Gabriela Denise de Araujo

ANÁLISE DO CONTEÚDO TEXTUAL DE MENSAGENS PROVENIENTES DE REDES SOCIAIS SOBRE TEMAS DE SAÚDE BASEADO NO INTER-RELACIONAMENTO DE DOENÇAS, MEDICAMENTOS E SINTOMAS

Tese apresentada à Universidade Federal de São Paulo – Escola Paulista de Medicina para obtenção do título de Doutor em Ciências

São Paulo
2020

Gabriela Denise de Araujo

ANÁLISE DO CONTEÚDO TEXTUAL DE MENSAGENS PROVENIENTES DE REDES SOCIAIS SOBRE TEMAS DE SAÚDE BASEADO NO INTER-RELACIONAMENTO DE DOENÇAS, MEDICAMENTOS E SINTOMAS

Tese apresentada à Universidade Federal de São Paulo – Escola Paulista de Medicina para obtenção do título de Doutor em Ciências, área de Gestão e Informática em Saúde

Orientador:
Prof. Dr. Ivan Torres Pisa

São Paulo
2020

REFERÊNCIA BIBLIOGRÁFICA

ARAUJO, GD. **Análise do conteúdo textual de mensagens provenientes de redes sociais sobre temas de saúde baseado no inter-relacionamento de doenças, medicamentos e sintomas**. 2020. 117f. Tese (Doutorado em Ciências) – Departamento de Informática em Saúde, Escola Paulista de Medicina, Universidade Federal de São Paulo.

Araujo, Gabriela Denise

Análise do conteúdo textual de mensagens provenientes de redes sociais sobre temas de saúde baseado no inter-relacionamento de doenças, medicamentos e sintomas. / Gabriela Denise de Araujo. -- São Paulo, 2020.
xx, 117f.

Tese (Doutorado) – Universidade Federal de São Paulo. Escola Paulista de Medicina. Programa de Pós-graduação em Gestão e Informática em Saúde.

Título em inglês: Textual content analysis on social media messages about health issues based on the interrelationship of diseases, medications and symptoms.

1. Mineração de Dados, 2. Redes Sociais, 3. Saúde Pública, 4. Redes Complexas 5. Modelo de Tópicos.

UNIVERSIDADE FEDERAL DE SÃO PAULO
ESCOLA PAULISTA DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E INFORMÁTICA EM SAÚDE

Coordenadora da Câmara de Pós-graduação e Pesquisa da Escola Paulista de Medicina: Profa. Dra. Monica Levy Andersen, livre docente

Coordenador do programa: Prof. Dr. Ivan Torres Pisa, livre docente

Gabriela Denise de Araujo

ANÁLISE DO CONTEÚDO TEXTUAL DE MENSAGENS PROVENIENTES DE REDES SOCIAIS SOBRE TEMAS DE SAÚDE BASEADO NO INTER-RELACIONAMENTO DE DOENÇAS, MEDICAMENTOS E SINTOMAS

Banca examinadora:

Prof. Dr. Paulo Bandiera Paiva
Universidade Federal de São Paulo, Escola Paulista de Medicina

Prof. Dr. Evandro Eduardo Seron Ruiz
Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto

Prof. Dr. Jesús Pascual Mena Chalco
Universidade Federal do ABC

Dr. Roberto Silva
Universidade de Santo Amaro, Unisa

Suplente:
Profa. Dra. Claudia Galindo Novoa Barsottini
Universidade Federal de São Paulo, Escola Paulista de Medicina

Profa. Dra. Fátima de Lourdes dos Santos Nunes Marques
Universidade de São Paulo, Escola de Artes, Ciências e Humanidades

Dedicatória

À minha família.

Resumo

Introdução: Analisar e interpretar dados disponíveis na web seja em redes sociais, blogs, ou sites editoriais estabelecendo relações, identificando informação útil e relevante é um desafio computacional atual e significativo. A era da informação favoreceu a disponibilização de uma enorme quantidade de dados na web que se tornaram naturalmente uma fonte rica de informação e de evidências sobre assuntos variados, inclusive sobre saúde. **Objetivos**: O objetivo deste estudo é elaborar um arcabouço metodológico para ações de monitoramento de assuntos de saúde provenientes de redes sociais e contribuir para a produção científica de estudos de vigilância em saúde. **Métodos**: Listas com termos e frases de saúde relacionadas a doenças, sintomas e medicamentos foram utilizadas para filtrar mensagens que apresentaram pelo menos dois termos de saúde. Técnicas de mineração de dados com redes complexas e modelagem de tópicos foram utilizadas para analisar as discussões relacionadas a saúde nas redes sociais. **Resultados**: Cerca de 141 milhões de mensagens do Twitter publicadas no território brasileiro em 2017 foram coletados e 95 mil foram classificados como saúde. Dessas, 27% continham termos relacionados a doenças, 56% relacionados a sintomas e 47% a medicamentos. Foi possível explorar o relacionamento entre os termos de saúde, a força das ligações e observar temas que se destacaram medindo sua importância relativa dentro da rede. Com a técnica de modelagem de tópicos, assuntos populares foram identificados e eventos nacionais sobre campanhas de saúde foram evidenciados. Tópicos não esperados também foram notados; como tratamentos de sintomas e alimentação. **Conclusão**: A participação dos usuários expondo suas opiniões e experiências sobre saúde nas redes sociais pode auxiliar no monitoramento de alguns aspectos da saúde pública e colaborar para uma vigilância participativa, oferecendo uma percepção aos gestores de saúde de como as pessoas interagem com temas de saúde na web. Os resultados mostraram que tópicos variados relacionados a saúde são discutidos em redes sociais e as metodologias utilizadas neste estudo são eficientes para evidencia-los e torna-los úteis em termos de informação.

Palavras-chave: Mineração de Dados (D057225). Redes Sociais (D060756). Saúde Pública (D011634). Redes Complexas. Modelagem de Tópico.

Abstract

Background: Analyzing and interpreting data available on the web, whether on social networks, blogs, or editorial sites, establishing relationships, identifying useful and relevant information is a current and significant computational challenge. The information age has favored the availability of a huge amount of data on the web that has naturally become a rich source of information and evidence on various subjects, including health. **Objectives:** The purpose of this study is to develop a methodological framework to monitor general public health information from social networks and to contribute to the scientific production of health surveillance studies. **Methods:** The messages containing at least two medical terms were selected using health terms and phrases related to diseases, symptoms, and medications. Data mining techniques, complex networks, and topic modeling were used to analyze health-related discussions on social networks. **Results:** About 141 million Twitter messages published in the Brazilian territory in 2017 were collected. Around 95 thousand were classified as health-related. Of these, 27% contained terms related to diseases, 56% related to symptoms and 47% to medications. It was possible to explore the relationship between health terms, the strength of connections and their types, and to observe themes that stood out by measuring their relative importance within the network. With the topic modeling technique, popular subjects were identified, and national health campaign events were highlighted. Unexpected topics were also noted; as symptom treatments and food. **Conclusion:** Users sharing their opinions and experiences on health topics on social networks can assist in monitoring some aspects of public health and collaborate for participatory surveillance, offering a perception to health managers of how people interact with health topics on the web. The results showed that varied topics related to health are discussed in social networks and the methodologies used in this study are efficient to highlight them and make them useful in terms of information.

Keywords: Data mining (D057225). Social Network (D060756). Public Health (D011634). Complex Network. Topic Model.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

A	Matriz de Adjacência
API	Application Programming Interface
ARS	Análise de Redes Sociais
BTM	Biterm Topic Model
CRISP-DM	Cross-Industry Standard Process of Data Mining
CEP	Comitê de Ética em Pesquisa
CID-10	Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde
DECS	Descritor em Ciências da Saúde
G	Grafo, G = (V,E)
KDD	Knowledge Discovery in Databases
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MD	Mineração de Dados
NoSQL	Non Structured Query Language
OMS	Organização Mundial de Saúde
PLN	Processamento de Linguagem Natural
pLSI	Probabilistic Latent Semantic Indexing
RI	Recuperação de Informação
SRES	Sistema de Registro Eletrônico em Saúde
TEA	Transtorno do Espectro Autista
UNIFESP	Universidade Federal de São Paulo

Lista de Publicações

Araujo GD, Mancini F, Guimarães MP, Pisa IT. Applying Sentiment Analysis of Twitter's Health Messages in Brazilian Portuguese. J. Health Inform. 2018 Janeiro-Março; 10(1): 17-24.

Apoio Financeiro

Este projeto de pesquisa recebeu apoio financeiro por meio da concessão da bolsa CAPES-Demanda Social, nível doutorado, entre Agosto/2015 e Julho/2019.

Sumário

1 Introdução

1.1 Contexto

A Web 2.0[1] proporcionou uma mudança na forma de interação das pessoas com a informação online. Elas expandiram seu papel na navegação e pesquisa de conteúdos para compartilharem experiências, críticas e opiniões em ambientes que facilitam a criação de conteúdo como as redes sociais. Facebook, Twitter e outras redes sociais eletrônicas têm desempenhado um papel fundamental para o compartilhamento e a rápida disseminação de informação, tornando-se um transmissor em massa e ininterrupto de informações. Destes conteúdos são evidenciados uma grande diversidade de opiniões e perspectivas em torno de quaisquer assuntos quotidianos fora dos editoriais mais tradicionais, como, jornais, revistas, informativos televisionados, entre outros. Esse novo comportamento, uma vez organizado e tratado, reflete um padrão de inteligência coletiva criado a partir de uma grande diversidade de pontos de vista.

