrados estiveram em intervalos próximos, aproximadamente -0.1. Além disso, observou-se tanto bons tópicos com valores de NPMI baixos quanto a situação contrária. Assim, o processo de avaliação automatizada foi impossibilitado para este trabalho. Acredita-se que o uso bases de dados extensas e externas para consulta, ao invés do próprio corpus, possa trazer resultados mais significativos. Porém, esta abordagem está fora do escopo deste estudo.

Então, foi usado o julgamento humano para a rotulagem dos tópicos. O processo, simples, consiste na leitura das palavras de cada tópico para determinar sua coerência semântica. Esse processo reduziu o conjunto inicial de **65** tópicos para um conjunto final de **40** tópicos coerentes, que abordam temas como economia, educação e saúde. A listagem completa de tópicos está na seção de anexos.

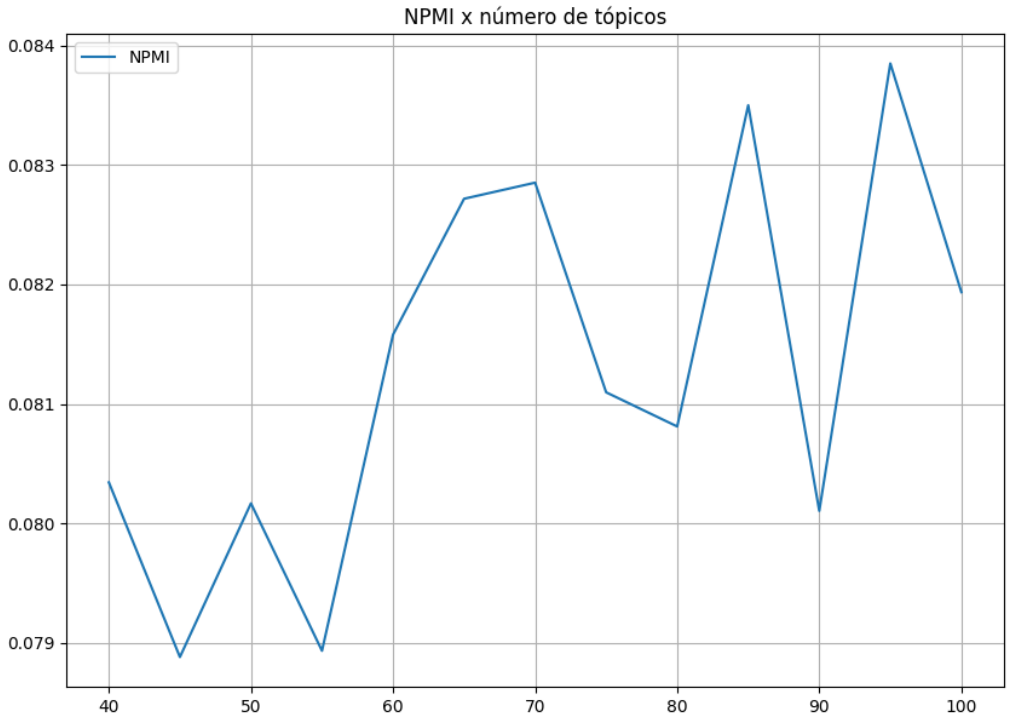

Figura 3.6: NPMI para k entre 40 e 100.

No entanto, o modelo criado ainda segmenta os discursos com base no conjunto inicial de tópicos. Então, fez-se necessário o recálculo da distribuição de tópicos de cada documento. Excluiu-se os tópicos fora da lista de tópicos coerentes e, após a exclusão, as distribuições foram normalizadas.

A Figura 3.7 apresenta um exemplo do procedimento de normalização proposto para um modelo hipotético de 10 tópicos. No exemplo, à esquerda, apresenta-se uma distribuição arbitrária de tópicos para um discurso. Supõe-se que os tópicos $3, 5, 9$ sejam pobres e, portanto, devem ser removidos. À direita está a nova distribuição de tópicos, em que há apenas tópicos de qualidade, com suas distribuições corrigidas.

Ainda com o intuito de melhor expor o funcionamento do LDA, exemplifica-se a distribuição temática de um dos discursos proferidos pelo senador Garibaldi Alves em 2007 [6]. A Tabela 3.3 apresenta os principais tópicos presentes no discurso.

Em geral, os dois ou três principais tópicos são suficientes para que se possa ter o devido entendimento contextual de um determinado discurso. Neste caso, pode-se entender, à priori, sem sequer ler o discurso integral, que o pronunciamento aborda uma reivindicação

[6] *Texto integral do discurso.*

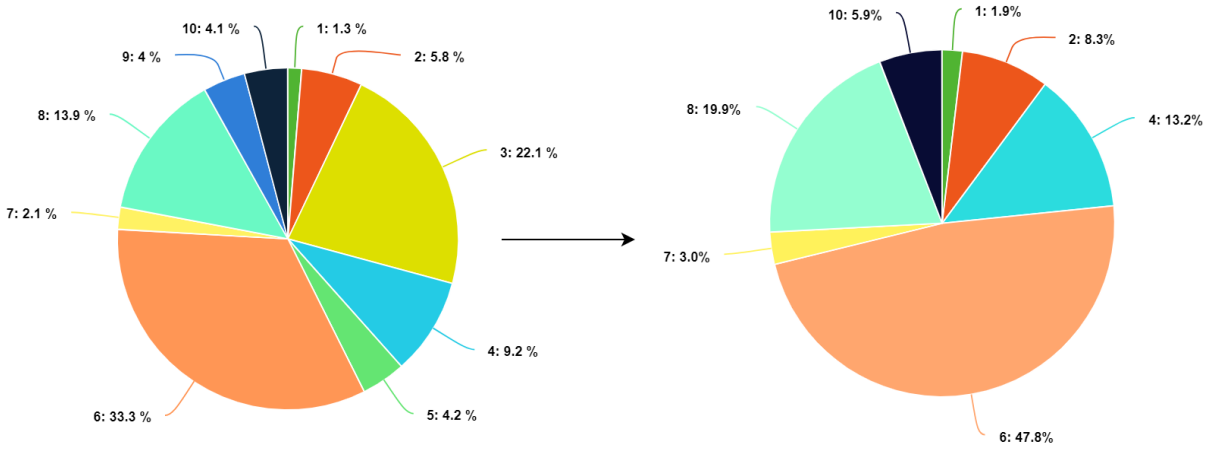

Figura 3.7: Exemplo de decomposição de tópicos antes e após filtragem de tópicos pobres.

Tabela 3.3: Principais tópicos do discurso 370187.

Tópico	**Contribuição**	**Tema**
14	63.54%	Educação
1	8.0%	Aumento de salário
31	3.05%	Pesquisa universitária
51	2.76%	Seca do Nordeste
55	1.43%	Previdência e aposentadoria
54	1.28%	Programas sociais

de melhores salários para a categoria dos professores. De fato, o resultado obtido é coerente. É possível verificar a corretude da distribuição após a leitura do texto na íntegra, disponível no site do Senado Federal.

3.4 Análise anual

O objetivo principal desse trabalho é comprovar a hipótese de que os tópicos dos discursos podem explicitar informações a respeito de contextos históricos, políticos e econômicos.

A metodologia proposta por esse trabalho sugere uma observação da evolução histórica dos tópicos ano a ano. Para cada tópico k e ano a, sugere-se a definição das seguintes métricas avaliativas:

- **Contribuição média**: Supõe-se que todos os discursos do ano a são combinados em 1 único discurso (discurso médio). Esta métrica observa a distribuição temática desse discurso geral e indica a contribuição percentual do tópico k.
- **Contagem de discursos dominantes**: É realizada uma contagem percentual do número de discursos do ano a em que o tópico k é o mais expressivo. Ou seja, é a

contagem de quantos são os discursos que falam majoritariamente do tópico k no ano a.

Estas métricas, combinadas, se mostram úteis para que se possa verificar a evolução de um tema ao longo dos anos, de acordo com os discursos proferidos pelos parlamentares. A Figura 4.5 ilustra as métricas citadas para o tópico 33, que pode ser descrito genericamente como "corrupção".

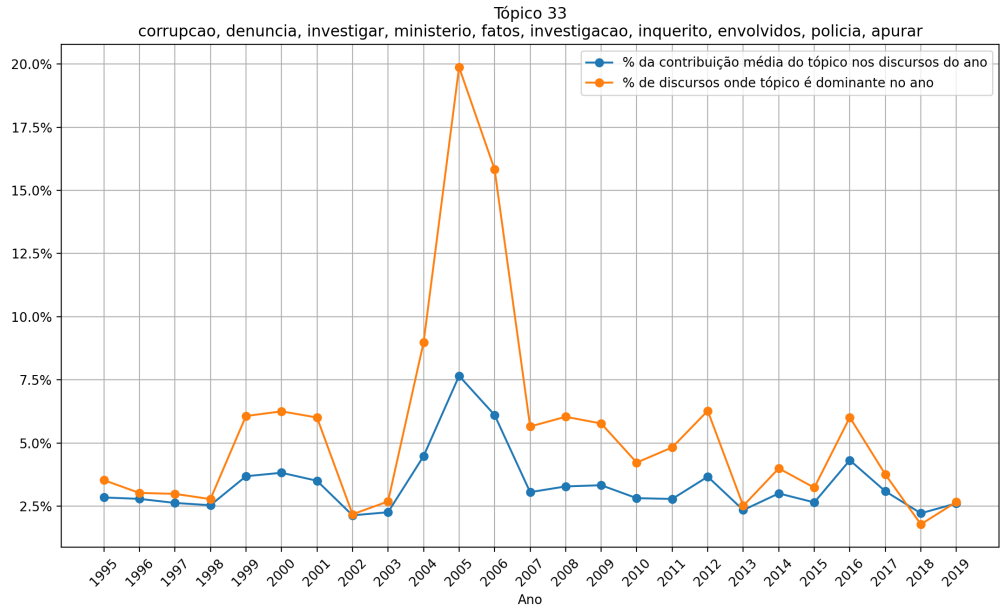

Figura 3.8: Tópico 33 entre 1995 e 2019.

Com base no exemplo, pode-se tirar as seguintes interpretações. No ano de 2005, 20% de todos os discursos tinham como principal tema o tópico 33. Além disso, no mesmo ano, o tópico 33 possui uma contribuição média de aproximadamente 7.5% em relação a todos os outros tópicos.

Após a criação dos gráficos de evolução temática para cada tópico, o próximo passo da análise é buscar uma correlação entre os picos mais expressivos dos gráficos com efetivos acontecimentos históricos. Para realizar essa tarefa, este trabalho utilizou duas abordagens manuais.

Para um pico em um determinado ano para um dado tópico, primeiro, realizou-se consultas em motores de buscas da internet da combinação das principais palavras do tópico com o ano em questão. Tentou-se com isso, localizar notícias ou outras evidências históricas que pudessem justificar tal pico. Para o exemplo do gráfico do tópico 33 supracitado, pesquisou-se por "corrupção denúncia investigar 2005" no site *Google*. As primeiras correspondências da busca foram suficientes para sugerir a conclusão de que o

Tabela 3.4: 5 discursos que mais abordam o tópico 33 no ano 2005.

Id do discurso	**Contribuição do tópico 33**
356041	69.47%
357131	63.15%
355990	62.92%
355013	61.52%
357093	61.44%

pico observado em 2005 se deu principalmente pelo esquema de corrupção do chamado "mensalão", que foi um dos maiores e mais icônicos escândalos de corrupção da história do Brasil.

A segunda abordagem adotada para validação dos gráficos foi a inspeção e leitura dos principais discursos de um determinado ano. Foi desenvolvida uma ferramenta que permite a pesquisa dos discursos mais representativos de um determinado tópico para cada ano. Como exemplo, considerando novamente o tópico 33 e o ano 2005, encontrou-se os seguintes discursos:

A Tabela 3.4 indica que o *discurso 356041* é o discurso mais representativo do tema corrupção em 2005. Após uma minuciosa leitura do texto integral, observa-se, novamente, coerência. O senador Aloizio Mercadante traz ao Senado uma reflexão sobre os acontecimentos mais recentes à época: a CPI dos correios e o esquema do mensalão.

A ideia da metodologia proposta, então, é encontrar em pelo menos uma das duas abordagens algum acontecimento histórico que possa justificar os picos observados nos gráficos de evolução temporal dos temas. No caso do exemplo indicado, ambas as abordagens se mostraram úteis.

Capítulo 4

Resultados

4.1 Base de discursos

Após o processo de coleta massiva, tratamento e processamento da base de discursos do Senado Federal, caracteriza-se os resultados encontrados. A Figura 4.1 apresenta a contagem dos discursos ano a ano. Como pode ser observado, existe certa discrepância em relação à quantidade de discursos. Os anos de 2005 e 2006 representam os anos com mais discursos, 5500 por ano, aproximadamente. Por outro lado, os anos de 1998, 2002 e 2018 são os anos em que menos houve pronunciamentos de senadores, 1700 por ano, em média. Contudo, idealmente, esperava-se contagens anuais com mais constância do que foi, de fato, observado. Hipotetiza-se que a discrepância possa ser um resultado natural causado pelas variação das atividades dos senadores no Senado Federal. Porém, também não descarta-se a possibilidade das contagens vistas serem causadas por um problema de inconsistência técnica do senado. Não é possível, ainda, concluir se o problema foi causado na etapa de digitalização de discursos ou na etapa de fornecimento dos dados por meio da *API*.

A etapa de limpeza, pré-processamento e tratamento dos discursos foi fundamental para a obtenção dos resultados encontrados. Observou-se que, em média, os discursos continham 9300 caracteres ou 1500 palavras. Após o processo de *stemming*, remoção de *stopwords* e filtragem de extremos, observou-se uma redução para, em média, 4500 caracteres ou 500 palavras. A redução percentual observada em número de caracteres e número de palavras foi de aproximadamente 50% e 67%, respectivamente. Conforme exposto anteriormente, a redução do vocabulário é fundamental para mitigar o ruído nos textos.

Por sua vez, a Figura 4.1 compara as 20 palavras mais frequentes, em ordem, dos discursos antes e após o processo de pré-processamento, ilustrando a necessidade e a importância da redução desse tipo de ruído.

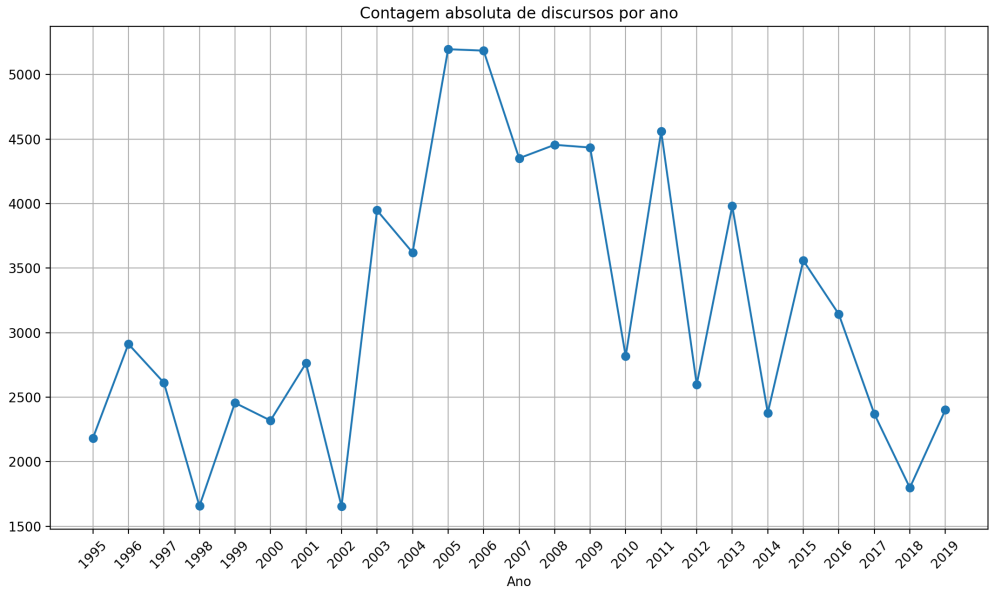

Figura 4.1: Contagem de discursos por ano entre 1995 e 2019.