O cenário atual concedeu a disponibilização de uma enorme quantidade de dados na web que se tornou naturalmente uma fonte rica de informação e de evidências sobre assuntos variados. Todavia, analisar os dados de redes sociais, identificar informação relevante e extrair conhecimento são desafios computacionais complexos e significativos[2]. Tal complexidade se dá por vários fatores, entre eles, o grande volume de dados envolvidos e a manipulação de dados não estruturados (textuais). No entanto, técnicas de recuperação de informação (RI), processamento de linguagem natural (PLN) e mineração de dados (MD) tem sido amplamente utilizadas com o propósito de descobrir conhecimento em dados não estruturados apresentando resultados satisfatórios. Assim, investigações de novos conceitos e metodologias para análise desses dados têm sido impulsionados com o intuito de monitorar hábitos comunicacionais e medir diferentes características da população, incluindo aspectos de saúde[2].

Há estudos[2,3,4,5] que mostram o potencial da utilização de técnicas de mineração de dados para extração e consumo de informações úteis sobre saúde das redes sociais. Dados que podem, eventualmente, auxiliar tanto às autoridades de saúde pública (identificação dos interesses/preocupações da população) quanto a própria população na busca por informações opinativas e relatos de experiências.

Atualmente o campo da saúde pública abrange tanto o indivíduo quanto o coletivo para realização de ações de promoção, proteção e recuperação da saúde[6]. O cenário epidemiológico se caracteriza basicamente por uma recorrência de doenças agudas, outrora

erradicadas e às condições crônicas. Esse cenário é decorrente da forma de gestão das práticas sanitárias que precisam de uma organização para promover ações oportunas que ajudem na redução dos riscos à saúde e aumentem a eficiência de resposta do sistema[7]. Para promover essas ações é necessário que os órgãos responsáveis tenham informações atuais e consistentes para realizarem o monitoramento da saúde pública.

Em 2016, durante os jogos olímpicos no Brasil, o Ministério da Saúde em parceria com as Secretarias de Saúde das sedes dos jogos e outras instituições nacionais e internacionais promoveram uma ação de vigilância participativa através de um aplicativo denominado Guardiões da Saúde[8]. Os visitantes e os residentes no Brasil ajudaram no monitoramento da saúde pública informando a sua condição de saúde, sendo possível identificar surtos e antecipar ações de serviço de saúde. Essa mesma ação pode ser promovida utilizando dados relacionados à saúde publicados pelas pessoas no dia-a-dia em sua rede social de preferência/costume, criando mais oportunidades de vigilância participativa na era digital.

Contudo, a tarefa de automatizar a extração de informações em dados textuais e organizá-las é desafiadora. O grande volume de dados e informações torna as ferramentas para processamento de texto cada vez mais necessárias. A extração de informação em dados textuais está diretamente relacionada a tarefas de agrupamento e classificação. Métodos de modelagem de tópicos[9] vêm sendo utilizados nessas tarefas com objetivo de descobrir a estrutura semântica de um conjunto de documentos baseada em padrões de uso de palavras. Essa estratégia possibilita extrair automaticamente os assuntos que estão sendo tratados em uma grande quantidade de dados, facilitando a compreensão para o usuário.

Outra abordagem que auxilia no entendimento dos dados buscando o significado das palavras e suas interações baseado na estrutura semântica dos documentos é utilização de grafos. Os grafos são estruturas matemáticas flexíveis e qualquer sistema que possa ser descrito como um conjunto de objetos e relacionamentos/interações entre eles pode ser representado como um grafo[10]. É possível descobrir informações significativas sobre os relacionamentos entre as palavras, inclusive um padrão nos discursos que são realizados nas redes sociais relacionado a um determinado assunto.

1.2 Justificativa

No campo da saúde pública a vigilância está diretamente relacionada às práticas de atenção e promoção da saúde dos cidadãos, por isso é uma área fundamental em qualquer país. O conceito de vigilância corresponde, principalmente, à observação contínua de tendência, incidência, análise e disseminação de doenças, mediante a coleta sistemática, avaliações de informes e demais dados relevantes[11]. A vigilância reconhece as principais doenças por notificações compulsórias e investiga epidemias que ocorrem ao longo do tempo em territórios específicos, auxiliando na situação de saúde da população identificando as doenças recorrentes e promovendo ações específicas para contornar situações de risco[12].

Desenvolver sistemas de vigilância envolve o acesso a um conjunto de informações relacionadas à saúde, geralmente relativas à morbidade, à mortalidade e ao estado imunitário e nutricional da população. Quanto maior a quantidade de dados de saúde úteis que possam auxiliar os processos de vigilância em saúde, melhor para identificação de alterações oportunas do padrão epidemiológico de doenças, detecção e monitoramento provendo respostas às emergências em saúde pública.

As redes sociais têm proporcionado um montante de conteúdos na web que amplia as possibilidades de obtenção de informações também relacionadas a saúde. A mineração de dados de saúde do Twitter, por exemplo, possibilita o rastreamento de relatos médicos ao longo do tempo e com geolocalização. A participação dos usuários expondo suas opiniões e experiências sobre saúde nas mídias sociais pode auxiliar no monitoramento de alguns aspectos da saúde pública e colaborar para uma vigilância participativa, oferecendo uma percepção aos gestores de saúde de como as pessoas interagem com temas de saúde na web.

Nesse contexto, a proposta deste trabalho é elaborar um arcabouço metodológico para ações de monitoramento de assuntos sobre saúde provenientes de redes sociais fomentando a produção científica de informações sobre vigilância em saúde utilizando as redes sociais.

Os resultados obtidos neste estudo seguem integrados aos estudos correlatos do grupo de pesquisa Saúde 360 (http://saude360.unifesp.br).

1.3 Objetivos

O objetivo geral desta tese é investigar o uso de técnicas de mineração de dados textuais, redes complexas e modelagem de tópicos para extrair informação sobre saúde a partir de mensagens publicadas em redes sociais e identificar relacionamentos entre os achados que possam descrever características da opinião da população brasileira e auxiliar em ações de monitoramento de assuntos de saúde.

Este estudo tem como norteadoras as seguintes questões:

- Quais são as doenças, sintomas ou medicamentos mais discutidos nas redes sociais no Brasil?
- É possível auxiliar a saúde pública do Brasil com dados de redes sociais?
- Quais fenômenos característicos da população brasileira podem ser descritos relacionando doenças, medicamentos e sintomas encontrados em mensagens publicadas em redes sociais em determinados períodos?

1.3.1 Objetivos Específicos

Os objetivos específicos são:

1. Desenvolver um método de seleção de mensagens de redes sociais relacionadas à saúde utilizando listas de termos de saúde associados à doenças, sintomas e medicamentos.
2. Descrever e analisar redes de co-ocorrência de palavras construídas a partir das mensagens selecionadas, identificando os relacionamentos entre doenças, medicamentos e sintomas.
3. Identificar os tópicos que sejam significativos para o entendimento dos assuntos que estão sendo tratados nas mensagens através do processo de modelagem de tópicos. Explorar os fenômenos característicos encontrados nos tópicos providos das mensagens do Twitter.

1.4 Organização do Documento

Esta tese está organizada nos seguintes capítulos:

- Capítulo 1: O capítulo corrente, contextualiza a pesquisa, descreve os objetivos, geral e específicos, e apresenta a justificativa do trabalho;
- Capítulo 2: Revisão da literatura que expões trabalhos similares a presente pesquisa e os aspectos conceituais;
- Capítulo 3: Descrição dos métodos utilizados para atingir os objetivos;
- Capítulo 4: Expõe os resultados alcançados pela pesquisa;
- Capítulo 5: Discussão sobre os resultados apresentados no capítulo 4;
- Capítulo 6: Conclusão do trabalho, apresentando as principais contribuições e trabalhos futuros.

2 Trabalhos correlatos, conceitos e definições

2.1 Revisão da Literatura

Foi realizada uma revisão da literatura do tipo análise exploratória, buscando artigos publicados no período de 2012 a 2019, utilizando as palavras-chaves tópicos de saúde, saúde pública, grafo, redes complexas, modelo de tópicos, mídia social e Twitter. A pesquisa foi realizada nas bases de dados PubMed e Periódicos Capes e a *query* de busca foi: "*("health topics" OR "public health") AND ("graph" OR "social network analysis" OR "topic model") AND ("twitter" OR "social media")*". A busca retornou um total de 34 artigos no PubMed e 1.588 na base de dados periódicos CAPES. Os critérios de exclusão considerados foram: artigo que não apresentasse a versão completa para leitura; artigo cujo tema não estava relacionado à saúde; artigo em outras línguas que não fosse inglês ou português, estudos de redes complexas constituídas por indivíduos; estudos de revisão sistemática e artigos de periódicos não revisados por pares. Após leitura e análise dos títulos e resumos, aplicando os critérios de seleção, 18 artigos foram selecionados para leitura completa.

Sanders-Jackson et al.[13] (2013) apresentaram em seu artigo um método de análise de dados de redes sociais que pode ser aplicado em grandes conjuntos de dados relacionadas ao tabaco e à saúde. No artigo os autores descrevem uma metodologia linguística de comparação da frequência relativa de termos e apresentam uma análise de rede semântica comparando a força das conexões entre as palavras. Com essa análise apresentam vários aspectos da rede que ajudaram a identificar frases-chaves ou conceitos relacionados que são importantes nas conversas online. Esse estudo mostrou que as pessoas estão falando sobre questões relacionadas ao tabaco no Twitter e que utilizar análise de redes semânticas pode auxiliar na caracterização de conversas online, assim intervenções futuras podem aproveitar as redes sociais e os principais eventos atuais para aumentar a conscientização sobre questões relacionadas ao tabagismo.

Em 2014, Paul e Dredze[14] realizaram um estudo para identificar informações gerais de saúde pública em milhões de mensagens do Twitter (tweets) relacionadas à saúde. Utilizaram duas abordagens de modelos de tópico a Latent Dirichlet Allocation (LDA) e Ailment Topic Aspect Model (ATAM). O ATAM é um modelo de tópico especializado para conteúdos de saúde criado pelos próprios autores e apresentou em um dos experimentos citados no artigo uma coerência maior na descoberta de tópicos de saúde comparado ao

LDA. Uma variedade de doenças e problemas de saúde foram encontrados nos tweets analisados. Realizaram uma avaliação da qualidade das duas abordagens comparando tendências temporais e geográficas com dados de fontes externas. Os autores também descrevem um fluxo de dados para coletar tweets relacionados a saúde e apontam que filtrá-los apenas com palavras-chaves não é suficiente, então, fazem uso de técnicas supervisionadas de aprendizado de máquinas para classificar os tweets previamente filtrados pelas palavras-chaves. Concluem que as redes sociais podem ser consideradas como fontes complementares as ferramentas de vigilância existentes, com algumas vantagens exclusivas, como por exemplo, acesso quase em tempo real a informação.

Ainda em 2014, um estudo foi realizado[15] para avaliar se o conteúdo das redes sociais e de buscas na web poderia ser utilizado para criar um modelo preditivo confiável de obtenção de *feedback* instantâneo sobre a incidência da gripe em Portugal. Um dos poucos trabalhos em língua portuguesa, os autores descrevem uma abordagem utilizando técnicas de aprendizado de máquinas para classificar os tweets que se referiam ao contexto gripe. Para avaliar o desempenho da abordagem utilizada foi feita uma comparação com os resultados epidemiológicos do *Influenzanet*, um projeto de monitoramento de saúde relacionado à gripe em Portugal. Apesar de obterem bons resultados de previsão de gripe, sugerem uma adaptação na fase inicial do modelo de regressão utilizado para obter melhores resultados.

Beykikhoshk et al.[16] (2015) realizaram um estudo para aprender sobre a população afetada por Transtorno do Espectro Autista (TEA), utilizando como fonte de dados o Twitter. Mais de 11 milhões de tweets foram coletados e diversos experimentos foram realizando utilizando várias técnicas como: distribuição de Zipf, frequências de palavras, análise de hashtag, análise de discurso (Part-of-speech) e métodos de classificação (lógica de regressão e Naive Bayes). Através dessas técnicas foi possível examinar uma série de aspectos linguísticos e semânticos das mensagens, bem como descobrir diversos tópicos relacionados a comportamento, preocupações e necessidades dos indivíduos com esse transtorno.

Xiang Ji et al.[17] (2015) publicaram um estudo usando mensagens do Twitter para monitorar a preocupação do público com relação a problemas comuns de saúde e identificar picos de preocupações realizando um controle de tendências. Esse estudo utilizou uma abordagem de classificação de sentimentos em duas etapas. Primeiro foi feita uma distinção de tweets de pessoas e de empresas utilizando técnicas de mineração de texto, após isso classificam o sentimento das mensagens de pessoas em

negativas e não negativas. Para isso foi utilizado três modelos diferentes de aprendizado de máquinas supervisionados. Os resultados experimentais mostraram que a abordagem de classificação de sentimento em duas etapas foi capaz de rotular os tweets em negativos, não negativos e de notícias com uma boa precisão. Porém, foi necessário um conjunto de dados rotulados manualmente por humanos para treinar os algoritmos.

Tangherlini et at.[18] (2016) coletaram e analisaram dados de dois sites populares de rede social dedicados aos pais/mães para entender o que é falado sobre vacinação e analisar a estrutura narrativa persuasiva que é discutida nesses sites. Para isso, utilizaram métodos probabilísticos para determinarem os tópicos de discussões, desenvolveram um modelo generativo de narrativa estatística-mecânica para extrair automaticamente as histórias e fragmentos subjacentes de milhões de postagens. Agregaram as histórias em um grafo de estrutura narrativa abrangente. Encontraram uma forte estrutura narrativa relacionada à busca de isenções e uma cultura de desconfiança do governo e das instituições médicas.

Chen et al.[19] (2016) propõem uma abordagem utilizando dois modelos de tópicos temporais para capturar estados epidemiológicos (ex.: “saudável”, “infectado” e “exposto”) ocultos dos usuários em seus tweets relacionados a gripe e agregar essas declarações por região para ter uma melhor estimativa das tendências. Os autores utilizaram o Twitter como fonte de dados e validaram a abordagem com um estudo de caso usando mensagens em espanhol de vários países na América Latina. Concluem nesse artigo que a abordagem utilizada obtém consistentemente melhores previsões de contagens de casos de gripe que abordagens que utilizam apenas vocabulário, e produzem melhores previsões de pico de gripe do que o Google Flu.

Um estudo[20] (2017) explorou o método de modelagem de Grafo de Adjacência de palavras para analisar um grande conjunto de dados sobre preocupações com a saúde de pessoas com epilepsia. Aplicado a um conjunto de dados complexo e numeroso, esse método identificou com êxito grupos de tópicos relacionados à epilepsia, possibilitando a separação de tópicos relevantes e potencialmente não relevantes. Os dados utilizados nesse estudo foram de um serviço norte-americano gratuito de mensagens curtas (SMS) e perguntas e respostas na web, conhecido como Chacha. Apenas pela análise do Grafo de Adjacência, aplicando técnicas de particionamento, foi possível visualizar a variedade de tópicos relacionados à epilepsia, analisar a amplitude e a profundidade dos tópicos e grupos de usuários específicos. Esse método pode ser útil para identificar questões de pesquisas orientadas ao paciente a partir de grandes conjuntos de dados.

Machey e Kalyanam[21] (2017) realizaram um estudo que investigou o papel da internet e das mídias sociais no marketing ilegal de uma medicação para dor conhecida como fentanil. Foram coletados e analisados dados do Twitter filtrados pela palavra-chave “fentanil”. Uma abordagem de aprendizado de máquina conhecida por Biterm Topic Model (BTM) foi utilizada para detectar os tópicos presentes no corpus e através dessa técnica foi possível descobrir mensagens relacionadas a comercialização ilegal da medicação e encontrar sites (hiperlinks relacionados nas mensagens) associados a anúncios e farmácias online que vendiam o medicamento e outras substâncias controladas. Em 2018, um ano depois, os mesmos autores publicaram um estudo[22] desenvolvido durante o evento HHS Opioid Code-a-Thon nos Estados Unidos apresentando a mesma abordagem BTM para detectar com precisão o marketing e a venda ilícita de opióides via Twitter. Dessa vez os resultados foram convertidos em ação e os autores também desenvolveram um protótipo com o objetivo de detectar, classificar e relatar tweets ilícitos de farmácias online que vendem substâncias controladas ilegalmente para a Food and Drug Administration e para a Agência de Repressão às Drogas dos EUA.

Um artigo[23] (2017) publicado por pesquisadores do Brasil apresentou um estudo transversal que analisou a associação entre variáveis obtidas no monitoramento das mídias sociais Facebook, Twitter, Instagram, Flickr, Youtube e Blog, e as variáveis obtidas pelos indicadores tradicionais da Diretoria de Vigilância Epidemiológica de Santa Catarina. Os assuntos monitorados foram: dengue, chikungunya, zika e microcefalia. Esse monitoramento foi restrito ao estado de Santa Catarina e realizado no período de um mês para coincidir com as semanas epidemiológicas da Vigilância Epidemiológica. Para verificar a associação estatística entre os dados foi utilizado o Coeficiente de Correlação de Pearson. As variáveis analisadas apresentaram uma alta correlação indicando que o monitoramento e a mineração de conteúdo compartilhado nas redes sociais podem ser um bom indicador para gestores da área de saúde.

Com a intenção de aprimorar a vigilância sazonal de gripe, um estudo[24] apresentou uma extensão metodológica para extrair de forma eficiente drogas/medicamentos amplamente consumidos durante a gripe sazonal, e tópicos relacionados ao uso desses medicamentos. Primeiramente, selecionaram todas as mensagens do Twitter que continham a palavra “gripe” e para identificar os tweets que mencionavam medicamentos, utilizaram o Sistema Unificado de Linguagem Médica (UMLS). Após isso, utilizaram o classificador Support Vector Machine (SVM) para

descobrir quais tweets indicavam que o usuário consumiu ou pretende consumir os medicamentos mencionados em seus tweets. Por fim, utilizaram a abordagem de modelagem de tópico para investigar os assuntos discutidos sobre os medicamentos encontrados. Os resultados apresentados mostraram que o medicamento que foi amplamente discutido foi a vacina de gripe. Também foi observado que certas pessoas preferem remédios naturais aos medicamentos convencionais.

O surto do vírus da Zika ocorrido entre 2015 e 2016 foi repercutido nas mídias sociais e diferentes comunidades em todo o mundo se envolveram no Twitter, discutindo a doença e os principais problemas associados a ela. Stefanidis et al.[25] (2017) analisaram as conversas no Twitter se baseando em três eixos: localização, atores e conceitos emergentes associados ao problema, para entender como uma emergência de saúde pública de interesse internacional ocorre nas mídias sociais. Os dados foram pré-processados e analisados com ferramentas de análise espaço-temporal e redes complexas para capturar tanto evolução do interesse pelo tópico quanto as conexões entre os locais, atores e conceitos através de grafos. Foi possível verificar a expansão do interesse do assunto zika desde sua localização original, no caso América Latina, até a América do Norte e depois o resto do mundo. As principais agências de saúde e controle de doenças tiveram uma presença proeminente nas discussões. Além desse estudo, uma pesquisa recente[26] (2019) realizou uma análise de tópico multilingual em mensagens do Twitter coletadas também durante o surto de Zika. Diferente do primeiro estudo citado, esse artigo cita o uso de um método de modelo de tópicos polilíngue aplicado em tweets em inglês, espanhol e português. Além da análise de tópicos também investigaram a prevalência do tópico por localização, observando o volume e a distribuição das mensagens ao longo do tempo. Os resultados mostraram que o assunto Zika foi discutido de forma completamente diferentes em diversos países, os tópicos encontrados relacionando cada país mostrou aproximação entre alguns lugares, mas total divergência em outros.