Em relação à seleção dos 40 tópicos criados pelo modelo LDA, observou-se a distribuição representada na Figura 4.2. A figura ilustra a contribuição de cada tópico em relação a todos os documentos. Em uma condição ideal, em que os tópicos possuem a mesma relevância e frequência, espera-se que cada um apresente uma contribuição média de 2.5%. Portanto, flutuações acima ou abaixo desse valor revelam se um determinado tópico é super um sub representado por meio dos discursos. Esta análise mostra que os temas mais frequentes foram “corrupção”, “legislação” e “programas de desenvolvimento”. Por outro lado, revelou-se que os temas abordados com menos frequência foram “energia elétrica”, “aviação” e “povos indígenas”.

4.2 Análise temporal

Em termos gerais, os resultados encontrados após a análise dos gráficos de evolução temporal dos tópicos se mostraram, no mínimo, promissores. Boa parte dos gráficos puderam explicitar claramente um ou mais eventos históricos que se associaram aos picos mais expressivos observados. Abaixo apresenta-se os gráficos para os tópicos com melhores resultados, bem como uma breve caracterização histórica para cada um. Ao final, são expostos exemplos de gráficos com a ausência de picos, ou com picos cujas justificativas não foram encontradas.

Tabela 4.1: 20 palavras mais frequentes nos discursos antes e após pré-processamento.

Palavra	**Frequência**	**Palavra**	**Frequência**
de	4906389	economia	157623
que	3841319	passado	155041
a	3543289	estados	131934
o	3208496	mundo	119034
e	2690539	social	113673
do	2373754	desenvolvimento	111025
da	1751475	vida	102866
para	1395332	direito	102858
em	1131310	partido	102214
é	1120506	saude	98657
um	1070077	recursos	98333
não	1067734	servicos	93645
com	1044999	milhoes	90255
-	1000014	problema	90236
uma	912218	melhor	85620
no	891820	republica	84354
os	818772	sociedade	83521
na	702825	lula	82872
O	697711	municipios	82423
se	683971	empresas	81568

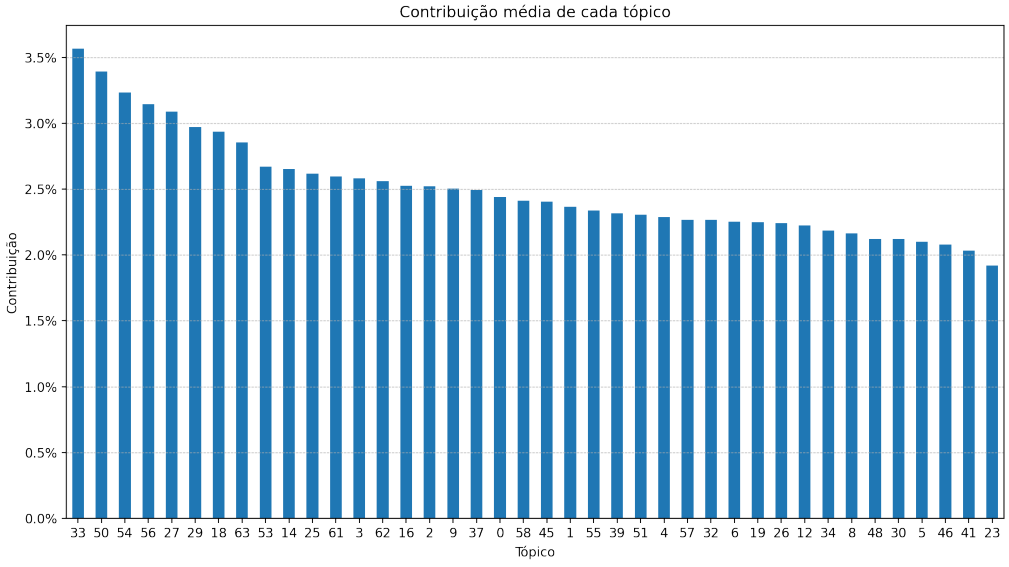

Figura 4.2: Contribuição média de cada tópico.

4.3 Resultados positivos

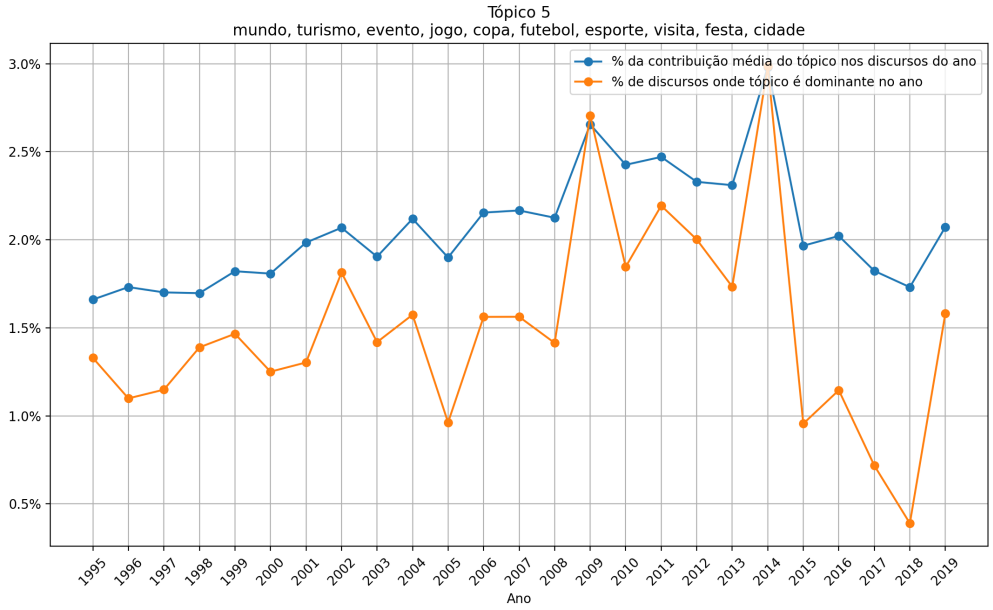

Figura 4.3: Evolução do tópico 5 entre 1995 e 2019.

O tópico 5, cujo tema é "esporte", é retratado na Figura 4.3. Observa-se que houve picos expressivos nos anos de 2009 e 2014. Os respectivos picos são justificados pela vitória da seleção brasileira na Copa das Confederações de 2009 e na participação da Seleção na Copa do Mundo de 2014. Ainda, ao comparar os anos de 2008 e 2009, pode-ser constar que houve um aumento de aproximadamente 90% no número de discursos que abordam principalmente o tema em questão.

A Figura 4.4 apresenta um exemplo curioso. Ao invés de explicitar um evento histórico em particular, o gráfico torna visível um padrão relacionado à uma característica política. O tópico 27 diz respeito a questões eleitorais, votos e candidaturas. O gráfico expõe que há um claro padrão periódico relacionado aos picos observados. Os picos acontecem aproximadamente de 4 em 4 anos, período que coincide com as eleições para presidente, governadores, deputados e senadores. Nos principais discursos desse tópico, observa-se que os senadores usam seu espaço político para expor as principais conquistas de seu partido e governantes durante o mandato que se encerra.

Dentre todos os exemplos, o tópico 33 é provavelmente o mais icônico. O gráfico da Figura 4.5 tem um pico expressivo entre os anos de 2004 e 2006. Este foi precisamente o período em que foram apurados esquemas de corrupção na CPMI dos correios. O principal desses escândalos foi o esquema do mensalão, o mais conhecido esquema de corrupção e desvio de dinheiro da história do Brasil.

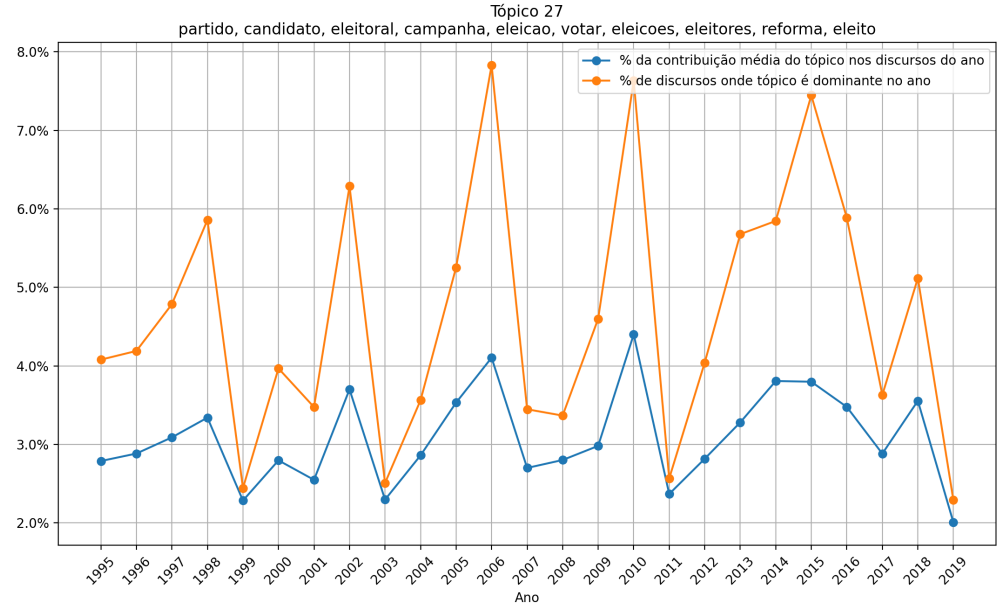

Figura 4.4: Evolução do tópico 27 entre 1995 e 2019.

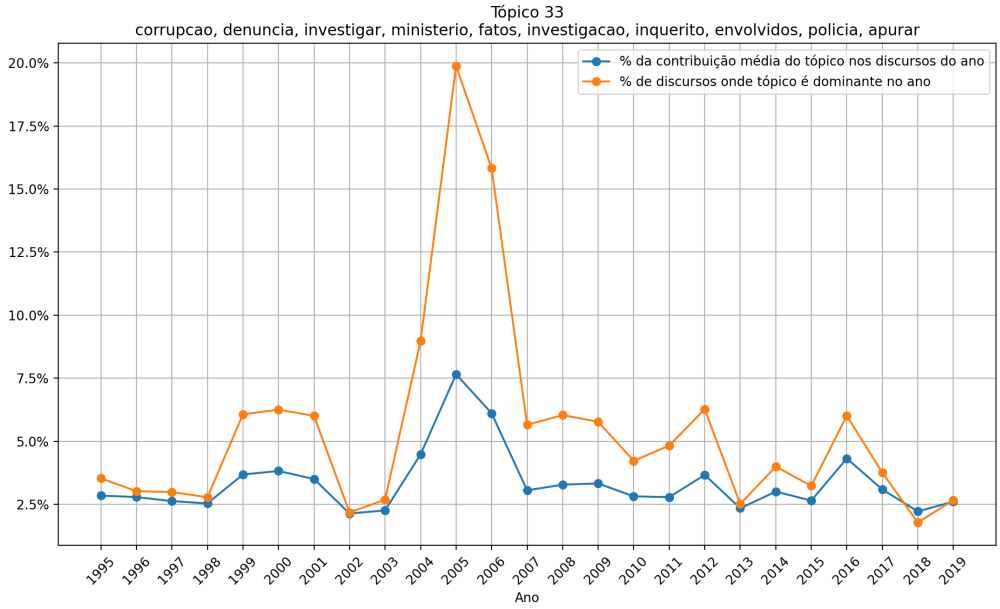

Figura 4.5: Evolução do tópico 33 entre 1995 e 2019.

O tópico 41, representado pela Figura 4.6, tem como tema a aviação e, especificamente, acidentes aéreos. Os picos encontram-se nos anos de 2007 e 2014. Estes foram anos em que ocorreram conhecidos acidentes aéreos no Brasil e no exterior. O primeiro pico está relacionado ao acidente do voo TAM 3054, considerado a maior tragédia da

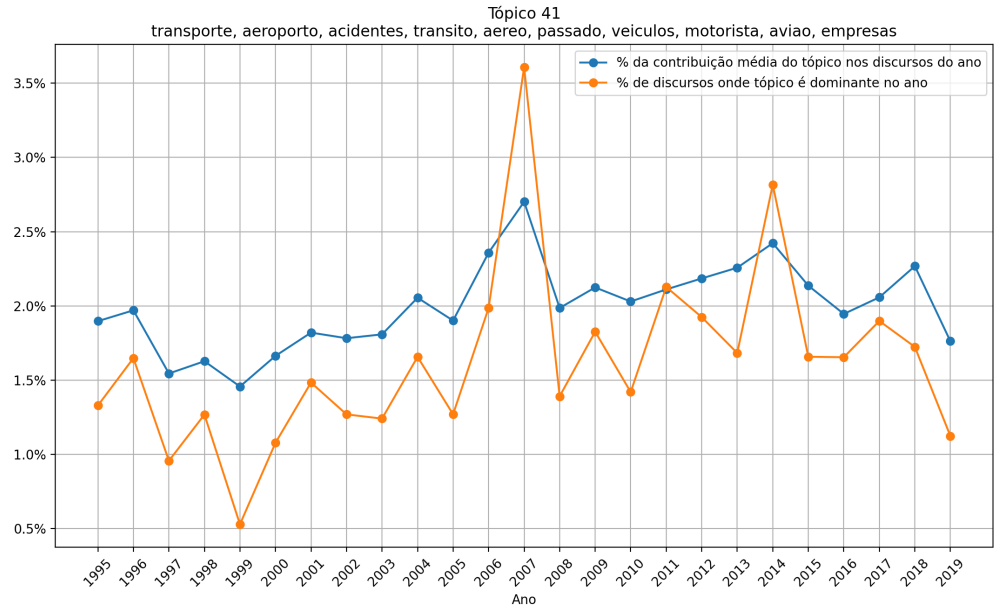

Figura 4.6: Evolução do tópico 41 entre 1995 e 2019.

aviação brasileira. Já o segundo, por outro lado, associa-se ao voo Malaysia Airlines 370, que causou comoção mundial com o desaparecimento de uma aeronave comercial cujos destroços nunca foram localizados. Nos discursos desse tópico, os senadores relembram as tragédias e reivindicam por mais segurança nas aeronaves e aeroportos brasileiros.