Ainda em 2017, Bian et at.[27] publicaram um estudo com uma análise detalhada de como informações promocionais de saúde relacionadas a síndrome de Lynch afetam as discussões de usuários leigos no Twitter em termos de conscientização e atitudes. Utilizaram técnicas de modelagem de tópicos (LDA) para descobrir os tópicos discutidos sobre a doença e análise de sentimentos (CNN - Convolutional Neural Network) para verificar o sentimento dos usuários nas mensagens publicadas sobre o assunto. Os resultados apresentaram evidências que confirmam os impactos positivos de iniciativas

e eventos de conscientização que foram amplamente promovidos por organizações e profissionais de saúde no Twitter.

Em 2018, um estudo[28] resolveu explorar dados sobre saúde de outra rede social bastante popular, o Instagram. Esse estudo buscou caracterizar tópicos de saúde que são discutidos nessa plataforma de compartilhamento de imagens como um passo para compreender como esses dados podem ser úteis para pesquisas em saúde pública. O Instagram não fornece API para coletar os dados públicos e não permite pesquisar por texto livre, apenas por *hashtag*. Assim, os autores desenvolveram um "*crawler*" que consulta o mecanismo de pesquisa de *hashtag* do Instagram via página web da rede social, coleta os dados da página e analisa as informações como conjunto de tags e legenda da imagem. Apesar das imagens postadas serem públicas o conteúdo ainda é considerado material sensível. A abordagem de modelo de tópicos polilíngüe foi utilizada em um conjunto de dados de aproximadamente 96 mil postagens. Foram identificados 47 tópicos relacionados a saúde sendo que os mais prevalentes estavam relacionados a dieta e exercícios. O Instagram apresenta um certo potencial como fonte de informações de saúde pública, embora limitações na coleta de dados e na disponibilidade de metadados possam restringir seu uso em comparação com plataformas como o Twitter.

Grover, Kar e Davies[29] (2019) exploraram as discussões no Twitter relacionadas às tecnologias utilizadas na saúde. O estudo apresenta as principais tecnologias no domínio da saúde por meio de uma análise de hashtag e as principais doenças (agudas, crônicas, transmissíveis e não transmissíveis) por meio da análise de palavras e sua associação pela co-ocorrência de palavras nos tweets. A associação mostrou que as tecnologias foram usadas no tratamento, identificação e cura de várias doenças. O estudo também realizou diferentes análises estatísticas e de redes sociais apresentando diversos grafos e matrizes de correlação com o intuito de mostrar um quadro geral de como as várias tecnologias são sendo relacionadas ao domínio da saúde.

Mais um estudo[30] focou no tema vacinação buscando caracterizar os indivíduos que compartilham mensagens sobre anti-vacinação, as informações que geralmente são publicadas e a disseminação desse conteúdo nas redes sociais. O conjunto de dados analisado era composto por mensagens publicadas no Facebook de 197 usuários em resposta a uma mensagem que promove a vacinação. Realizaram uma análise quantitativa, análise descritiva, análise de redes sociais e uma avaliação qualitativa. A análise das redes sociais descobriu que os tópicos e as pessoas tendiam a se agrupar

em quatro subgrupos distintos (“confiança”, “alternativa”, “segurança” e “conspiração”). A identificação de subgrupos distintos alerta que não se pode utilizar uma abordagem geral para campanhas ou programas educacionais sobre vacinação, combater um único tema ou argumento provavelmente não será bem-sucedido com todas as crenças anti-vacina.

Na maioria dos estudos apresentados acima a fonte de dados dominante é o microblogging Twitter. A justificativa foi a facilidade de recuperar o conteúdo dessa plataforma já que o Twitter disponibiliza várias bibliotecas de serviço, conhecidas como Application Programming Interface (API), que possibilita a coleta dos tweets de formas variadas e por seus dados serem públicos. O artigo de Klein[23] cita o Twitter como a mídia social mais relevante para rastreabilidade de dados, ao contrário do Facebook que não permite a coleta de mensagens em seu banco de dados de pessoas físicas, apenas de *fanpages*. Todavia, as plataformas de rede social em geral constituem uma fonte de dados rica para tarefas de processamento de linguagem natural, como extração de relações, análise de sentimentos, análise de redes sociais e até mesmo reconhecimento de entidades nomeadas. Assim como a maioria dos estudos esta tese também irá utilizar as mensagens do Twitter como fonte de dados.

Vários desses estudos citados adotaram metodologias para encontrar tópicos e assuntos relacionados a um determinado tema de saúde utilizando técnicas de modelagem de tópicos. Mesmo assim, cada estudo apresentou uma forma diferente de selecionar os tweets “relevantes” ou aqueles tweets que se referem ao tema de busca, e diversos métodos de aprendizado de máquinas supervisionados foram utilizados para tal seleção.

Geralmente, os métodos de aprendizado de máquinas supervisionados precisam de um conjunto de dados de treinamento que necessariamente fora previamente rotulado. Isso incorre em uma influência humana no processo, classificando o conjunto de dados manualmente, o que leva tempo e disponibilidade de pessoas para participarem e classificarem os dados de forma correta.

O presente estudo apresentará uma abordagem diferenciada para seleção de mensagens relacionadas a saúde, evitando utilizar métodos supervisionados que necessitem de uma classificação manual.

A maioria dos estudos que buscam analisar dados de saúde em redes sociais são, geralmente, focados em um tema específico da saúde, por exemplo uma doença ou um medicamento. A seleção das mensagens é realizada com um conjunto pequeno e

específico de palavras chaves (normalmente unigrama), utilizando também abordagens supervisionadas com treinamento de dados. Esta tese difere destes estudos pois tem o propósito de explorar vários assuntos de saúde, relacionando; doenças, medicamentos e sintomas, e investigará os dados de redes sociais utilizando abordagens de modelagem de tópicos e análise de redes sociais com foco no conteúdo.

2.2 Conceitos e Definições

2.2.1 Descoberta de Conhecimento em Bases de Dados e Mineração de Dados

O processo de descoberta de conhecimento em grandes bases de dados é conhecido como Knowledge Discovery in Databases (KDD), uma área interdisciplinar com foco em metodologias para extrair conhecimento útil em dados. Segundo Fayyad et. al.[31] KDD refere-se a um processo não trivial para identificar padrões nos dados que sejam válidos, novos, potencialmente úteis e compreensíveis. A mineração de dados (MD), do inglês *data mining*, é parte do processo de KDD. A Figura 1 mostra uma ilustração das etapas do processo de KDD.

A etapa que recebe destaque nesse processo é a de mineração de dados. A MD pode ser definida como um processo automático ou semiautomático de explorar grandes bases de dados, encontrar padrões valiosos nos dados capazes de embasar a assimilação de informação importante e gerar conhecimento[32]. Basicamente trata-se da aplicação de algoritmos computacionais que são capazes de receber um conjunto de dados de entrada com fatos ocorridos no mundo real e devolver um padrão de comportamento como saída.

Segundo Fayyad, as tarefas de MD são dividas em dois grandes grupos, preditivas e descritivas[31]. Tarefas preditivas estão relacionadas a classificação e regressão/estimação. Já as tarefas descritivas possibilitam identificar padrões e tendências nos dados, geralmente utilizadas em atividades exploratórias de dados, como técnicas de regras de associação, agrupamentos e sumarização.

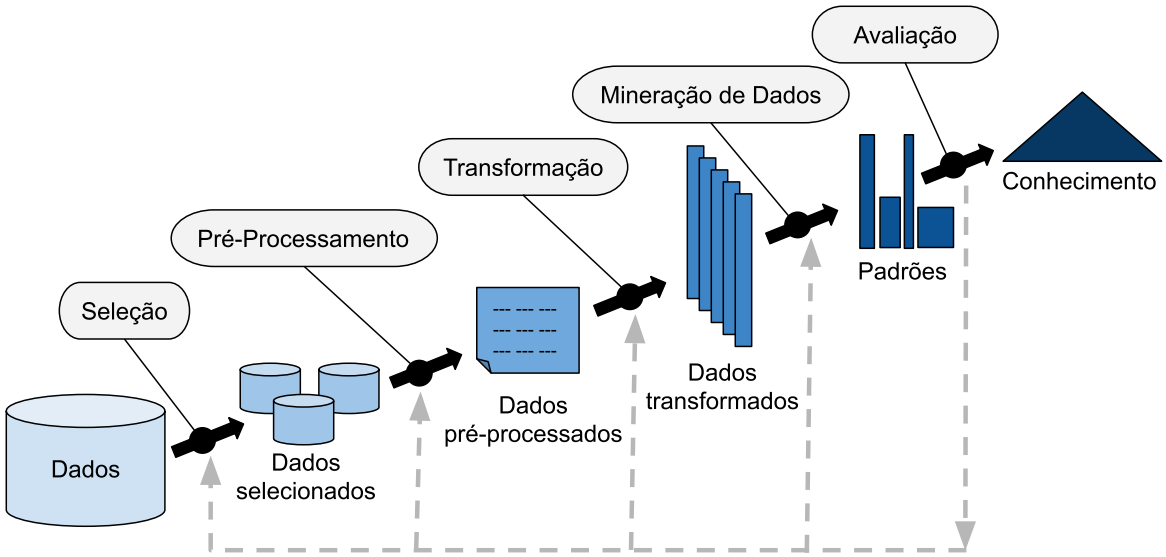


Figura 1- Visão geral das etapas do processo de KDD. Fonte: Fayyad et al[31] (Adaptado pelo autor).

Há diversos processos que definem as fases e atividades da MD, todos em geral apresentam a mesma estrutura. O modelo chamado Cross-Industry Standard Process of Data Mining (CRISP-DM)[33] é um dos mais populares. Esse modelo é composto por seis etapas organizadas de maneira cíclica, como mostra a Figura 2. Apesar de ser composto por etapas no qual o resultado de uma depende da anterior, o fluxo não é unidirecional; é possível ir e voltar entre as etapas com o intuito de executá-las novamente obtendo melhores resultados.

Essas etapas auxiliam na organização das atividades que precisam ser realizadas para o processo de MD. Resumidamente, na primeira etapa (entendimento do negócio) o foco é entender o objetivo que se quer atingir com a MD. Após isso, na próxima etapa (entendimento dos dados) é preciso conhecer os dados, realizando uma análise descritiva dos mesmo para assegurar quais serão importantes para atingir os objetivos da MD. Na terceira etapa (preparação dos dados) abrange todas as atividades que serão realizadas para criação do conjunto de dados que será utilizado na ferramenta de modelagem. A etapa de modelagem é onde se aplicam os algoritmos de MD e os parâmetros são calibrados a fim de encontrar o melhor valor.

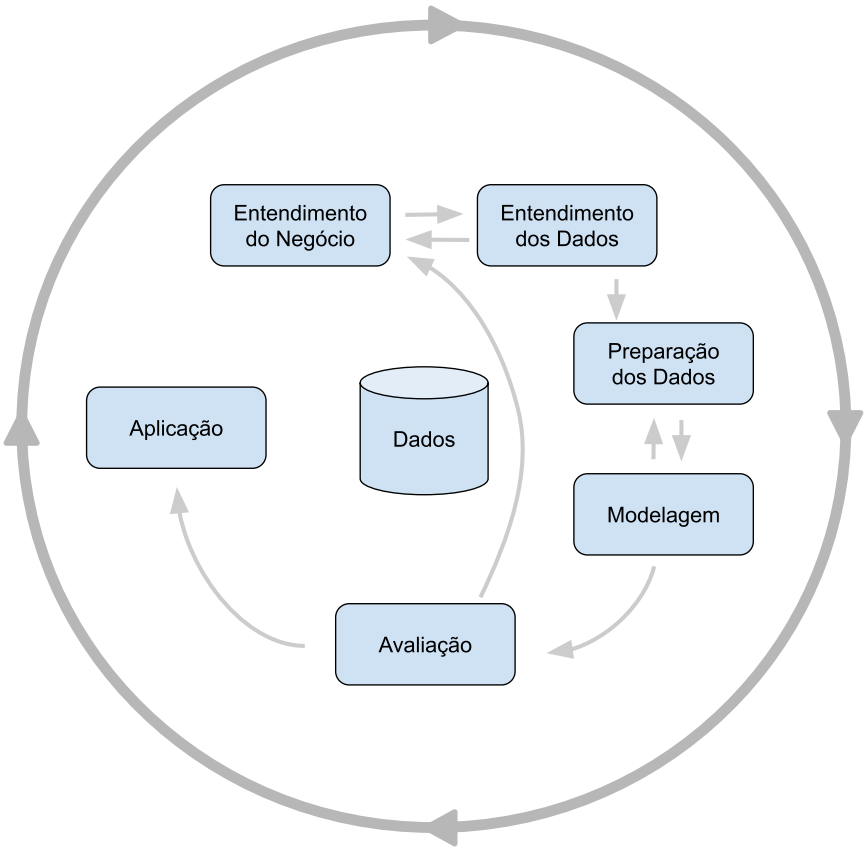


Figura 2 - Ilustração do modelo do processo de mineração de dados CRISP-DM. Esse modelo descreve abordagens geralmente utilizadas em mineração de dados para solucionar problemas. Fonte: Shearer C[33] (Adaptado pelo autor).

A penúltima etapa (avaliação) é considerada uma etapa crítica do processo de mineração na qual certas ferramentas gráficas são utilizadas para a visualização e análise dos resultados observando se os resultados obtidos vão ao encontro com os objetivos propostos. E a última etapa é apresentar e fazer uso dos modelos criados.

A análise de dados não estruturados pode ser considerada uma tarefa complexa se comparada à análise de dados estruturados, assim técnicas específicas para o tratamento deste tipo de dados são necessárias. Essas técnicas são encontradas na área de Descoberta de Conhecimento em Textos (KDT)[34], também conhecida como Mineração de Textos (MT), que pode ser relacionada como uma extensão da KDD. A grande diferença entre essas áreas está na manipulação dos dados, que no caso da KDT são dados textuais registrados em linguagem natural.

A MD ampliou sua aplicabilidade para realizar a mineração de estruturas mais complexas, recebendo novas denominações como, por exemplo, a mineração de grafos que visa descobrir padrões em dados representados como grafos[35].

2.2.2 Redes Complexas

Diversas análises de textos baseiam-se nas relações entre as palavras, seja examinando quais palavras tendem a seguir outras imediatamente ou que tendem a coocorrer dentro dos mesmos documentos. Uma forma de representação de um conjunto de objetos e suas ligações é por meio de um tipo de dado chamado de rede complexa.

Uma rede complexa corresponde a um grafo com elementos discretos representados por um conjunto de nós ligados por arestas[35]. Através dessa representação é possível calcular diferentes métricas para compreensão dos objetos e suas ligações, buscando o entendimento dos dados e o significado de suas interações.

O conceito de redes é amplo e retrata como os objetos estão conectados uns aos outros, descrevendo suas integrações. Existem redes de pessoas, por exemplo, amigos, famílias e mídias sociais; mas também há redes de sistemas de comunicação, sistemas semânticos, malhas elétricas, redes culturais entre outras[36].

A análise de redes sociais (ARS) explora uma série de métricas para realizar a análise da estrutura de conexões de atores em redes[20]. Para representar uma rede existe um subcampo da matemática chamado de teoria dos grafos.

A teoria dos grafos, segundo a literatura, foi idealizada pelo matemático Leonard Euler que apresentou uma teoria de apoio para estudar as relações entre os objetos de um determinado conjunto e se tornou a principal linguagem matemática para descrever as propriedades das redes complexas[35].

Um grafo é uma abstração matemática utilizada para estudar as relações existentes entre objetos. Várias situações do mundo real podem ser representadas por grafos. Ademais desempenham um papel importante na linguística e são úteis para representar o conhecimento semântico[10].

Define-se um grafo $G = (V, E)$ como um conjunto de vértices V (ou nós) e um conjunto de arestas E (ligações ou conexões). Um grafo pode ser visualizado como uma coleção de pontos (ou círculos) e linhas unindo os pontos. A Figura 3 mostra um exemplo de um grafo com seus elementos, os nós e as arestas.

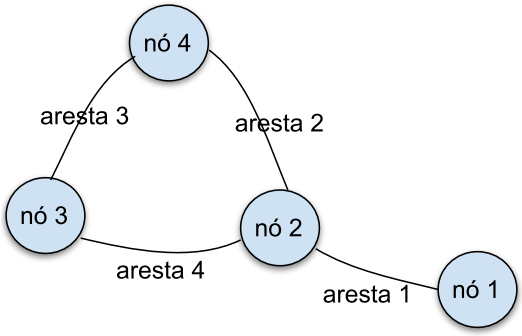


Figura 3 - Exemplo de grafo com quatro nós e quatro arestas. Esse grafo possui os vértices V={1,2,3,4} e arestas E = {(1,2), (2,3), (3,4), (2,4)}.

Um grafo também pode ser representado por uma matriz de adjacência[35], que é uma matriz A com linhas e colunas rotuladas pelos vértices de G. O elemento da matriz na linha i coluna j é 1 se existir uma aresta entre os vértices i e j, ou seja, se $(i, j) \in E$, caso contrário, será 0. As medidas binárias de relações são a forma mais simples para classificações de relações no grafo. Certas propriedades importantes dos grafos baseadas na sua matriz de adjacência são:

- Grafos direcionados: as relações representadas pelas arestas possuem direção/sentido. A representação do fluxo de informação em grafos direcionados pode ser unidirecional ou bidirecional.
- Relação de vizinhança ou adjacência: dois nós são vizinhos se eles estão conectados por uma aresta. Em grafos direcionados a relação de adjacência não é simétrica, ou seja, o nó “A” está conectado ao nó “B”, mas o nó “B” não possui uma conexão de volta. Já em grafos não direcionados a relação é sempre simétrica.
- Grafos ponderados: os grafos direcionados e não direcionados podem ser ponderados, ou seja, tem um número associado a cada aresta que pode ser considerado como a "força da conexão".
- Grau de um nó (*degree*): o número de arestas que incidem em um nó é o grau. Em grafos direcionados o grau do nó é dividido em "grau de entrada" (*in degree*), número de arestas que chegam em um nó, e "grau de saída" (*out degree*) número de arestas que partem de um nó. Por exemplo, o nó 2 na Figura 3 possui três arestas incidentes, (1,2), (2,3), (2,4), e, portanto, tem um grau de 3.
- Caminho: duas arestas são adjacentes se compartilham um nó. Uma sequência

de arestas adjacentes no grafo é um caminho.

- Comprimento do caminho: é o número de arestas das quais ele é composto.
- Subgrafo: um grafo $G' = (V', E')$ é um subgrafo de G, se o V' e E' são subconjuntos de V e E.
- Grafo induzido: é um subgrafo $G' = (V', E')$ que possui um subconjunto de nós cujo todas as arestas que ligam esse subconjunto de nós estão presentes no grafo original G. Ou seja, é um subgrafo induzido se ele possui todas as arestas que aparecem no grafo G sobre o mesmo conjunto de vértices.
- Diâmetro: o diâmetro de um grafo é o maior caminho dentre todos os caminhos existentes do grafo.