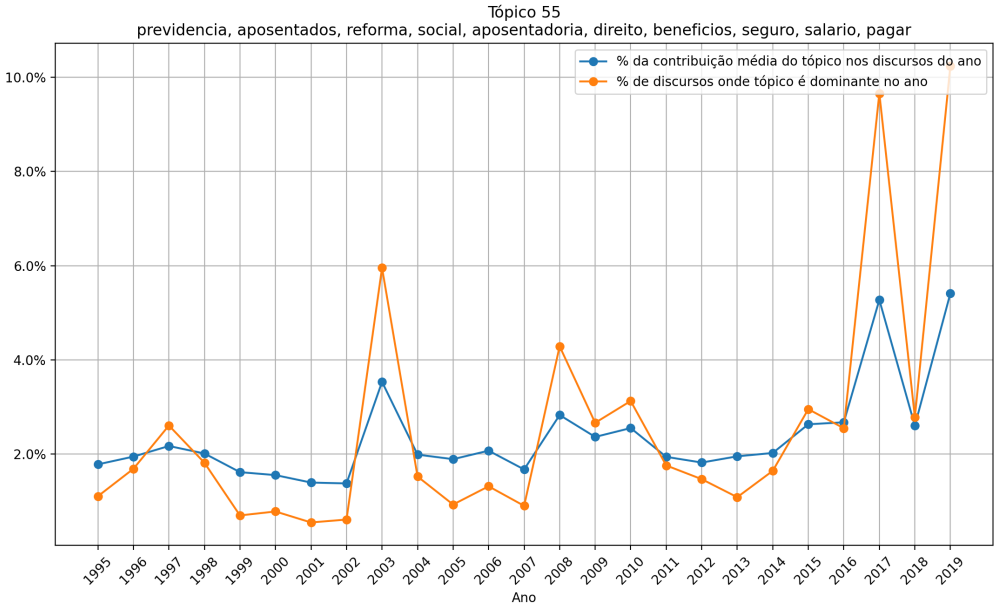

Figura 4.7: Evolução do tópico 55 entre 1995 e 2019.

O tópico mostrado na Figura 4.7 aborda a reforma da previdência e reivindicações previdenciárias. Os picos mais expressivos, nos anos de 2003, 2017 e 2019, alinham-se com as principais Propostas de Emenda Constitucional apresentados sobre o tema. São eles a PEC nº 40 de 2003, a PEC nº 287 de 2016 e a PEC nº 6 de 2019 respectivamente.

4.4 Resultados negativos

Apesar da preliminar coerência observada em grande parte dos tópicos encontrados, existiram também tópicos cujos gráficos não trouxeram boas conclusões. O exemplo inicial é o tópico 63, apresentado pela Figura 4.8, que aborda de cultura em geral. No gráfico observado, não existem picos delimitados ou qualquer outro tipo de padrão. Neste caso, o entendimento que se extrai é que não existiram grandes acontecimentos em torno do universo semântico em questão.

Outra possibilidade é o fato de que o tema não atrai a atenção de um número grande de políticos, mais voltados para uma agenda fortemente concentrada em questões econômicas.

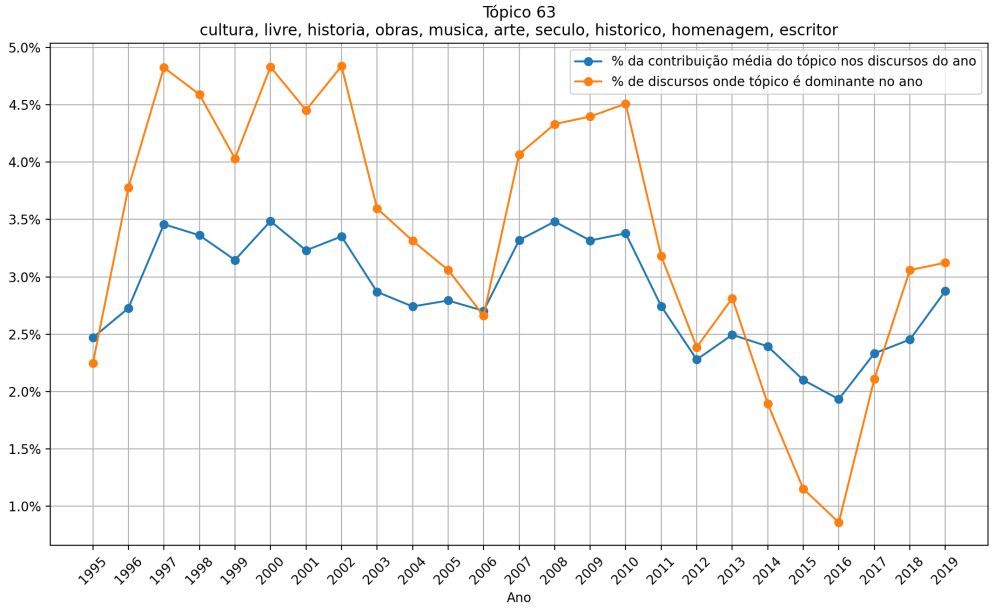

Figura 4.8: Evolução do tópico 63 entre 1995 e 2019.

O gráfico da Figura B.14 trata do tema “povos indígenas”. Aqui, novamente, existem frequentes picos cujas razões desconhece-se. Este exemplo mostra a necessidade de outro fator a se considerar durante a análise. Este fator é a contribuição média do tópico em questão. Observa-se que, em média, a contribuição temática do tópico 23 corresponte à 2%, considerada abaixo da média. Ou seja, pode-se concluir que o tópico 23 aborda um

tema pouco frequente nos pronunciamentos dos senadores e, portanto, não pode evidenciar picos expressivos.

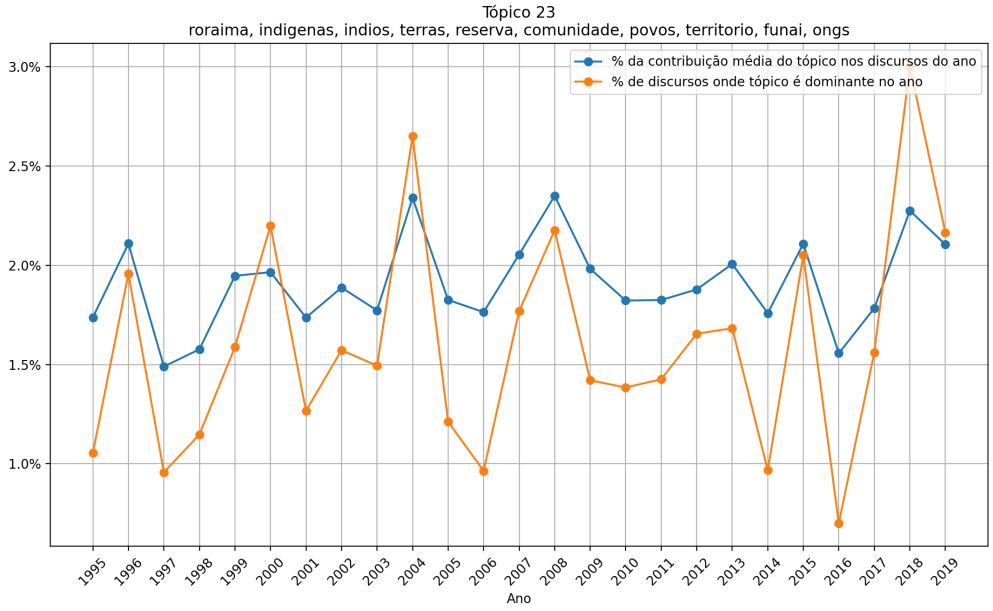

Figura 4.9: Evolução do tópico 63 entre 1995 e 2019.

Capítulo 5

Conclusão

O presente trabalho se empenhou em atingir dois objetivos principais. O primeiro é a criação e disponibilização de uma extensa base de discursos de senadores, que pode ser acessada por meio do seguinte endereço público: *https://github.com/VictorLandim/brazilian-senators-speeches*. O segundo objetivo é a verificação de se a decomposição temática dos discursos por meio do LDA é capaz de evidenciar fatos históricos ao longo dos anos.

A *API* disponibilizada pelo Portal de Dados abertos do Senado Federal foi o ponto de partida para o processo de coleta dos pronunciamentos dos senadores. Por meio dos diversos serviços expostos, pôde-se coletar uma quantidade substancial de documentos que embasaram este trabalho. É natural esperar que, eventualmente, se encontre inconsistências em grandes bases de dados como a base de dados do senado. Exemplifica-se a inconsistência citada no discurso: *370187*. Neste caso, observa-se que a transcrição do senado inclui não só a fala do senador que discursa, mas também do diálogo que ocorre entre o senador, o presidente da sessão e também entre senadores aparteantes que se manifestam. Esse tipo de dinâmica acaba por diluir a intenção temática do pronunciamento e torna a decomposição de tópicos mais dispersa. Apesar disso, o material coletado se mostrou como uma boa fonte de dados textuais de qualidade e acompanhados de metadados úteis à estudos nas áreas de Economia e Ciência Política, por exemplo.

Em seguida, criou-se o modelo LDA com os parâmetros descritos na seção 3. Após a filtragem, os tópicos resultantes mostraram-se coerentes. O conjunto de tópicos determinado explicitou de forma clara temas específicos como Saúde, Seca, Corrupção e Cultura.

Em se tratando dos resultados relacionados à análise temporal, avalia-se os expostos na seção 4 como promissores. Apesar da imprecisão em observar correspondências históricas para cada um dos tópicos, temas populares no contexto brasileiro como Corrupção, Futebol e Eleições puderam ser facilmente identificados temporalmente. Por fim, especula-se que a implementação proposta seja, ainda, incapaz de evidenciar as nuances contidas em

temas com baixa relevância proporcional nos pronunciamentos. Há que se considerar que, a despeito de o parlamento dar uma resposta para os anseios da população e aos acontecimentos conjunturais, trata-se de uma instituição muito plural, que representa interesses e segmentos diversos. Diversos são, portanto, os tópicos abordados pelos senadores.

Diante do exposto, entende-se que se cumpriu os objetivos propostos neste estudo. É disponibilizada de forma pública uma extensa base de discursos de senadores, que contém, além de metadados para cada discurso, a transcrição integral bem como uma variante pré-processada. Além disso, mostrou-se de forma conclusiva que os tópicos resultantes de modelos probabilísticos permitem que se evidencie eventos históricos quando é analisada a evolução anual dos tópicos.

5.1 Trabalhos futuros

Julga-se possível que outros estudos possam propor avanços em diferentes frentes no contexto deste trabalho. Primeiramente, a técnica de pré-processamento de lematização pode ser aplicada ao invés da técnica de *stemming*, aqui aplicada. Sabe-se que a lematização confere melhores resultados em relação às outras técnicas, mas a sua alta exigência de tempo de processamento, causada por sua maior complexidade computacional, desencorajou seu uso. No contexto de modelos de tópicos, julga-se apropriada a experimentação com implementações de modelos de tópicos mais modernos, como os *Interactive Topic Models* [17]. Esse tipo de abordagem permite que, na prática, a atuação humana direcione os tópicos criados. Por fim, no contexto da análise temporal dos tópicos, sugere-se o uso de extensas bases de notícias como entrada para outro conjunto de modelos LDA. Dessa forma, é possível comparar as variações temáticas nos pronunciamentos com aquelas nas notícias, de forma automatizada.

Referências

[1] Han, Jiawei e Micheline Kamber: *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, 500 Sansome Street, Suite 400, San Francisco, CA 94111, 2006. 1, 5, 7

[2] Sarkar, Dipanjan: *Text Anaytics with Python*. Apress, Bangalore, Karnataka - India, 2016. 1, 2, 5, 6, 7

[3] Tong, Zhou e Haiyi Zhang: *A text mining research based on lda topic modelling*. Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada, 2016. 1

[4] Boyd-Graber, Jordan: *Applications of topic models*. Department of Computer Science, umiacs, Language Science - University of Maryland, 2017. 1

[5] Blei, David M., Andrew Y. Ng e Michael I. Jordan: *Latent dirichlet allocation*. J. Mach. Learn. Res., 3(null):993–1022, março 2003, ISSN 1532-4435. 2, 11

[6] Paulo Faleiros, Alneu de Andrade Lopes Thiago de: *Modelos probabilÍsticos de tÓpicos: Desvendando o latent dirichlet allocation*. Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP, 2016. 2, 8, 9, 11

[7] Beghin, Nathalie e Carmela Zigoni: *Avaliando os websites de transparência orçamentária nacionais e subnacionais e medindo impactos de dados abertos sobre direitos humanos no brasil*. Instituto de Estudos Socioeconômicos, 2014. 2

[8] Blei, David M.: *Probabilistic topic models, surveying a suite of algorithms that offer a solution to managing large document archives.* Communications of the acm, 2012. 8, 9, 10, 11

[9] Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov e David Mimno: *Evaluation methods for topic models*. Communications of the acm, 2012. 11

[10] Newman, David, Jey Han Lau, Karl Grieser e Timothy Baldwin: *Automatic evaluation of topic coherence*. Em *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, página 100–108, USA, 2010. Association for Computational Linguistics, ISBN 1932432655. 11

[11] Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-graber e David Blei: *Reading tea leaves: How humans interpret topic models*. Em Bengio, Y., D. Schuurmans, J. Lafferty, C. Williams e A. Culotta (editores): *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf. 11, 12, 22

[12] Newman, David, Sarvnaz Karimi e Lawrence Cavedon: *External evaluation of topic models*. ADCS 2009 - Proceedings of the Fourteenth Australasian Document Computing Symposium, janeiro 2011. 12, 13

[13] Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders e Andrew McCallum: *Optimizing semantic coherence in topic models*. Em *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, página 262–272, USA, 2011. Association for Computational Linguistics, ISBN 9781937284114. 13, 14

[14] Bouma, G.: *Normalized (pointwise) mutual information in collocation extraction*. 2009. 14

[15] Lau, Jey Han, David Newman e Timothy Baldwin: *Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality*. Em *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 530–539, Gothenburg, Sweden, abril 2014. Association for Computational Linguistics. https://www.aclweb.org/anthology/E14-1056. 14

[16] Brasileiro, Senado Federal: *Serviços de dados abertos*, 2021. https://legis.senado.leg.br/dadosabertos/docs/, acesso em 2021-04-18. 15, 16, 20

[17] Boyd-Graber, Jordan, David Mimno e David Newman: *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*. CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, Florida, 2014. http://umiacs.umd.edu/~jbg//docs/2014_book_chapter_care_and_feeding.pdf. 19, 38

[18] Moreira, Davi: *Com a palavra os nobres deputados: ênfase temática dos discursos dos parlamentares brasileiros*. DADOS, Rio de Janeiro, 2020. 22

Apêndice A

Tópicos do modelo LDA para 65 tópicos

A.0.1 Tópicos ricos:

0 mundo unidos paises estados internacional americano nacoes europa humanos desenvolvimento norteamericano china organizacao africa america japao alemanha globo capital planeta

1 minimo salario riograndedosul sindical gaucho greve aumento reajuste sindicato recebeu ganhar movimento salarial categoria horas acordo pagar entendimento central realidade

2 obras rondonia matogrosso estradas transporte rodovias infraestrutura ferrovia trecho construcao quilometros investimentos portos velho ponte construir passado regiao ligados milhoes

3 industria produtos mercado setor economia comercio producao comercial exportador desenvolvimento importantes exportacoes investimentos crescimento preco produzir venda exportacao internacional custo

4 ambiente ambiental floresta amazonia natureza desenvolvimento preservacao areas sustentavel desmatamento mundo ibama recursos planeta preservar reserva economia agua ecologico humanos

5 mundo turismo evento jogo copa futebol esporte visita festa cidade clube realizar torcida turistico participar melhor oportunidade belo turistas alegria

6 empresas petrobras petroleo vale privada privatizacao companheiros investimentos estatal venda espiritosanto presentes doce bilhoes preco exploracao interesse refinaria lucro patrimonio

8 policia armas militar seguranca forca exercito civil riodejaneiro defesa marinha fronteira comandante bombeiros soldados combate ordem general fogo oficial passado

9 programa pobres renda social familia bolsa fome milhoes basica minimo miseria economia beneficios populacao desigualdades combate distribuicao ricos melhor alimentos

12 servicos acesso consumo cidadao internet informacao telefone dados informacoes comunicacao defesa operar passado empresas eletronico sociedade digital permite tecnologia larga

14 educacao escola ensino professor jovens alunos estudo qualidade criancas fundamental profissionais medio educacional formacao curso melhor aula universidade educadores tecnica

16 municipios prefeito estados municipal estadual vereadores uniao cidade recursos populacao participacao federacao atender recebeu fundo passado responsabilidade encontro pequenos dificuldades

18 tribunal justica supremo processo juiz direito decisao julgamento judiciario advogado constituicao juridica superior ministros judicial acao instancia corte defesa decidir

19 mulheres criancas violencia direito homem adolescentes sexual sociedade filhos familia infantil luta vida menino domestica idade humanos feminina mundo general

23 roraima indigenas indios terras reserva comunidade povos territorio funai ongs areas vivendo fronteira garimpeiros demarcacao rapido direito regiao criar existem

25 jornal imprensa jornalista comunicacao noticia globo publicada reportagem televisao revista radio folha saopaulo correio pagina midia entrevista semana informacao manchete

26 paises america argentina mercosul relacoes venezuela exterior embaixador latina acordo fronteira economia integracao bolivia paraguai internacional cuba tratamento uruguai chile

27 partido candidato eleitoral campanha eleicao votar eleicoes eleitores reforma eleito reeleicao disputa partidaria escolha turno processo deputados eleger democracia cargo

29 democracia luta historia democratico liberdade movimento direito militar partido regime ditadura popular republica forca vivendo social historico vargas esquerda periodo

30 amazonia regiao amazonas amapa manaus tocantins zona pesca estados desenvolvimento territorio paraense belo conhecimento vivendo acre criacao borracha registro polo

32 universidade pesquisa tecnologia ciencia curso estudo conhecimento instituto desenvolvimento professor instituicoes cientifica superior tecnica central faculdade criar instituicao formacao formas

33 corrupcao denuncia investigar ministerio fatos investigacao inquerito envolvidos policia apurar prova denunciar crime escandalo etica operacao processo grave sigilo republica

34 doenca saude drogas casos combate tratamento cancer causa ministerio problema prevencao bebidas mundo controle morte risco vida alcool consumo grave

37 produtos agricultura producao rural agricola agricultores produzir alimentos pequenos milhoes safra setor campainha soja preco plantar familiar toneladas milhares embrapa

39 emprego empresas pequenos economia desemprego gerar empresarios setor milhoes empreendimentos geracao criar mercado negocio social renda crescimento micro empresarial simples

41 transporte aeroporto acidentes transito aereo passado veiculos motorista aviao empresas passageiros carreira seguranca onibus aviacao horas companheiros varig civil viagem

45 vida humanos igreja deus homem fraternidade mundo sociedade catolica espirito campanha religiosa papa irmaos padre social sentido pastor cristo vale

46 energia eletrica usinas setor investimentos consumo geracao energetica gerar hidreletrica combustivel natureza alcool construcao custo producao produzir preco fonte tarifas

48 reforma parana agraria santacatarina rural assentamentos propriedade movimento campainha terras familia incra fazenda proprietarios semterra catarinense possamos problema ocupo pequenos

50 constituicao estabelecer constitucional direito legislacao normas legal juridica apresentei texto decreto regras principio garantir determinado emenda dispositivo limite definir prazo

51 agua bahia nordeste paraiba pernambuco ceara seca alagoas sergipe regiao nordestino riograndedonorte obras barragem transposicao baiano recursos estados cidade hidricos

53 banco financiamento divida pagar central credito emprestimos economia juros bndes caixa fundo recursos bilhoes contrato operacoes dinheiro tesouro titulo devemos

54 desenvolvimento programa social acoes objetivo realizar plano iniciativa atividades projetos participacao integra promover acao parceria diversos estrategia entidades destacar comunidade

55 previdencia aposentados reforma social aposentadoria direito beneficios seguro salario pagar fatos contribuicao servicos previdenciario pensionistas minimo idade vida recebeu fundo

56 economia crescimento crise juros taxa aumento investimentos inflacao realidade bilhoes plano medidas capital gastos fiscal divida mercado alto deficit desemprego

57 direito negros humanos deficiencia estatuto racial pessoa idosos social sociedade igualdade escravo luta portos preconceito branco discriminacao vida populacao cotas

58 saude medico atender hospital profissionais servicos populacao medicina tratamento pacientes plano assistencia ministerio unico equipe vida melhor cirurgia remedio qualidade

61 crime violencia seguranca policia sociedade morte presentes penal pena justica criminosos vitimas jovens cometeu assassinado combate trafico criminalidade organizado drogas

62 imposto tributaria estados receita fiscal pagar reforma aumento arrecadacao cargo cpmf renda icms uniao perda aliquota bilhoes contribuicao arrecadar tributario

63 cultura livre historia obras musica arte seculo historico homenagem escritor memoria escreveu portugal letras nasceu lingua poeta vida academia belo

A.0.2 Tópicos pobres:

7 recursos orcamento milhoes bilhoes investimentos gastos uniao emenda destino orcamentaria fundo ministerio financiamento aplicar despesas realidade contas corte obras verbas

10 vida familia amigo deus filhos falo vivendo luta querido morreu amor homem olhos carinho conhecimento passado coracao sabem sentido volta

11 indice crescimento dados pesquisa populacao resultado aumento ultimos periodo milhoes mostra estudo apresentei melhor taxa divulgado passado realidade reducao ibge

13 precisamos ideia mudar comeca problema maneira melhor pensamento passado falo futuro certo conseguiu chamado levar diferente criar discutir claro fizeram

15 requerimento pedido encaminhar esclarecer recebeu ordem informacoes assinado respeito solicito informar republica carta saopaulo documento enviado ocorreu tomada assunto resposta

17 respeito atitude etica opiniao defender partido interesse aceitar responsabilidade erro direito manifestar maneira devemos episodio tentativa comportamento oposicao entendimento manifestacao

20 documento interno termos registro regimento artigo integra inserido passado inciso jornal constante apanhamento publicada considerado ocupo citar intitulada saopaulo requeiro

21 homenagem luta registro especial representantes parabens reconhecimento palmas agradecer cumprimento comemorar honra orgulho cumprimentar presenca prestar destacar data importancia oportunidade

22 desenvolvimento regiao regioes regional nordeste economia estados investimentos centrooeste sudeste matogrossodosul recursos desigualdades criar fundo condicoes incentivos federacao sudene criacao

24 realidade atual necessidade constitui torno natureza verdadeira pratica sociedade propria devemos certo processo vale social essencial moderna consequencias fatos verdade

28 parlamentar campainha resistencia tema democracia socialismo impeachment golpe falo processo crise tirar claro dilma michel interrupcao responsabilidade decreto jato crime

31 acontecendo republica verdade congresso lider sentido passado entendimento querido falo respeito aceitar criar homem certo presidencia levar terminar vivendo fizeram

35 problema situacao grave solucao sofrimento crise perda vivendo resolver encontro apelo atingir dificuldades ocorreu providencias medidas acontecendo causa enfrentar tomada

36 congresso legislativo executivo parlamentar republica deputados medidas constituicao parlamento assembleia poderes representantes provisorias sociedade constitucional passado iniciativa devemos aprovada interesse

38 homem vida honra amigo companheiros deixou conhecimento lider respeito sentido figura palavras partido perda conviver oportunidade luta testemunho durante marca

40 cidade capital distritofederal goias populacao urbana riodejaneiro saopaulo construir construcao central moradores belo vida goiania habitantes interior vivendo casos horizonte

42 piaui homem deus melhor verdade mentira prefeito historia teresina ganhar republica livre cidade parnaiba dizia velho conhecimento nasceu mundo estudo

43 dilma presidenta acre ministros ministerio rousseff branco ajudar melhor sentido junto passado semana registro acompanhar respeito especial ultimos frente maneira

44 assunto problema prazo oportunidade permite ouco devemos preocupacao estados conhecimento tema abordar permiteme sentido respeito discutir verdade atencao trazer acredito

47 dinheiro pagar compra tirar milhoes gastos mandou venda ganhar verdade mostra falo acontecendo recebeu passado colocou mentira pior entrada jogo

49 lula oposicao republica partido lider palacio planalto passado ministros palocci nacao mostra expresidente base petista verdade saopaulo presidencia prometeu tentativa

52 sociedade avanco fundamental precisamos processo reforma melhor mudanca busca tema economia necessidade possamos responsabilidade social desafio construir oportunidade sentido papel

59 comissao assunto reuniao discutir participar debates tema apresentei comissoes plenario possamos representantes deputados presidentes parlamentar membros semana relatar especial realizar

60 servicos ministerio administrativa contas administracao orgaos cargo funcionarios controle uniao gestao contrato agencia fiscalizacao tecnica prestar responsabilidade funcionamento concurso pessoal

64 votar aprovada emenda votacao lider deputados plenario acordo veto entendimento pauta aprovacao apelo amanha pedido congresso apresentei urgencia unanimidade materias

Apêndice B

Gráficos de evolução temporal para tópicos ricos

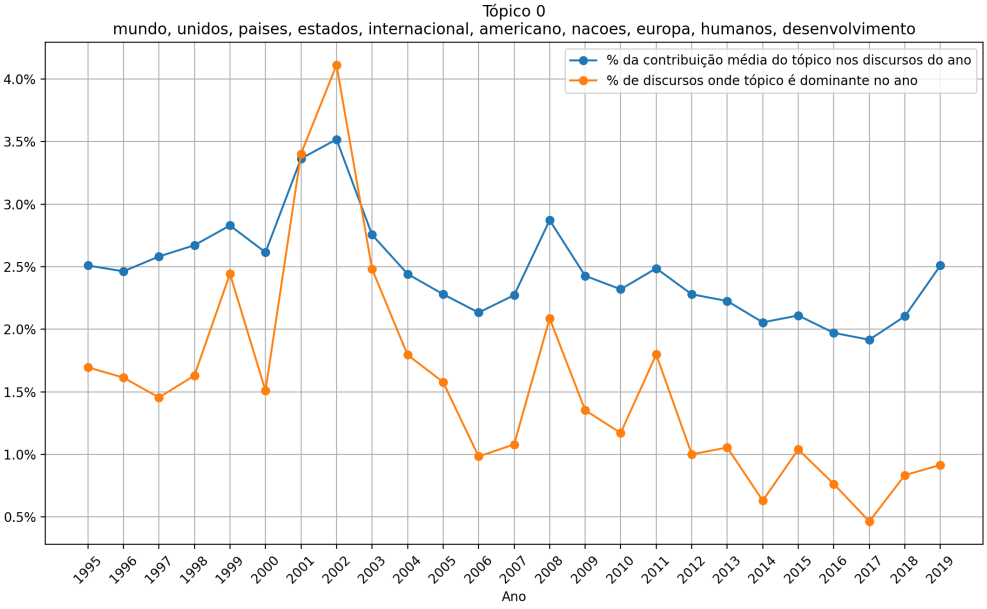

Figura B.1: Evolução do tópico 0 entre 1995 e 2019.

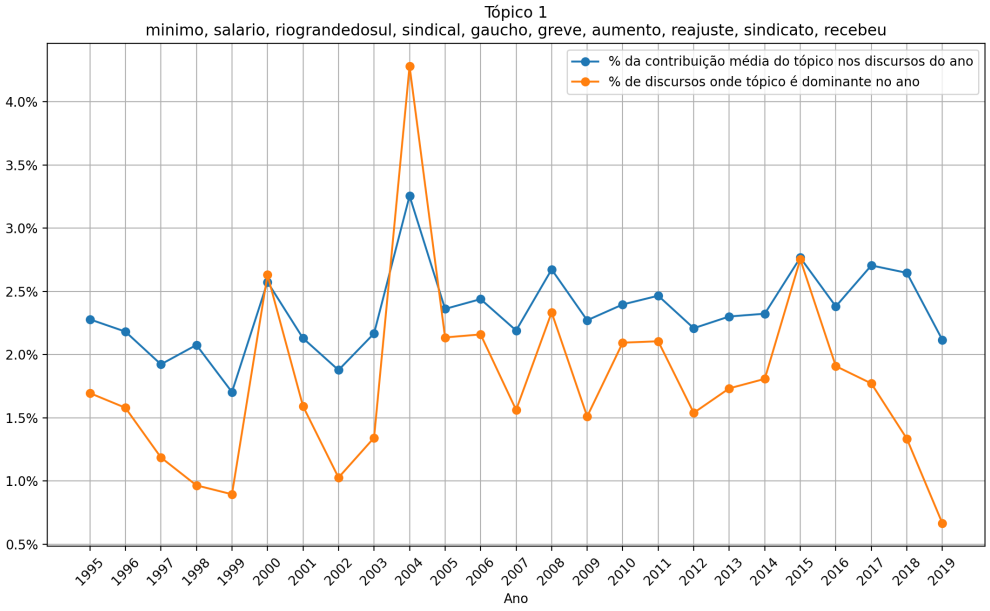

Figura B.2: Evolução do tópico 1 entre 1995 e 2019.

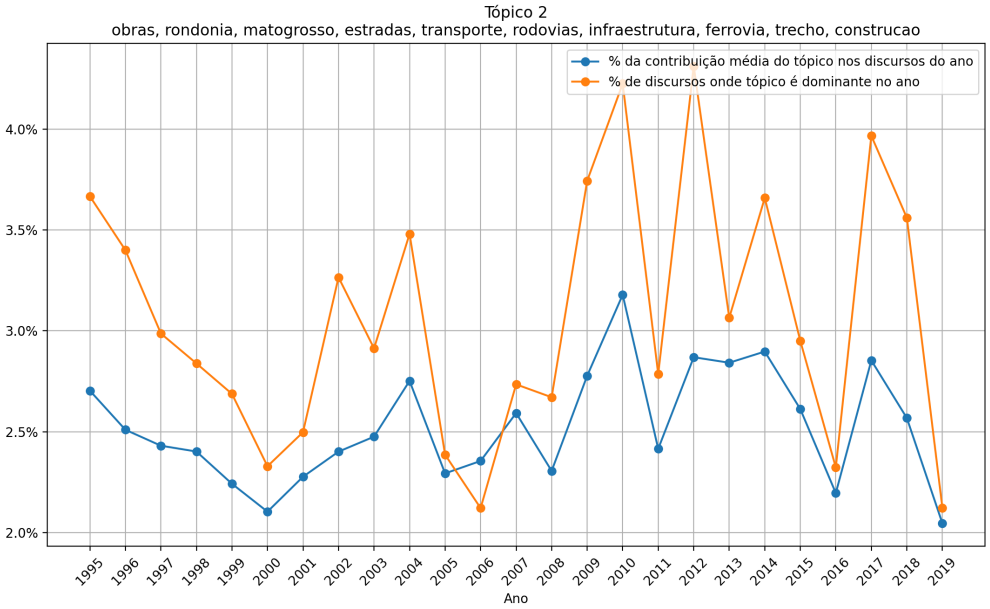

Figura B.3: Evolução do tópico 2 entre 1995 e 2019.