Além dessas propriedades há também medidas[38] de ARS que auxiliam na análise das características topológicas e dinâmicas da rede, entre elas temos:

- Centralidade de grau (*degree centrality*): é uma medida que reflete a atividade relacional direta de um nó. É o número de relações/ligações que um nó tem com outros nós, representando o poder do nó na rede.
- Centralidade de intermediação (*betweenees centrality*): Identifica e caracteriza os nós com maior vantagem ou poder numa rede. É uma medida que fornece o número de menores caminhos de todos os vértices para quaisquer outros vértices que passam por aquele nó em consideração.
- Centralidade de proximidade (*closeness centrality*): Comprimento médio das distâncias entre um nó específico e todos os outros nós no grafo.

Em redes complexas com uma grande quantidade de nós é possível detectar grupos de nós que tendem a serem mais conectados entre si do que com o restante da rede. Esses grupos são chamados de comunidade. Uma comunidade de nós tende a ter um número elevado de triângulos, uma tripla de nós conexos entre si, e apresentar uma organização hierárquica dentre os grupos. Identificar comunidades é um problema semelhante ao de detecção de agrupamentos[10], porém como se trata de redes complexas os algoritmos de detecção de comunidades devem levar em conta a ligação entre os elementos.

Além disso, é possível calcular o coeficiente de agrupamento de uma rede através da medida transitividade (*transitivity*)[10]. A transitividade de um grafo é baseada no número relativo de triângulos no grafo comparado ao número total de trincas de nós

conectadas. Essa medida tem o objetivo de indicar quão próximo o grafo está de ser um grafo completo, ou seja, de ter todos os nós conectados. Uma rede totalmente conectada possui a medida de transitividade igual a 1.

Por outro lado, redes de grande escala possuem a tendência em apresentar uma estrutura chamada “centro-periferia”[39]. Essa estrutura apresenta uma rede composta por um grande e denso conjunto de nós no centro ligados entre si, que não tem nenhuma estrutura de comunidade hierárquica, dessa forma não podem ser quebrados em comunidades menores.

Quando se trata de linguagem escrita, três tipos de redes são predominantes: as redes de co-ocorrência, sintáticas e semânticas[40]. As redes de co-ocorrência, que serão abordadas nesta tese, basicamente conectam as palavras adjacentes, dessa formam conseguem capturar a maior quantidade possível de ligações relevantes entre as palavras. As redes sintáticas associam apenas as palavras que possuem relação de dependência na sintaxe linguística. Já as redes semânticas conectam conceitos no qual suas ligações representam relações semânticas entre os conceitos.

2.2.3 Modelagem de Tópicos

Modelagem de tópicos (MT) é um campo ativo de pesquisa em aprendizado de máquinas e processamento de linguagem natural cujo principal objetivo é descobrir tópicos em grandes coleções de documentos[41]**.** Essa abordagem visa a organização e agrupamento de dados de conteúdo textual. É utilizada para construir modelos generativos a partir de dados textuais não estruturados, fornecendo uma estrutura probabilística para ocorrência da frequência de termos em documentos em um determinado corpus[42]. A Figura 4 mostra uma representação visual do processo da MT, extraída e adaptada do artigo de David Blei[9].

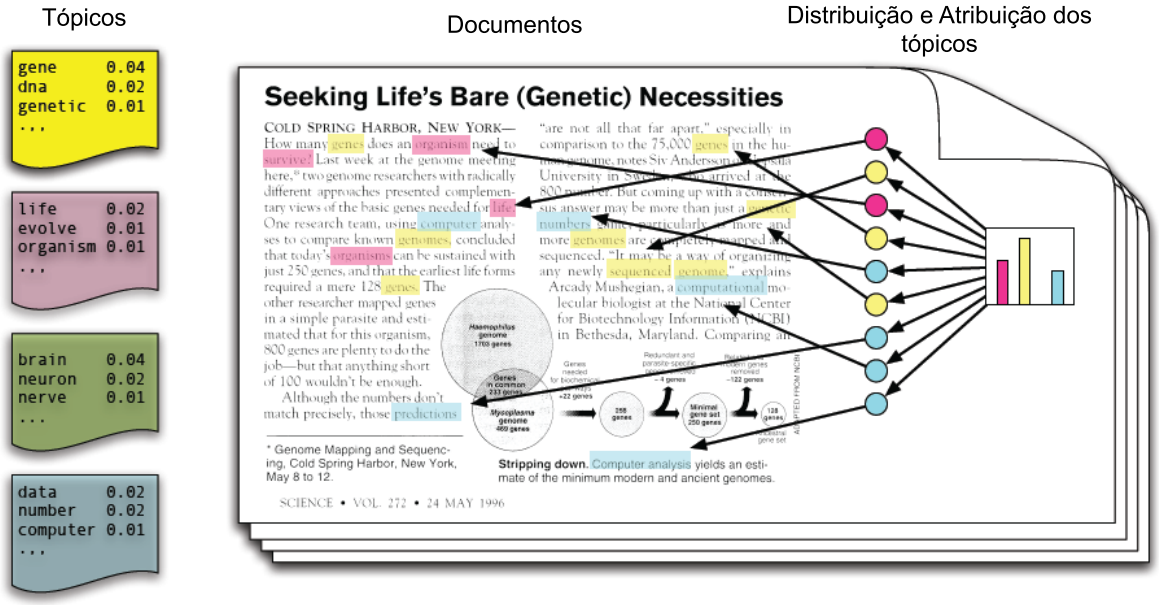


Figura 4 - Ilustração do processo de modelagem de tópicos. A figura mostra a relação entre as palavras destacadas que compõem o documento e os tópicos que emergiram dessas palavras. Fonte: Blei. 2012. (Adaptado pelo autor).

Basicamente a MT identifica padrões em um corpus, extraindo tópicos como distribuições multinomiais sobre palavras que supostamente descrevem os assuntos ou temas apresentados nos documentos. Os tópicos, portanto, podem ser considerados padrões encontrados pela co-ocorrência de palavras com um certo grau de similaridade. A MT é considerada uma técnica de aprendizado não supervisionado, ou seja, os algoritmos não necessitam de nenhum conhecimento prévio dos elementos e os tópicos surgem da análise dos textos[9].

Existem técnicas de MT como a Indexação Semântica Latente[43] (Latent Semantic Indexing - LSI), a Indexação Semântica Latente Probabilística[44] (probabilistic Latent Semantic Indexing - pLSI), a Análise de Semântica Latente[45] (Latent Semantic Analysis - LSA), e a técnica considerada mais popular, que serviu de base para construção de vários outros modelos que é a Alocação Latente de Dirichlet[9] (Latent Dirichlet Allocation - LDA).

A técnica LDA se estendeu do modelo pLSI e faz parte do campo de pesquisa amplo em modelagem probabilística[9]. O LDA é um modelo Bayesiano que trata os dados como oriundos de um processo generativo contendo variáveis ocultas (as estruturas de tópicos). Esse modelo assume que os documentos são produzidos a partir de uma mistura de tópicos. Esses tópicos são formados por palavras baseado nas distribuições de probabilidade das mesmas. Assim, dado um conjunto de

documentos o LDA retrocede e tenta descobrir quais tópicos poderiam gerar esses documentos. Esse processo assume que os documentos têm uma distribuição sobre os tópicos e que esses tópicos possuem uma distribuição de probabilidade sobre um vocabulário fixo. Com a distribuição posterior é possível descobrir os tópicos que melhor explicam o corpus e estes tendem a refletir um padrão de palavras tematicamente consistentes[46].

A distribuição que é utilizada para amostrar a distribuição de tópicos é a de Dirichlet[9]. Os hiper-parâmetros (parâmetros das distribuições de tópicos) são dados a priori no modelo. São eles “alfa” (α) e “beta” (β), sendo que:

- α: representa a densidade do tópico no documento, ou seja, a frequência média da ocorrência de cada tópico em um documento.
- β: representa a densidade da palavra no tópico.

Quanto maior o valor de alfa, mais tópicos compõem os documentos. Por outro lado, quanto maior o valor de beta, os tópicos são compostos por uma grande quantidade de palavras do corpus[46]. Além desses parâmetros para aplicar o modelo é necessário selecionar o número de tópicos k que serão extraídos do corpus e também o número de iterações I que se refere ao número máximo de iterações concedidas para o algoritmo LDA convergir. O objetivo, no treinamento do modelo, é encontrar parâmetros α e β que potencializam a probabilidade de o corpus ser gerado pelo modelo.

Os métodos de inferência do modelo LDA mais utilizados na literatura são o Amostrador de Gibbs (Gibbs Sampling)[29] e Inferência Variacional (Variational Inference)[30]. O primeiro será abordado nesta tese.

3 Métodos

Neste capítulo estão descritos o tipo de estudo realizado, os materiais necessários para o desenvolvimento da pesquisa e os métodos empregados para contemplar o objetivo principal e os objetivos específicos.

3.1 Tipo de Estudo

Trata-se de um estudo exploratório-descritivo[47] que faz uso de técnicas de mineração de dados para investigar novos conhecimentos relacionados à saúde em textos publicados em redes sociais.

Os estudos exploratórios baseiam-se em investigações de um fato ou fenômeno do qual se conhece pouco, auxiliando na compreensão e clarificação dos conceitos por trás desse fenômeno. Este tipo de estudo faz uso, normalmente, de procedimentos sistemáticos para análises de dados ou observações empíricas. Já as pesquisas descritivas possuem caráter informativo e explicativos com objetivo de descrever as características de determinada população, fenômeno ou estabelecer relações entre as variáveis. A combinação destes dois tipos de estudo, os estudos exploratórios-descritivos, apresenta como característica descrever completamente determinado fenômeno utilizando tanto descrições quantitativas e/ou qualitativas quanto informações detalhadas por intermédio da observação participante.