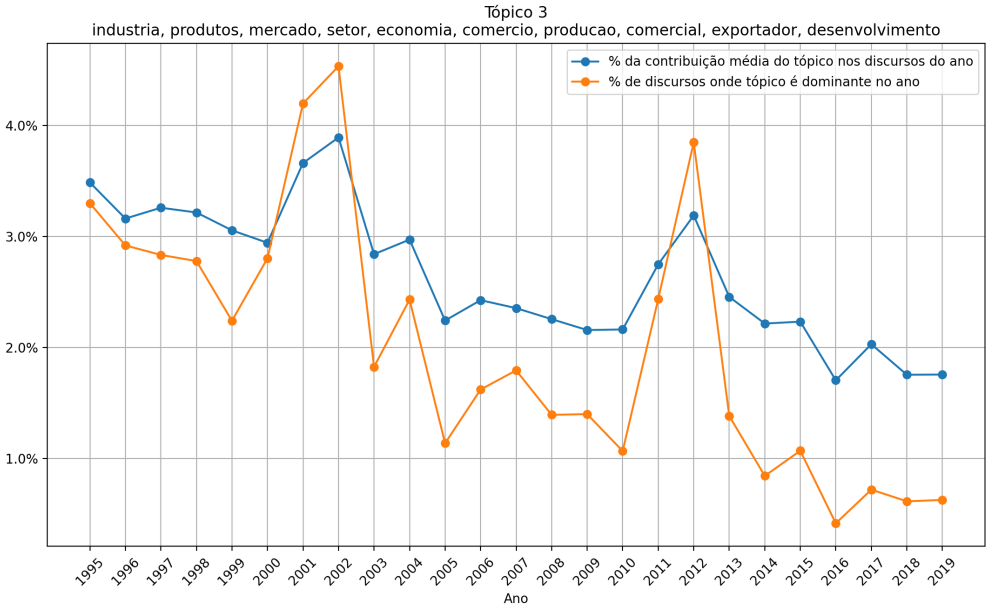

Figura B.4: Evolução do tópico 3 entre 1995 e 2019.

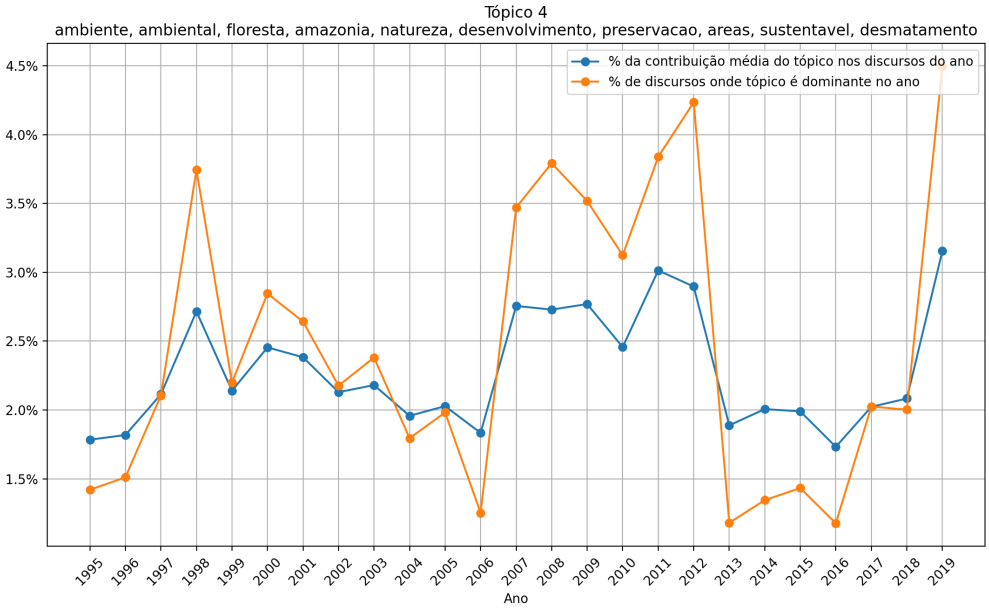

Figura B.5: Evolução do tópico 4 entre 1995 e 2019.

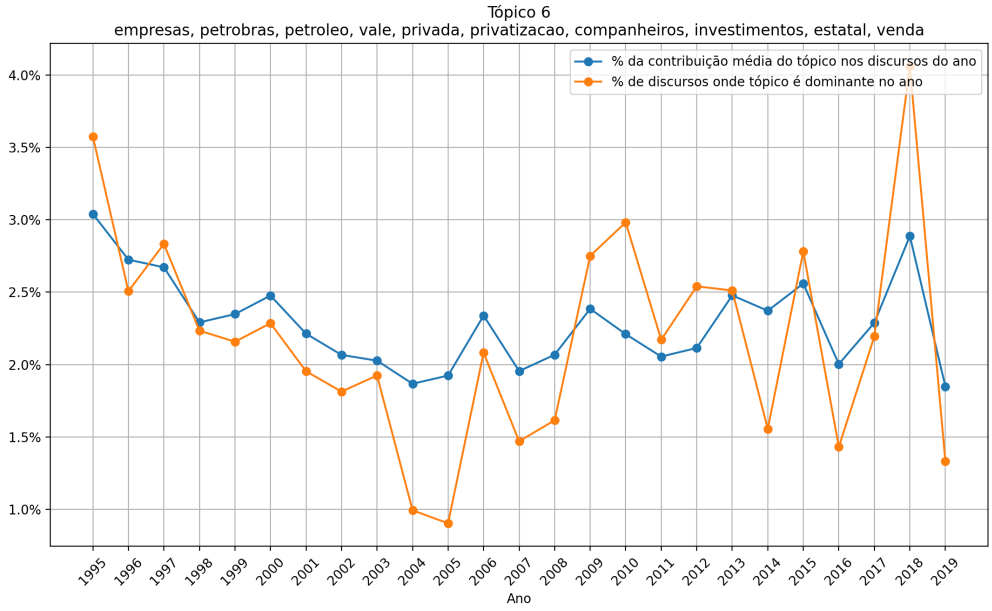

Figura B.6: Evolução do tópico 6 entre 1995 e 2019.

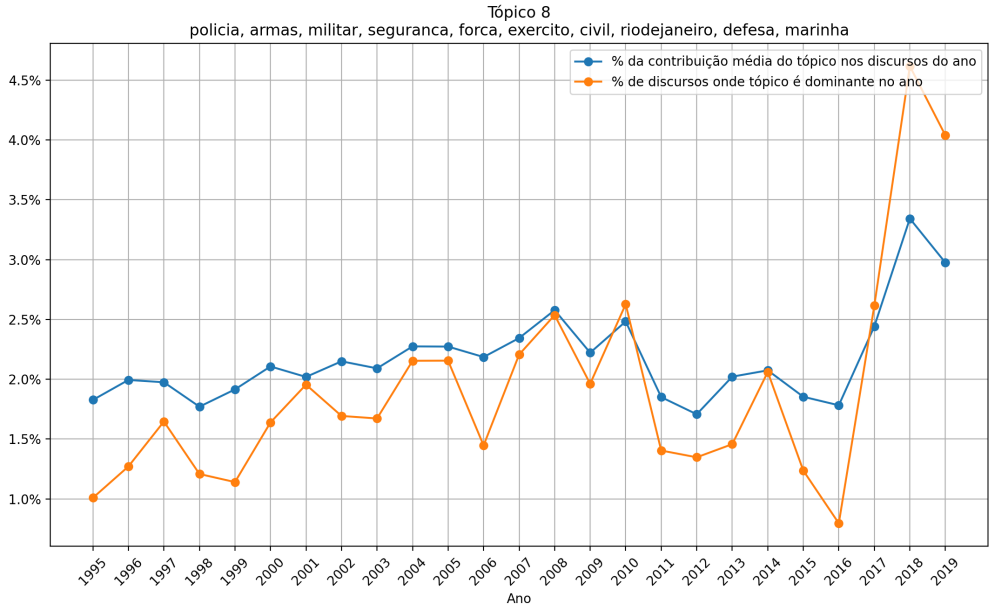

Figura B.7: Evolução do tópico 8 entre 1995 e 2019.

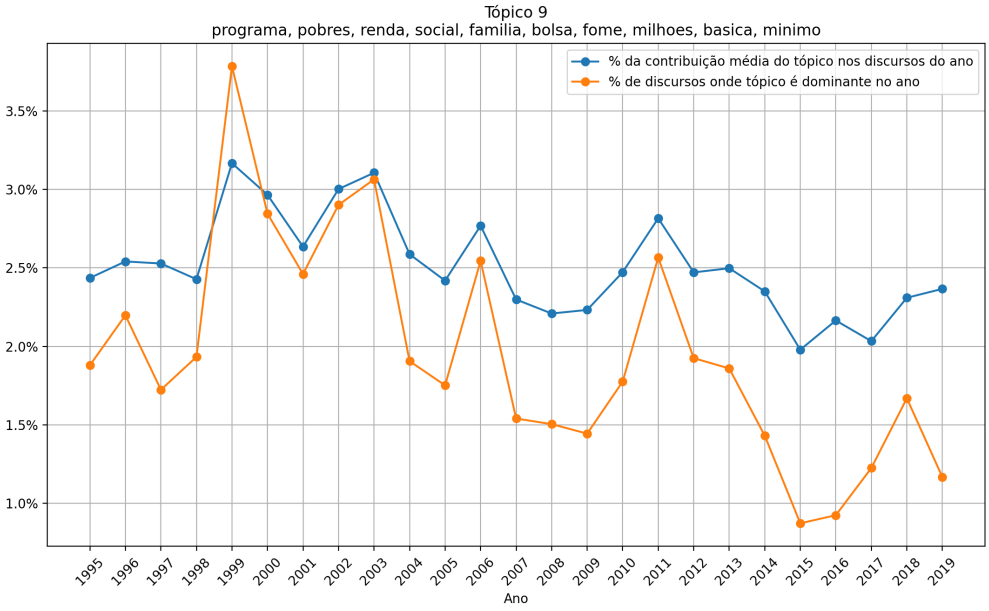

Figura B.8: Evolução do tópico 9 entre 1995 e 2019.

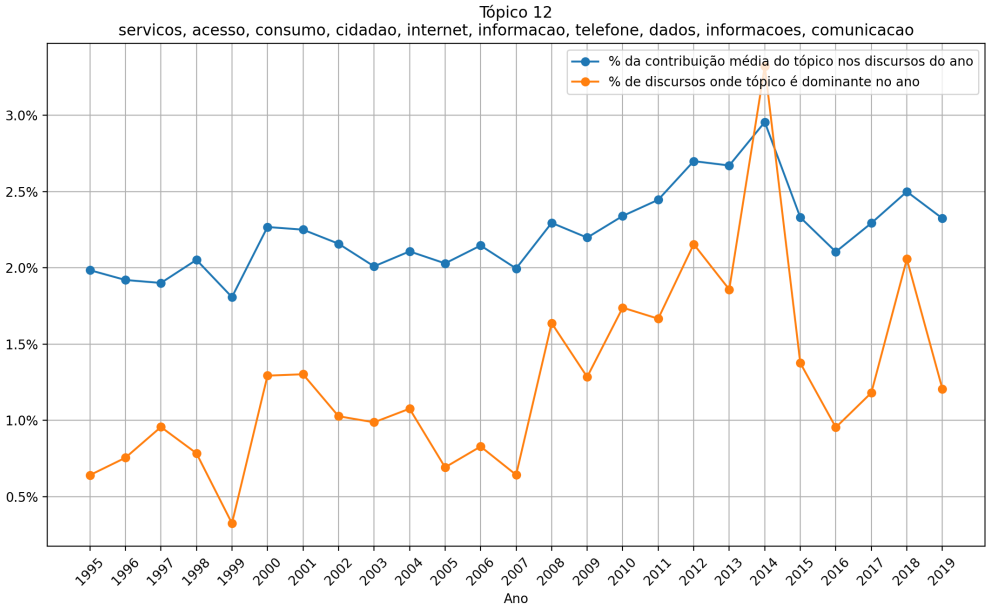

Figura B.9: Evolução do tópico 12 entre 1995 e 2019.

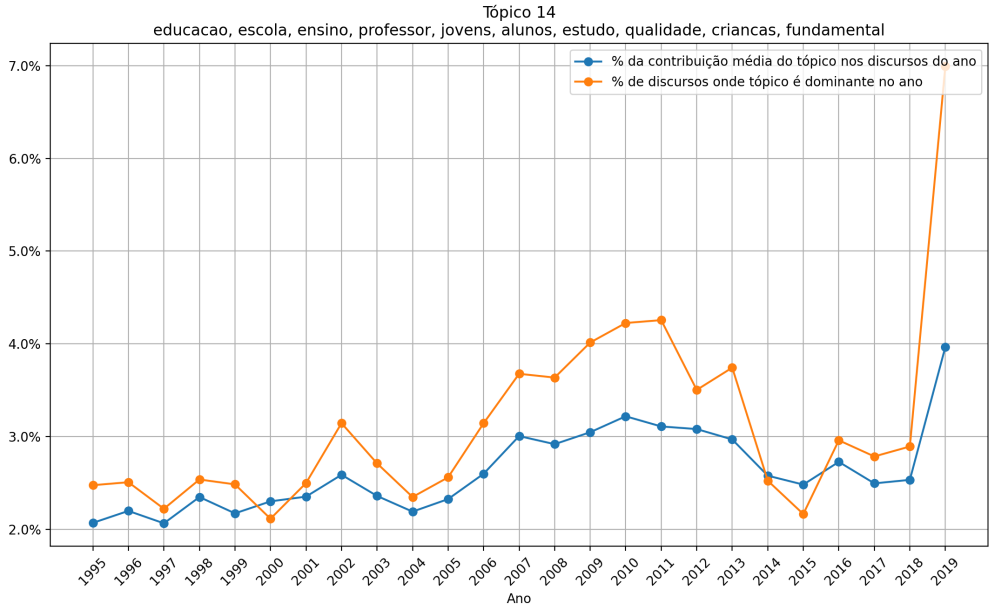

Figura B.10: Evolução do tópico 14 entre 1995 e 2019.

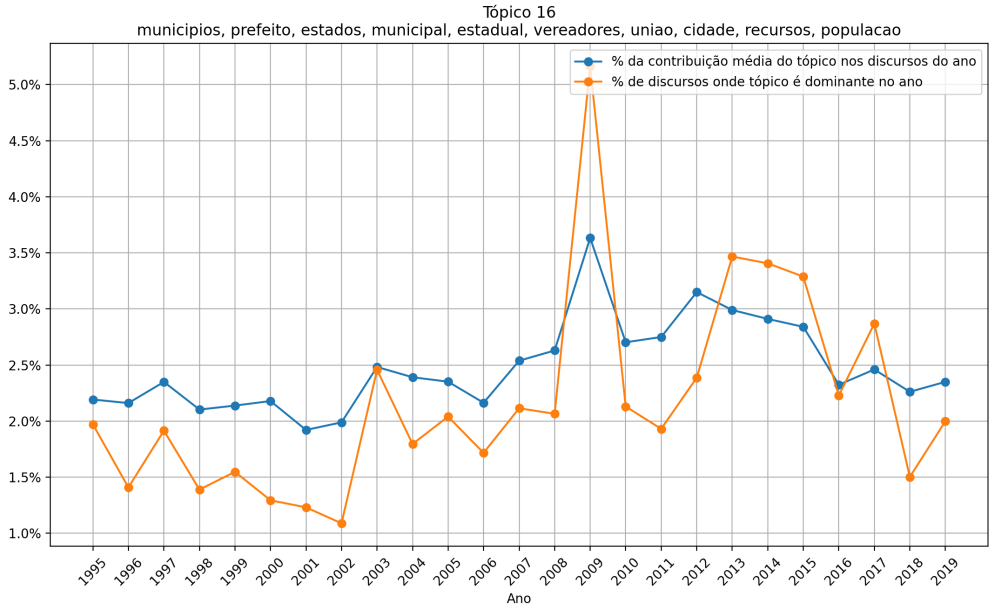

Figura B.11: Evolução do tópico 16 entre 1995 e 2019.

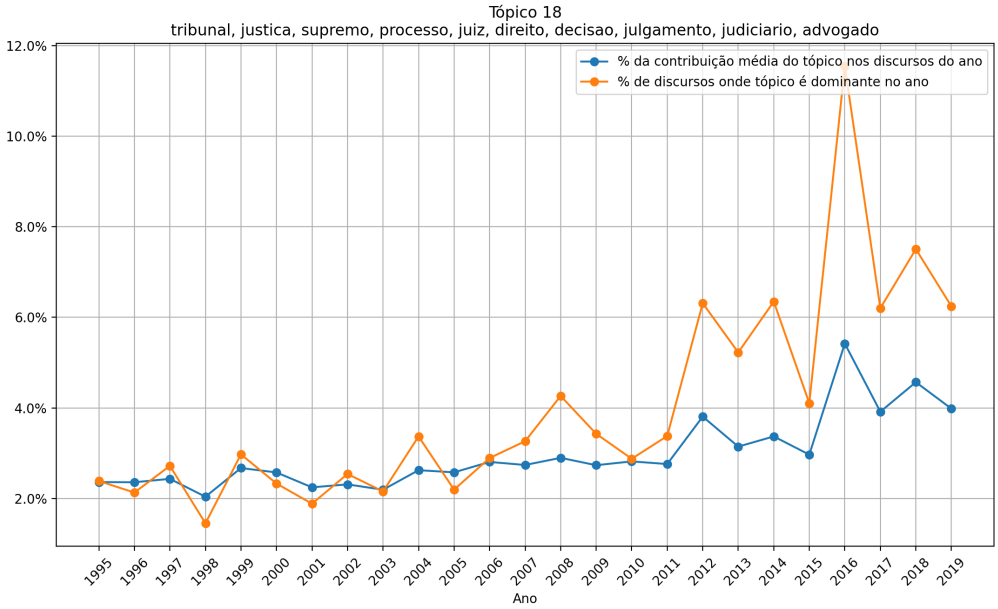

Figura B.12: Evolução do tópico 18 entre 1995 e 2019.

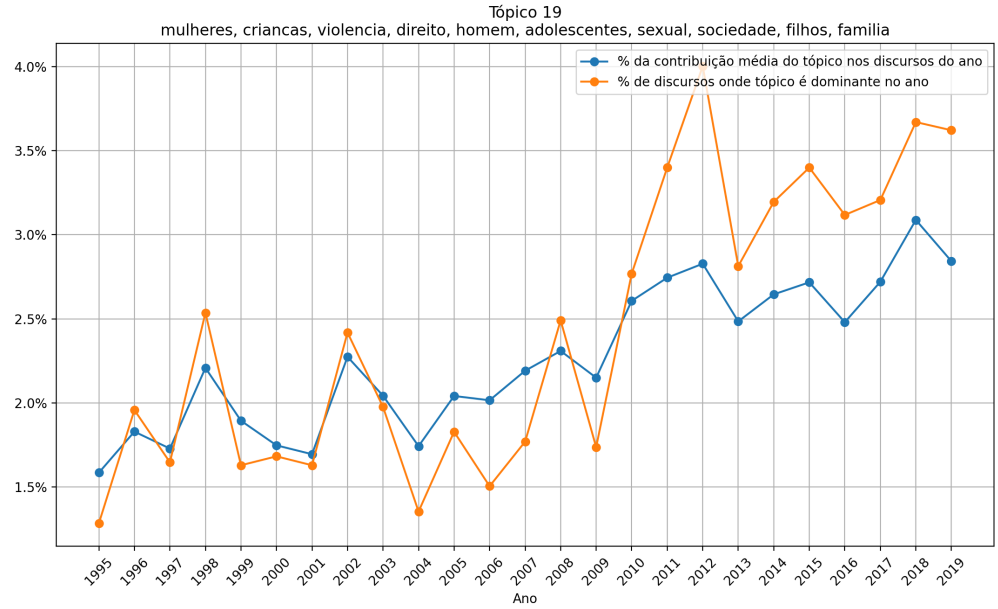

Figura B.13: Evolução do tópico 19 entre 1995 e 2019.

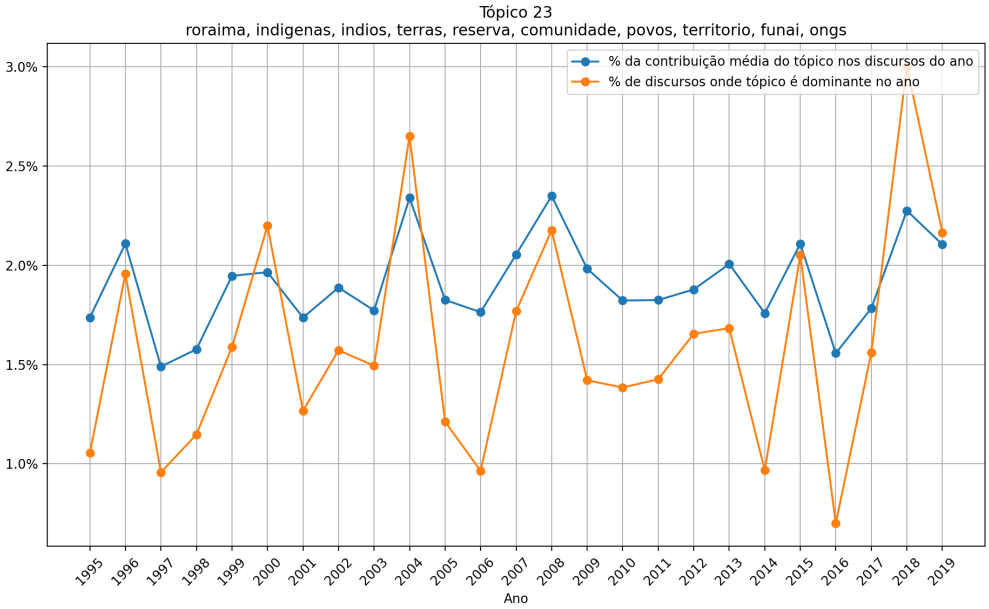

Figura B.14: Evolução do tópico 23 entre 1995 e 2019.

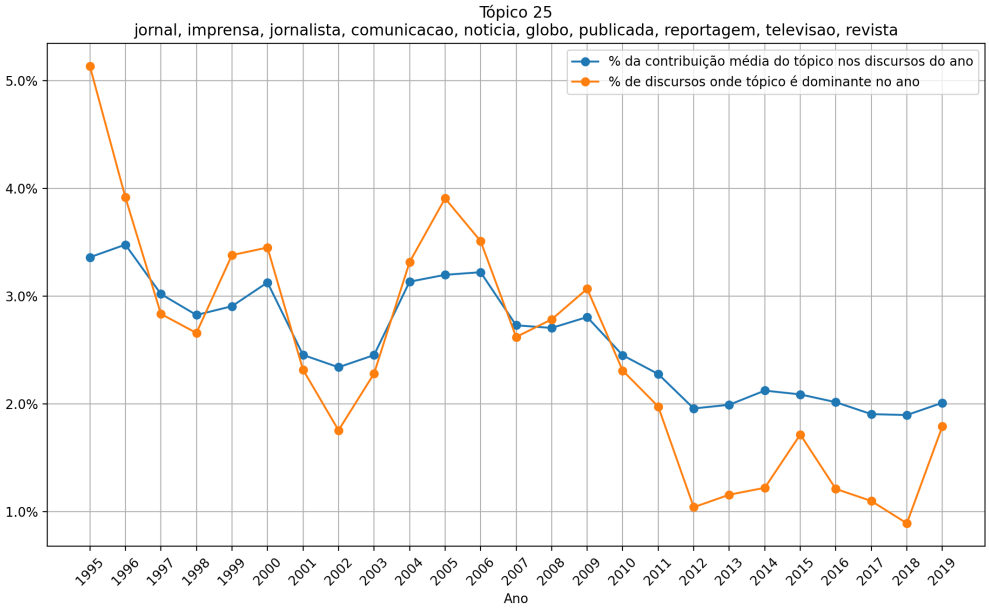

Figura B.15: Evolução do tópico 25 entre 1995 e 2019.

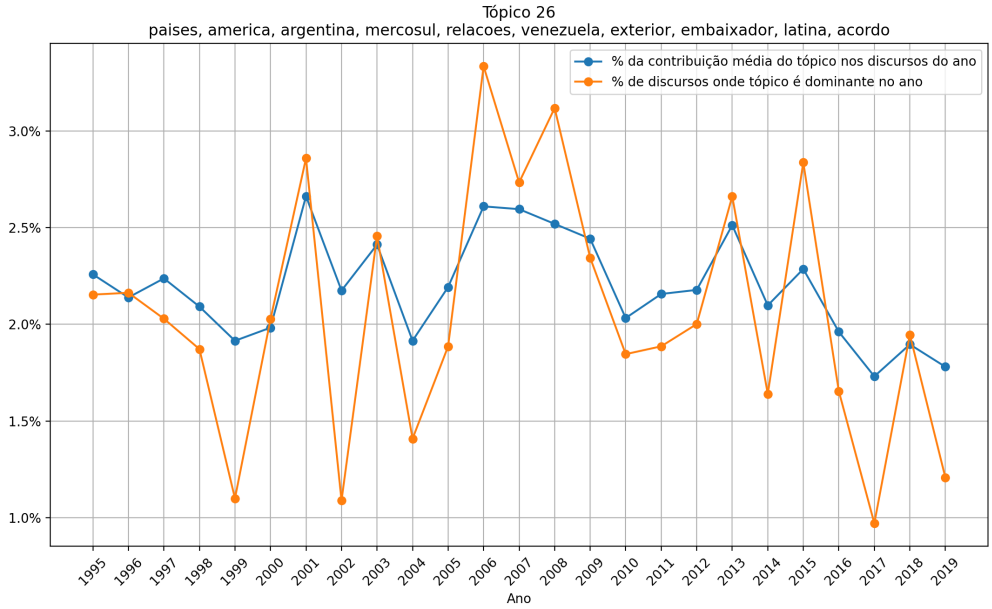

Figura B.16: Evolução do tópico 26 entre 1995 e 2019.

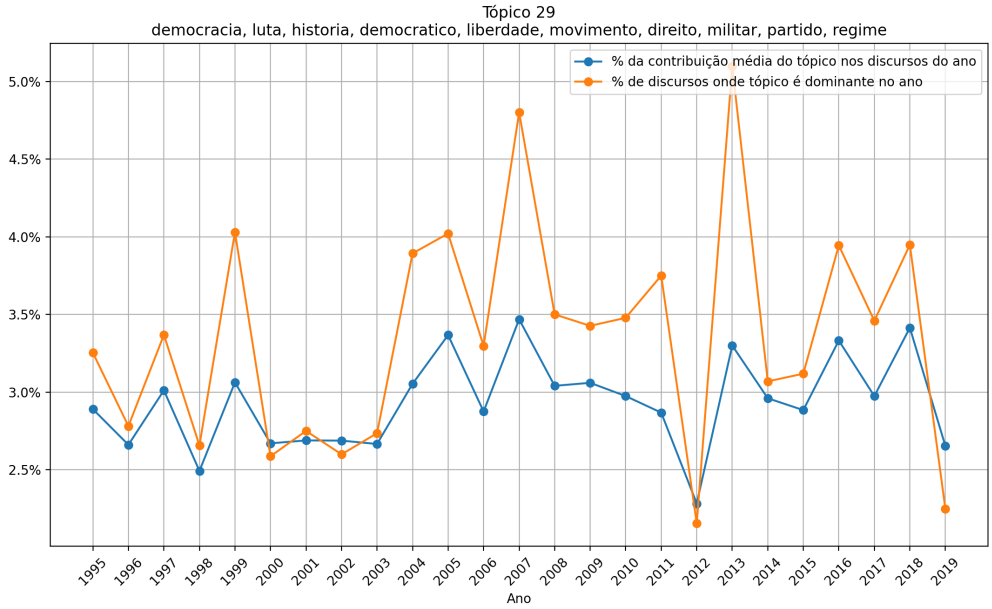

Figura B.17: Evolução do tópico 29 entre 1995 e 2019.

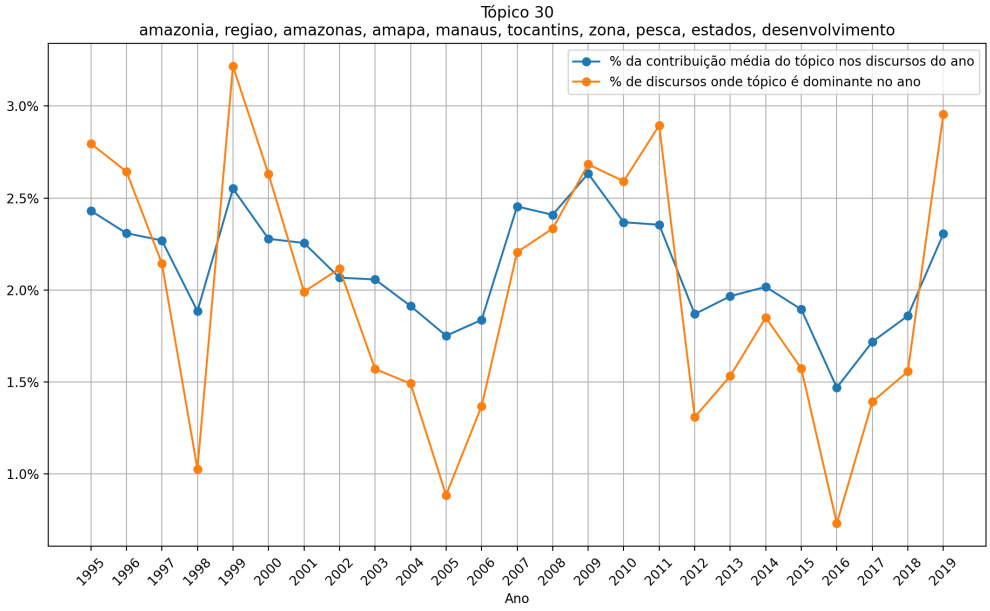

Figura B.18: Evolução do tópico 30 entre 1995 e 2019.