3.2 Comitê de Ética em Pesquisa e Conflito de Interesses

O presente estudo foi aprovado pelo Comitê de Ética em Pesquisa (CEP) da UNIFESP, número do processo CEP 0654/2015 (Anexo A) em 15 de julho de 2015. A rede social foco deste estudo, o Twitter, possui uma política de privacidade no qual qualquer usuário registrado pode publicar um tweet que é público por padrão, e na utilização de qualquer serviço do Twitter o usuário fornece seu consentimento à coleta, transferência, adaptação, conservação, divulgação e qualquer outra forma de utilização ou tratamento da sua informação (http://twitter.com/privacy).

Portanto, neste estudo não houve violação de direitos no acesso das mensagens dos usuários do Twitter. Os autores das mensagens analisadas não foram identificados neste estudo.

3.3 Fluxo de Desenvolvimento da Pesquisa

Esta pesquisa baseia-se na execução de três etapas principais para cumprir os objetivos propostos. A etapa 1 remete a coleta e seleção dos dados, e na escolha das listas de termos de saúde que foram utilizadas para filtrar os dados coletados. A etapa 2 apresenta os métodos de pré-processamento e mineração de dados para a análise do conteúdo como descoberta de padrões e comportamentos, e por fim a etapa 3 descreve o processo de análise exploratória dos tópicos utilizando a abordagem de modelo de tópicos. A Figura 5 apresenta uma ilustração das etapas metodológicas para elaboração desta pesquisa.

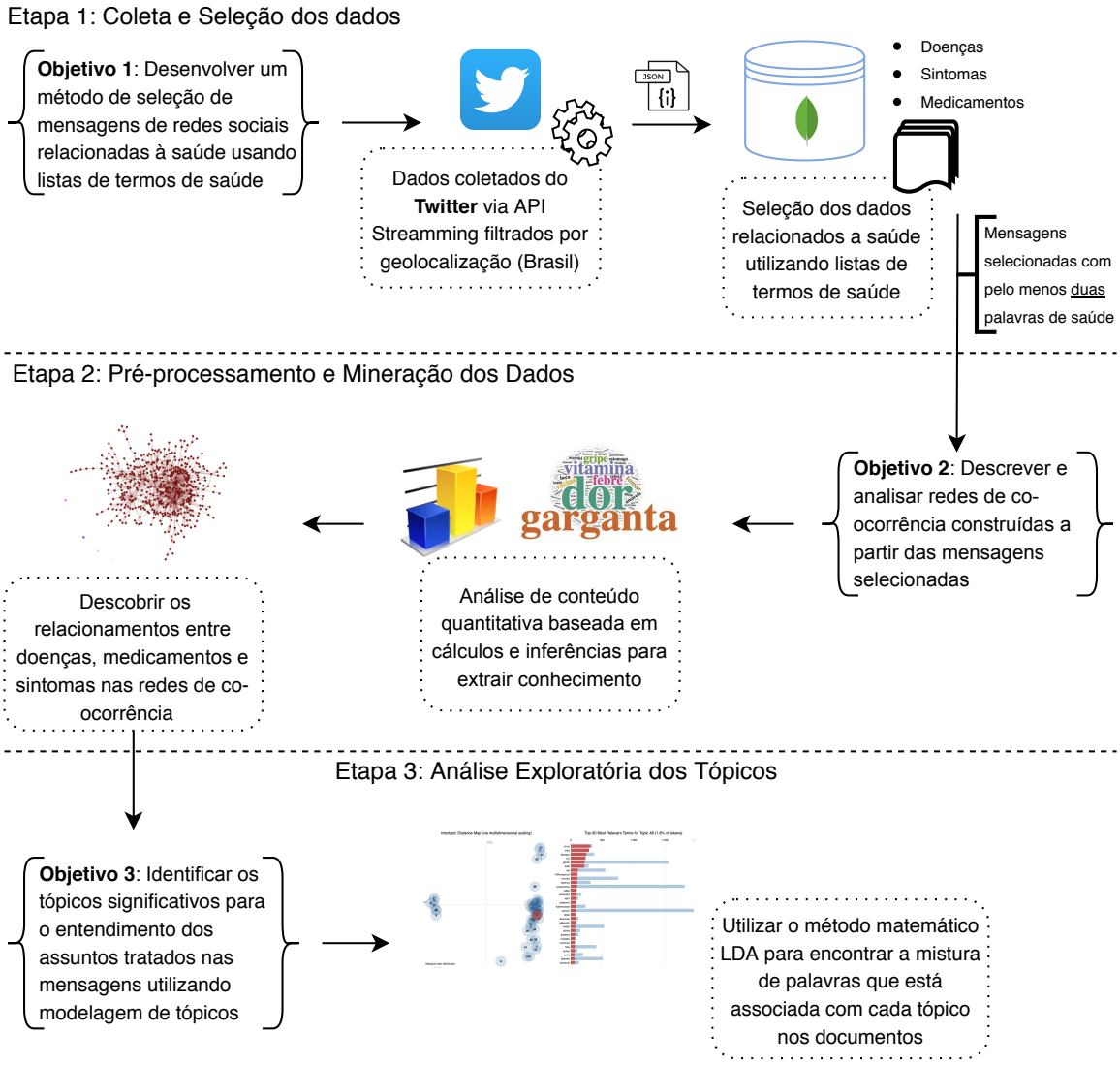


Figura 5 - Ilustração das etapas metodológicas propostas para realização deste estudo englobando os três objetivos específicos.

3.4 Coleta e Seleção dos Dados

Inicialmente foi realizada a coleta e armazenamento dos dados. Para armazenar os dados coletados do Twitter uma base de dados não relacional[48] (NoSQL) foi selecionada. Os bancos de dados orientados a documentos são geralmente utilizados para o armazenamento de grandes volumes de dados, possuem uma maior flexibilidade e escalabilidade, e processam dados mais rápido do que os bancos de dados relacionais por apresentarem uma modelagem dos dados simples. Este tipo de banco de dados pode manipular e trabalhar com estruturas e conjuntos de dados maiores do que o tradicional devido à sua natureza distribuída e a forma diferenciada em que armazenam os dados fisicamente. Além disso, já vem sendo utilizados pelas redes sociais Twitter e Facebook. Existem vários bancos de dados NoSQL como o Cassandra, MongoDB, CouchBase etc. Nesta pesquisa foi utilizado o banco de dados MongoDB por ser de código aberto e possuir uma comunidade ativa que oferece suporte e melhorias.

Foi desenvolvida uma aplicação em NodeJS (https://nodejs.org/) que interage com a API do Twitter para coletar todas as mensagens publicadas no Brasil. A API do Twitter disponibiliza a utilização de filtros de pesquisa e coleta dos tweets, e nesse caso foi utilizado o filtro de geolocalização para retornar apenas as mensagens publicadas no Brasil.

3.4.1 Listas de Termos de Saúde

Com a intenção de identificar as mensagens relacionadas a saúde quatro listas com termos e frases de saúde foram selecionadas para filtrar as mensagens coletadas.

A Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (CID-10)[49] foi a primeira lista selecionada. A CID-10 é publicada pela Organização Mundial de Saúde (OMS) e visa padronizar a codificação de doenças e outros problemas relacionados à saúde, como, distúrbios, lesões etc. Por ser uma publicação oficial da OMS, os países membros precisam adotá-la para finalidade de apresentações estatísticas das causas de morte (mortalidade) ou das doenças que levam a internações hospitalares ou atendimentos ambulatoriais.

Atualmente é a classificação diagnóstica padrão internacional para propósitos epidemiológicos gerais e administrativos[50]. Possui uma ampla variedade de conceitos

clínicos diferentes por natureza. Além disso, em sistemas de informação em saúde essa classificação pode ser utilizada para monitorar a incidência e prevalência de doenças, avaliar o estado de saúde dos pacientes no país e dimensionar os programas e ações de saúde que serão prestados aos cidadãos, como ações preventivas e de promoção à saúde.

A CID-10 é composta por 22 capítulos e cada capítulo descreve um grupo de doenças semelhantes ou causadas pelo mesmo fator. Nesta pesquisa foi utilizada a tabela CID-10 fornecida pelo Datasus tanto para a lista de doenças quanto para a de sintomas. Na CID-10 o Capítulo XVIII - Sintomas, sinais e achados anormais de exames clínicos e de laboratório, apresenta os sintomas de doenças em geral a partir dos códigos R00 a R99.

Além da CID-10 foi escolhido um guia farmacêutico que possui informações de todos os medicamentos, laboratórios, soluções parenterais e materiais hospitalares, conhecido como Brasíndice (https://www.brasindice.com.br/), para filtrar as mensagens que possam conter algum medicamento.

Ademais foi utilizada uma lista adaptada das categorias de fármacos do Wikipédia (https://pt.wikipedia.org/wiki/Categoria:Fármacos) que contém os tipos/categorias de medicamentos/fármacos como analgésicos, antibióticos, antigripais, antiinflamatórios, diuréticos, soníferos, vitamínicos, anticoncepcionais, fitoterápicos, antisépticos, anticonvulsivantes, antiespasmódicos, antitussígenos, antimicóticos, antihelmínticos, antipirético, laxante, vacinas, antidepressivo, antihistamínico, antifúngico etc., com o intuito de abranger ainda mais o escopo de busca por tweets de saúde.

Por fim foi utilizada uma lista de termos médicos extraída de um sistema de registro eletrônico em saúde (S-RES) utilizado por mais de 4 mil profissionais da área da saúde no Brasil. Essa lista contém termos de saúde comumente utilizados por médicos no atendimento dos pacientes. Geralmente estão relacionados a condições de saúde, sinais e sintomas. Todos os termos desta lista foram validados pelo cruzamento com os termos do Descritor em Ciências da Saúde (DECS - https://decs.bvs.br). Essa lista foi retirada do sistema ClinicWeb em 2015. Essa lista foi denominada nesta pesquisa como “termos médicos”.