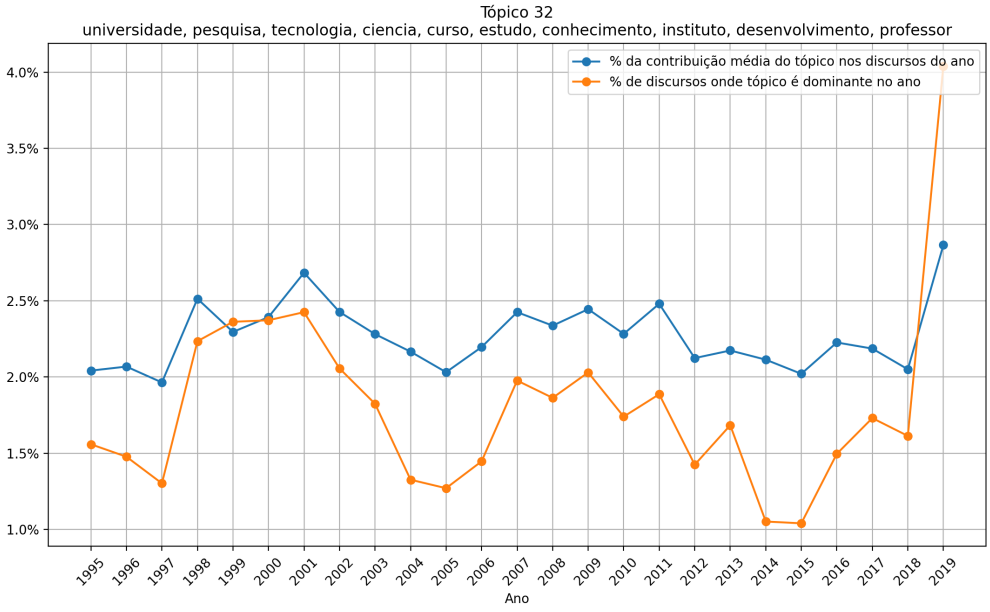

Figura B.19: Evolução do tópico 32 entre 1995 e 2019.

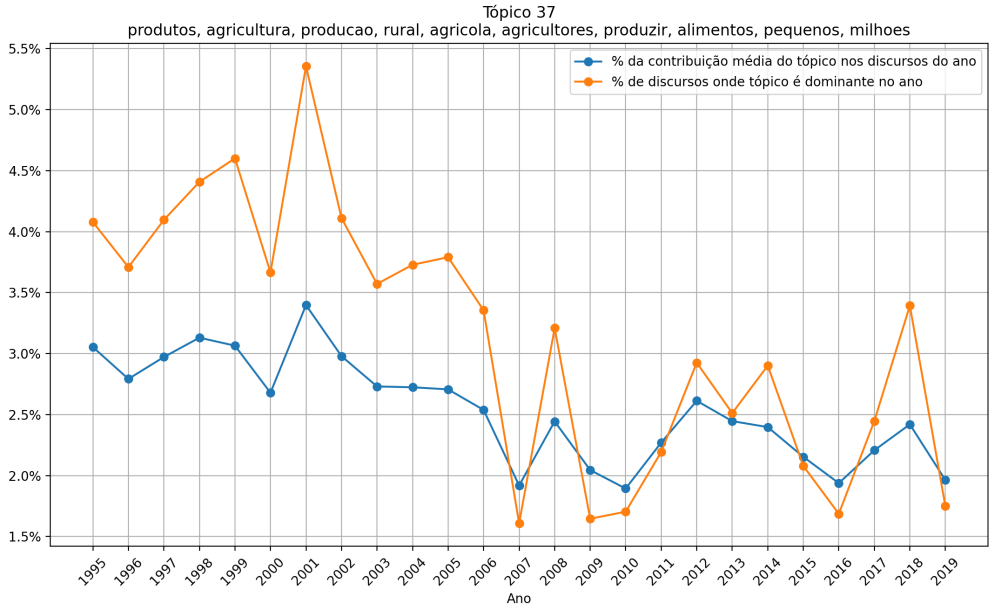

Figura B.20: Evolução do tópico 37 entre 1995 e 2019.

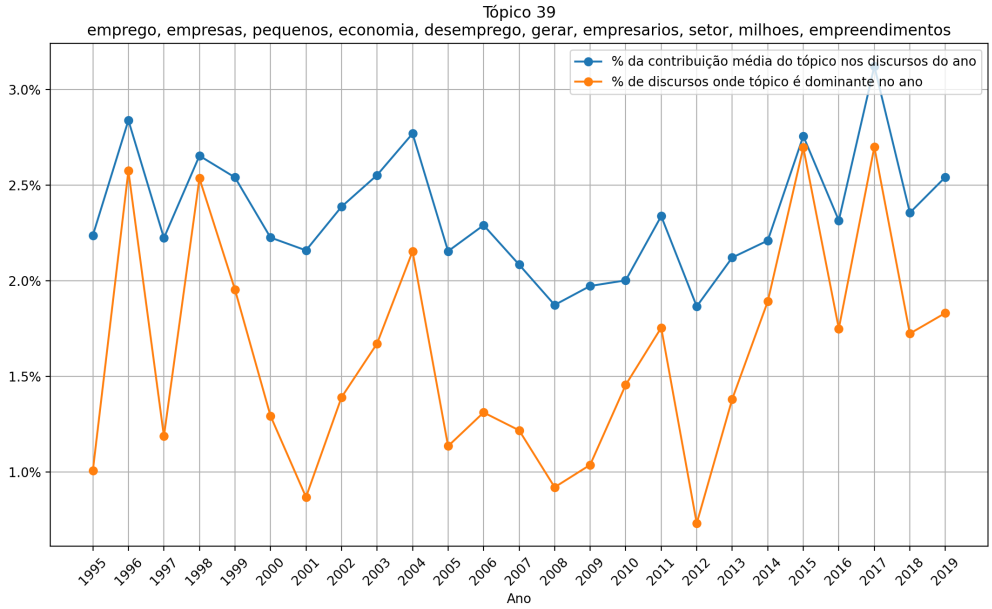

Figura B.21: Evolução do tópico 39 entre 1995 e 2019.

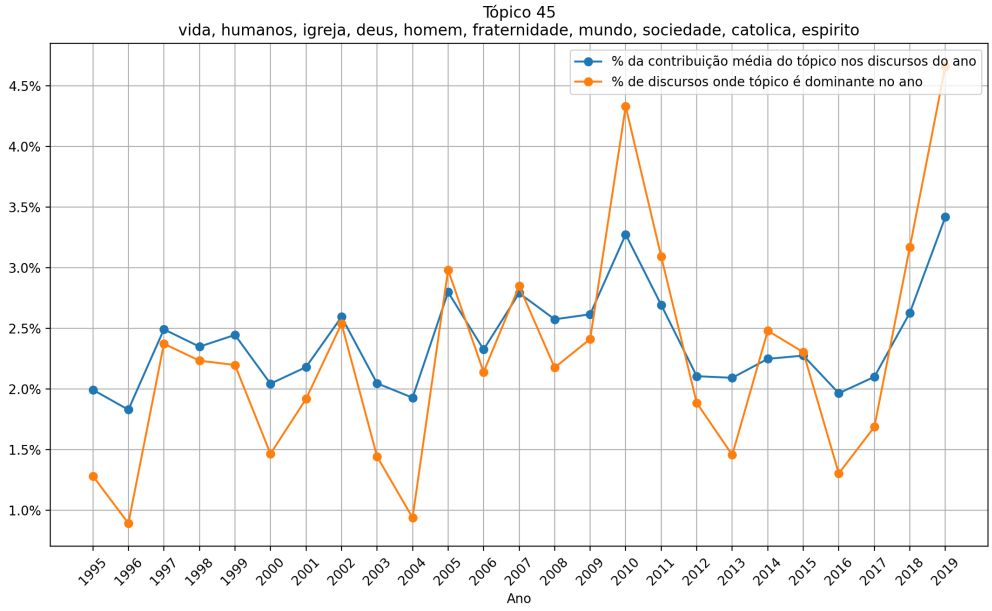

Figura B.22: Evolução do tópico 45 entre 1995 e 2019.

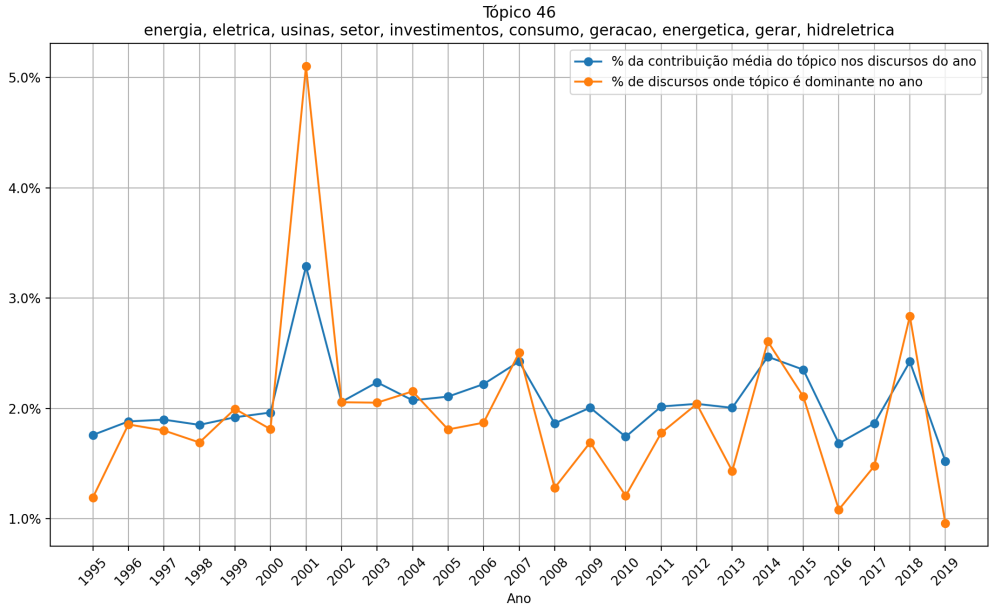

Figura B.23: Evolução do tópico 46 entre 1995 e 2019.

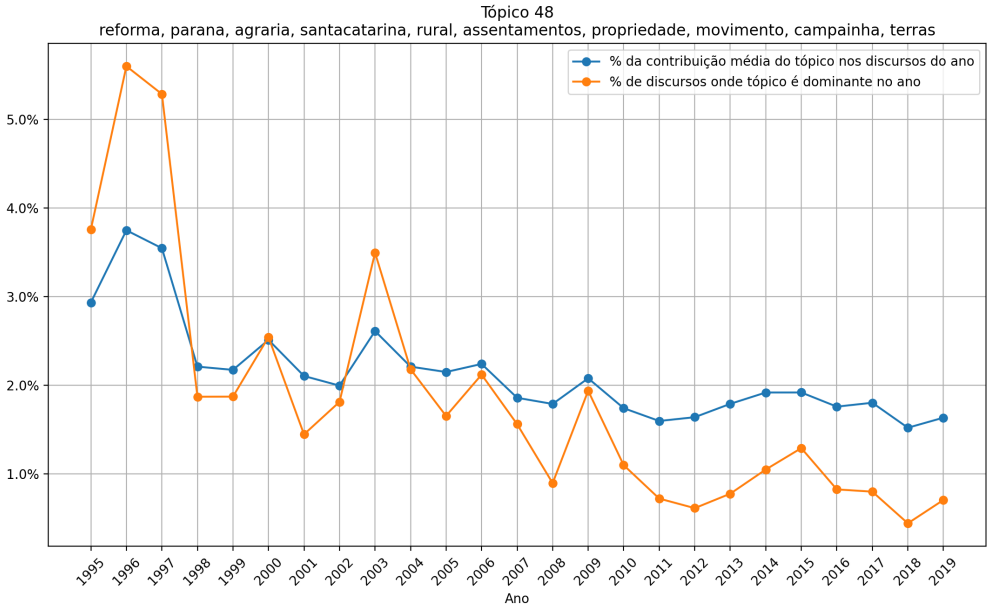

Figura B.24: Evolução do tópico 48 entre 1995 e 2019.

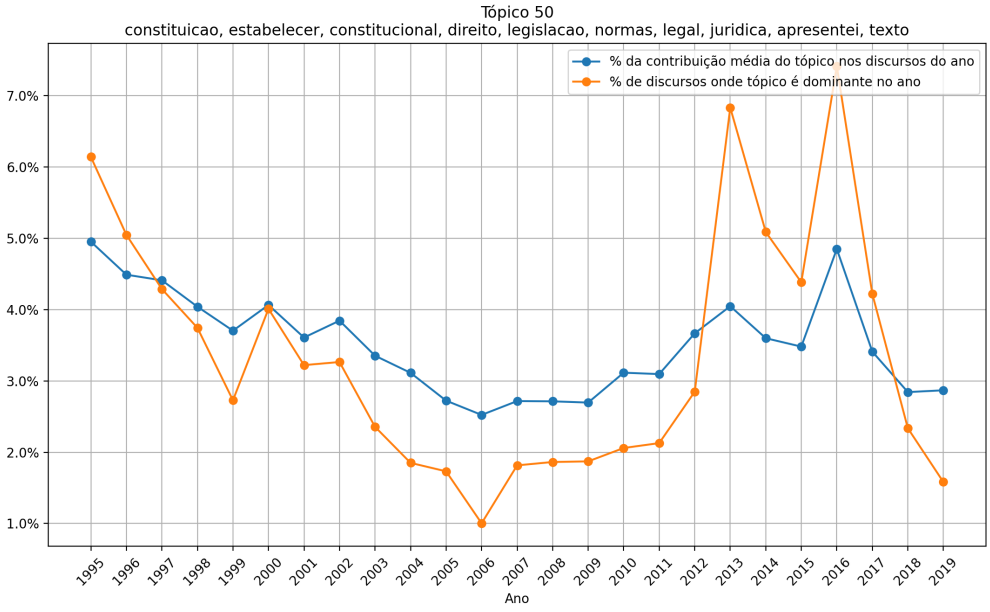

Figura B.25: Evolução do tópico 50 entre 1995 e 2019.

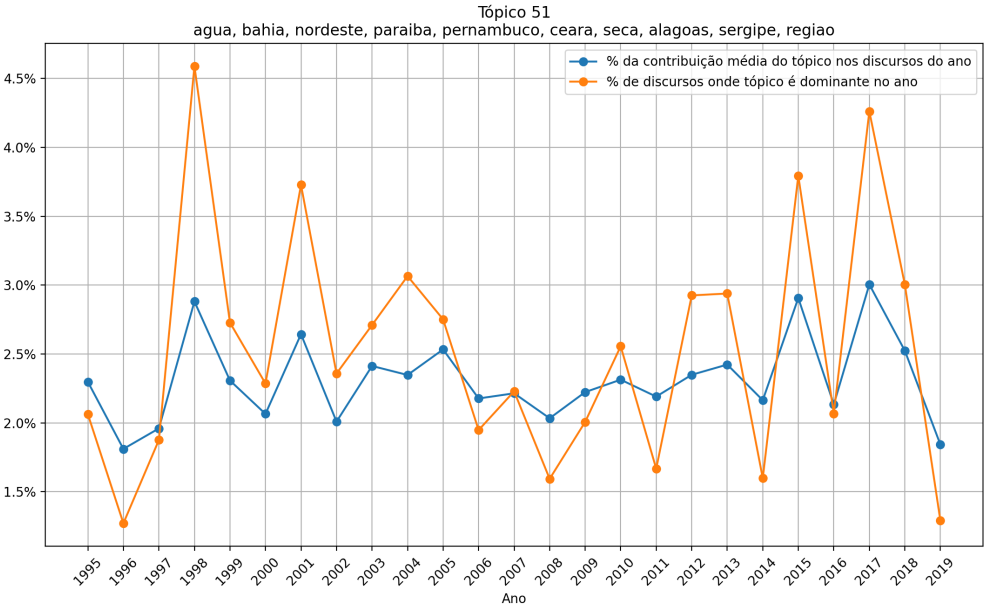

Figura B.26: Evolução do tópico 51 entre 1995 e 2019.

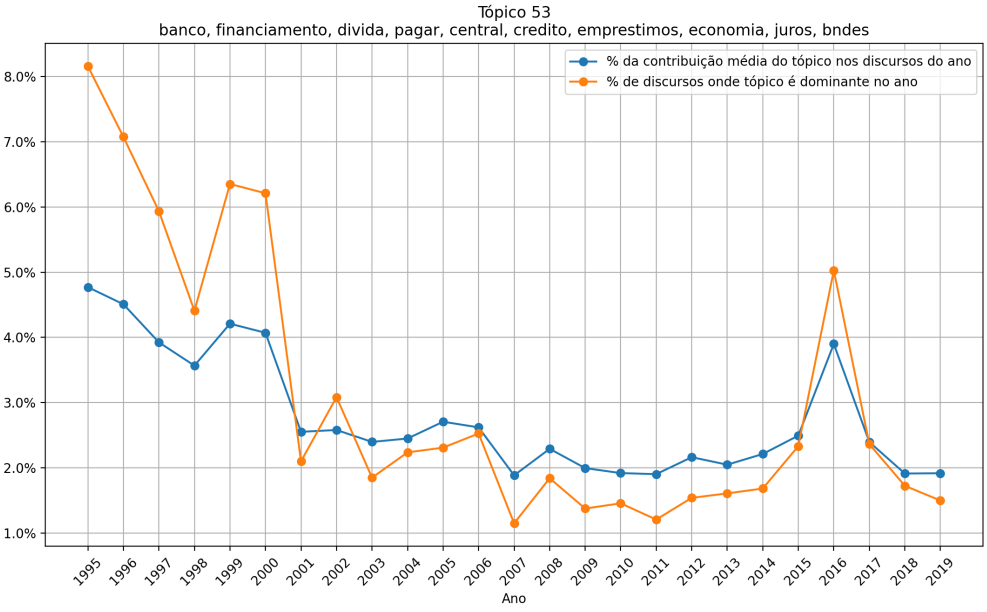

Figura B.27: Evolução do tópico 53 entre 1995 e 2019.

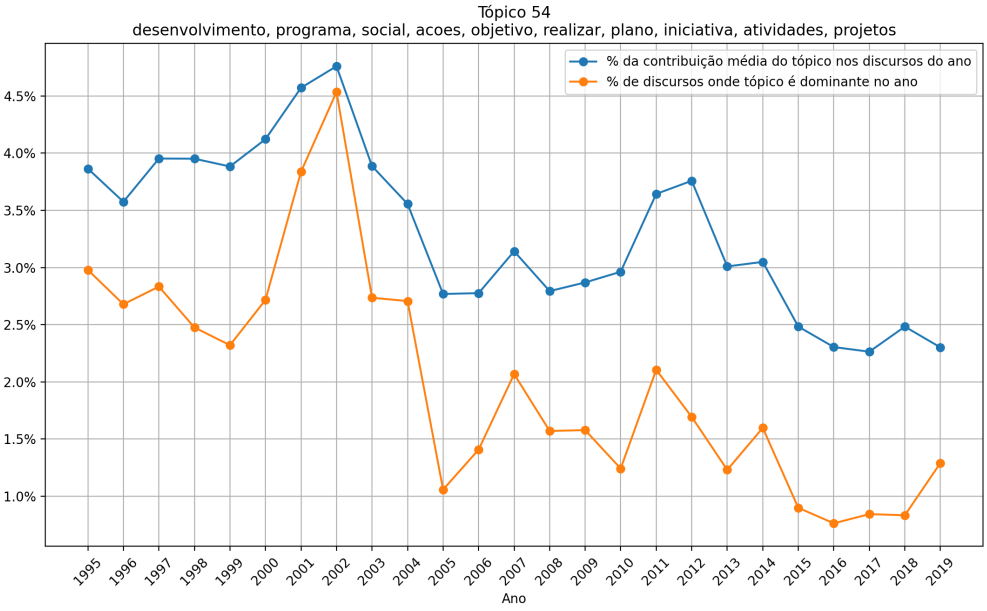

Figura B.28: Evolução do tópico 54 entre 1995 e 2019.

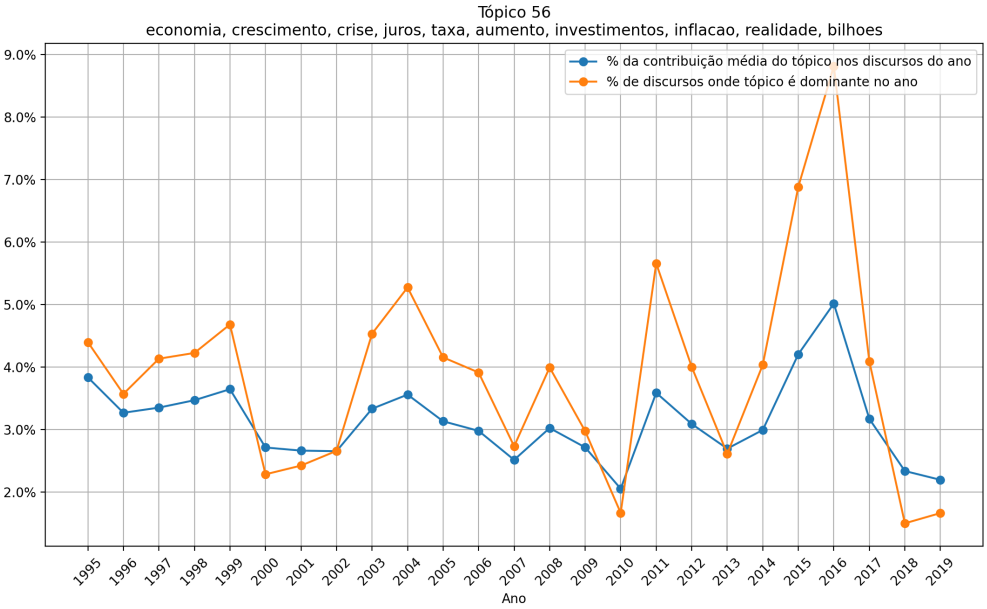

Figura B.29: Evolução do tópico 56 entre 1995 e 2019.

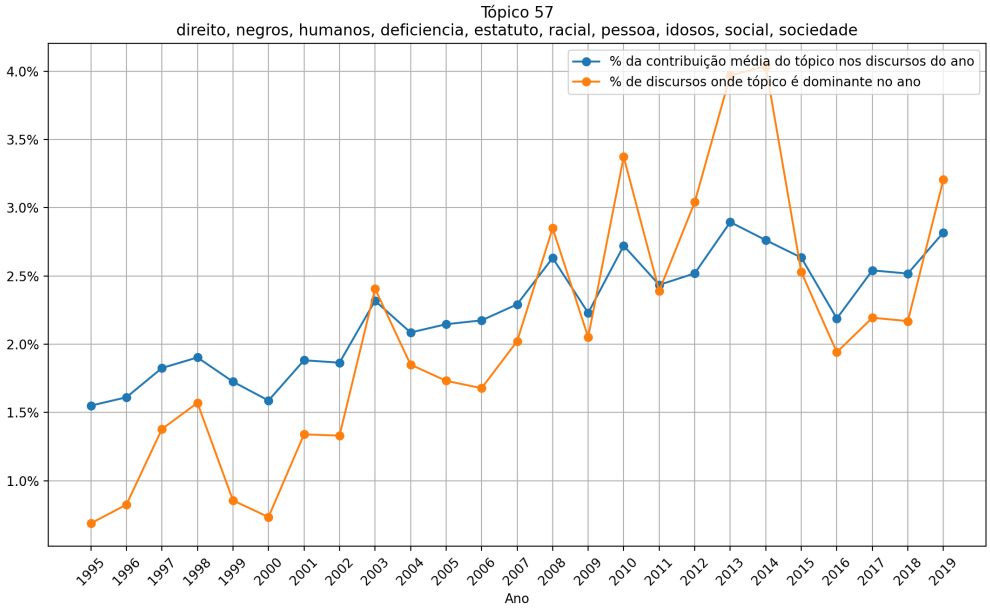

Figura B.30: Evolução do tópico 57 entre 1995 e 2019.

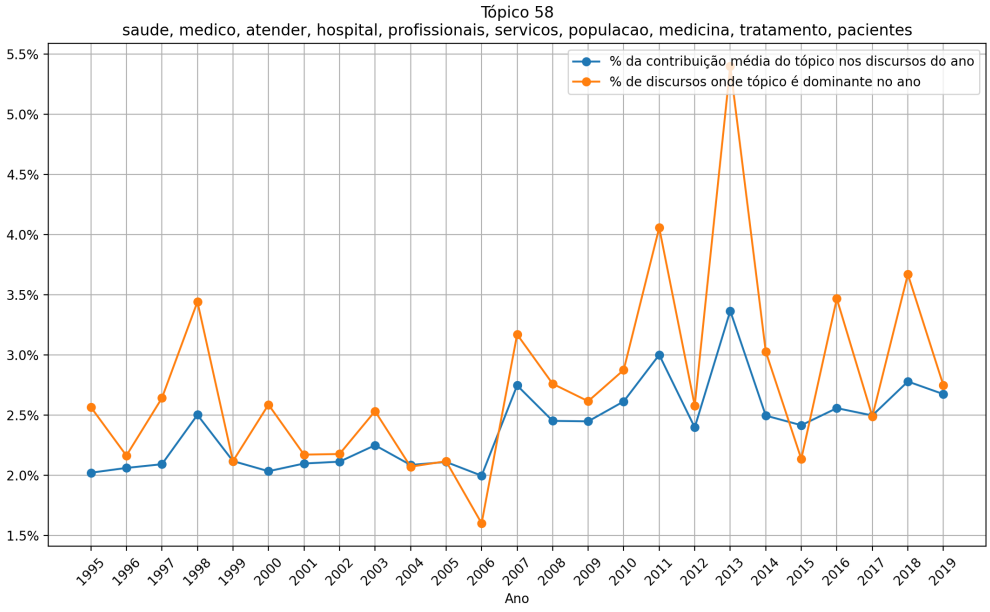

Figura B.31: Evolução do tópico 58 entre 1995 e 2019.

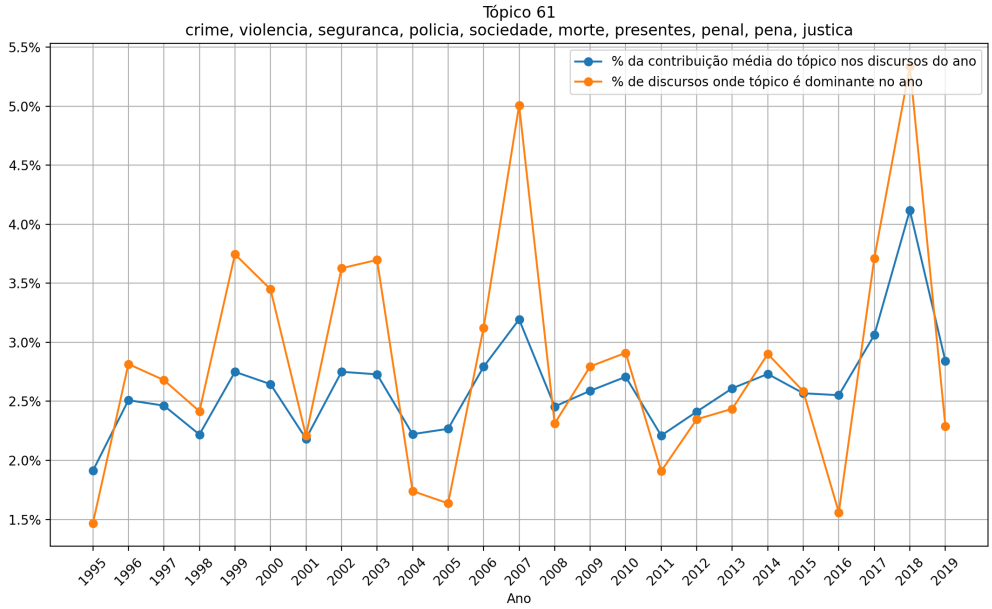

Figura B.32: Evolução do tópico 61 entre 1995 e 2019.

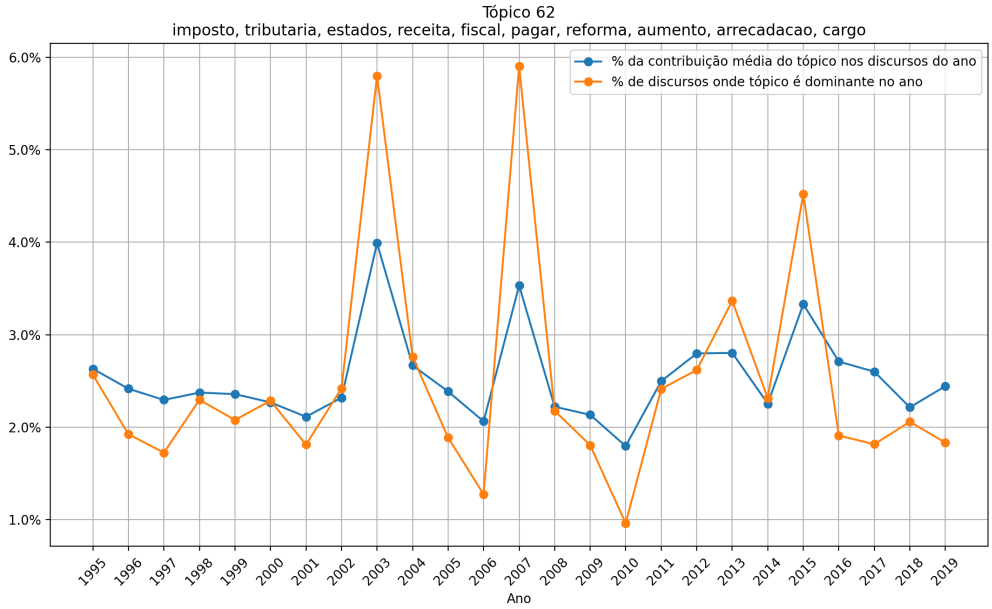

Figura B.33: Evolução do tópico 62 entre 1995 e 2019.