3.4.2 Seleção das Mensagens sobre Saúde

A partir das listas descritas na seção anterior foi criado um mecanismo para selecionar as mensagens armazenadas que apresentavam algum termo de saúde das listas.

Nos textos da língua portuguesa encontramos variações nas palavras com relação ao gênero, número ou grau e dos inúmeros tempos verbais. Por isso, para selecionar o maior número de mensagens o mecanismo de busca por mensagens que continham termos de saúde fez uso da técnica *stemming*[51]. O *stemming* é um processo que consiste em reduzir uma palavra flexionada ou não a sua parte essencial. Por exemplo, a palavra “gripada” pode ser reduzida para “grip”. Dessa forma, fazendo uso dessa técnica há uma probabilidade maior de encontrar mensagens relacionados aos termos de saúde procurados. O próprio MongoDB possui essa técnica disponibilizada em seu mecanismo de “$text” (https://docs.mongodb.com/manual/reference/operator/query/text/) que faz a busca textual em seus documentos.

Com isso foi desenvolvido um processo automático para selecionar as mensagens do Twitter previamente armazenadas no banco de dados MongoDB. Esse processo foi realizado em partes já que o banco de dados continha um número grande de dados armazenados. Primeiramente, as palavras unigramas, ou seja, com um único termo, foram selecionadas para filtrar as mensagens do banco de dados. A Tabela 1 mostra uma relação da quantidade de atributos n-grama e unigrama em cada lista de termos de saúde.

Tabela 1 - Total de atributos n-grama e unigrama relacionados em cada lista de termos de saúde.

Categorias	**Quantidade total de atributos n-grama**	**Quantidade total de atributos unigrama**
Doenças	13.826	521 (4%)
Sintomas	386	50 (13%)
Medicamentos	14.960	6.239 (42%)
Categorias de Medicamentos	33	33 (100%)
Termos Médicos	657	200 (30%)
Total de atributos	**29.862**	**7.093**

Após selecionar as mensagens com as palavras unigramas, o algoritmo foi

então aplicado para os atributos n-grama. Pelo fato da tabela CID-10 ser composta por vários registros com frases/sentenças foi necessário realizar um processo de tokenização na tabela para poder utilizar alguns termos unigramas que representassem doenças e sintomas assim ampliar a seleção das mensagens.

Segundo Culotta[52], encontrar uma única palavra de saúde em uma mensagem do Twitter não é suficiente para filtrar mensagens relacionados a saúde. Portanto, após filtrar as mensagens que continham pelo menos uma palavra das listas de saúde selecionadas, apenas as que possuíam duas ou mais ocorrências de palavras de saúde (por exemplo: um sintoma e doença) ou atributos n-gramas foram mantidos. Segundo Jimeno-Yepes et al.[53] essa abordagem de seleção contendo pelo menos duas palavras de saúde certifica uma maior probabilidade de o conteúdo daquela mensagem estar realmente relacionado a saúde.

3.5 Pré-processamento e Mineração dos Dados

A Figura 6 mostra o esquema com as etapas e tarefas para análise do conteúdo textual selecionado.

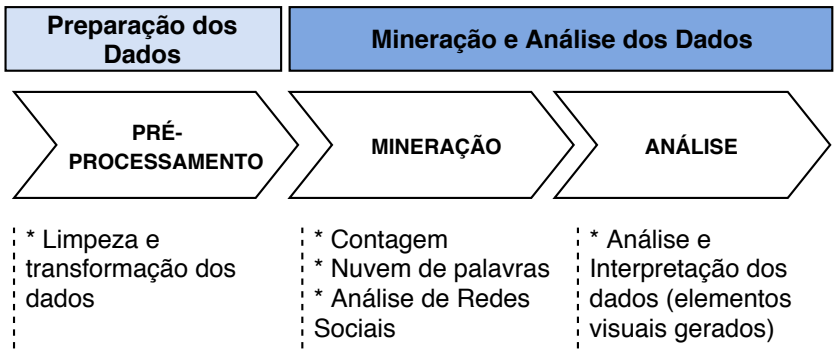


Figura 6 - Etapas e atividades desenvolvidas para análise das mensagens selecionadas.

3.5.1 Preparação dos Dados

Com os dados selecionados o próximo passo é realizar o pré-processamento ou transformação dos dados. O pré-processamento de dados têm a responsabilidade de melhorar a qualidade dos dados para aumentar a eficiência do processo de mineração de dados. Com as mensagens com conteúdo de saúde selecionadas foi utilizado o software R (https://www.r-project.org/) para realizar todos os processos de limpeza e transformação dos dados.

O software R utilizado neste estudo foi o da versão o R 3.2.4, sendo escolhido por ser um software com um grande poder de análise de dados fornecendo ferramentas robustas para manipular e preparar os mesmos. Contém recursos para análise de mineração de texto (*text mining*), como por exemplo o pacote “tm” que possui uma introdução ao *text mining* e “wordcloud” para nuvem de palavras. Além disso o R tem um grande conjunto de funções e pode ser aperfeiçoado com o uso de novos pacotes, ou seja, é um programa poderoso com relação a análises estatísticas e mineração de dados. Por meio deste software é possível realizar a limpeza e transformação dos dados realizando os seguintes processos:

1. Remoção das *stopwords*[54] que são termos não representativos para um documento como artigos, preposições, conjunções e pronomes.
2. Remoção de URLs que não são importantes para uma análise do conteúdo textual e relacionamento entre palavras.
3. Remoção das tags de usuário (quaisquer palavras que come- cem com "@").
4. Remoção das *hashtags* (palavras que comece com "#").
5. Remoção de pontuação, caracteres especiais e números.
6. Remoção de letras repetidas mais de duas vezes nas palavras (Ex.: A palavra “Gripeeee” fica “Gripe” após a remoção das letras repetidas). Exceções como dois “r”, dois “s” e repetições da letra “o” foram devidamente tratadas.
7. Padronizando as palavras em caixa baixa.

Essa é a etapa mais onerosa, apesar das implementações existentes para grande parte das tarefas de pré-processamento, essas tarefas consomem boa parte do tempo do processo de extração do conhecimento. Além disso, com o entendimento dos dados alguns cuidados devem ser realizados, como tratamento exceções e particularidades dos dados.

3.5.2 Mineração e Análise dos Dados

Com o corpus formado o próximo passo é realizar a mineração dos dados. Primeiro uma análise de conteúdo quantitativa baseada na identificação dos termos mais frequentes e suas correlações foi realizada. Métodos de contagem de ocorrência de palavras da área de RI foram utilizados. Os seguintes procedimentos foram realizados para conceber a análise dos dados.

Apesar de na etapa de preparação dos dados as *hashtag*, tags designadas pelo

usuário, serem removidas para posterior análise do conteúdo textual, antes de removê-las foi realizada uma análise de *hashtag.* As hashtag são reconhecidas por apresentar o caracter “#” no início da palavra e são, por convenção, entendidas como rótulos de metadados que de alguma maneira estão relacionados ao conteúdo da mensagem que está presente. A rede social Twitter foi um dos principais disseminadores dessa convenção que posteriormente se tornou um recurso para filtrar os conteúdos postados na rede por assunto.

Contagem da ocorrência dos termos (*term occurrence*[55]) é uma abordagem simples que contabiliza quantas vezes as palavras aparecem nas mensagens e cria-se uma estrutura de dados organizada (vetor de características) de todos os termos das mensagens e suas respectivas quantidades de ocorrência. A partir desta estrutura é possível visualizar gráficos de linha e nuvem de palavras que destacam os termos que mais ocorreram possibilitando notar os assuntos de saúde em evidência nas mensagens no recorte de tempo selecionado.

As nuvens de palavras surgiram como um método de visualização simples e atraente para investigação de texto e podem servir como um ponto de partida para uma análise profunda[56]. Elas são usadas em vários contextos como um meio de fornecer uma visão geral de um texto, ajudando a julgar se um dado texto é relevante para uma necessidade específica de informação. Geralmente isso é feito de maneira estática como resumo puro do texto. As palavras aparecem em fontes de vários tamanhos podendo também estar em diferentes cores, indicando o que é mais relevante e o que é menos relevante no contexto. Esse recurso pode ser utilizado em atividades de interpretação e produção de textos.

Análise de redes sociais consiste em encontrar relacionamentos entre os objetos de estudo, que neste trabalho são as palavras das mensagens publicadas nas redes sociais, e identificar as formas de interação e suas correlações criando uma rede de coocorrência de palavras. Essa rede é representada por meio de grafos que são objetos matemáticos consistindo de pontos, conhecidos por nós ou vértices e suas ligações, chamadas de arestas. Os nós da rede serão as palavras e as arestas as conexões entre as palavras que representam uma coocorrência entre dois termos que ocorriam ao mesmo tempo em uma mesma unidade (tweet). Medidas de centralidade são utilizadas para medir a importância dos nós, podendo indicar a popularidade dos termos. As medidas que foram analisadas a partir de métricas da análise de rede são:

a. centralidade de grau (*degree*)

b. centralidade de intermediação (*betweenness*)
c. centralidade de proximidade (*closeness*)

3.6 Análise Exploratória dos Tópicos

Após a análise quantitativa dos dados a abordagem utilizada para identificar os tópicos significativos para o entendimento dos assuntos tratados nas mensagens foi a modelagem de tópicos. Para a realização da modelagem de tópico utilizou-se o modelo popular de inferência, o LDA, com o método Amostrador de Gibbs (Gibbs Sampling). Esse método é utilizado principalmente por ser razoavelmente simples e fácil de implementar e por sua aplicação em diversos problemas. O método de Gibbs aplicado no LDA é colapsado, possui uma implementação no software R pela função "lda.collapsed.gibbs.sampler" e essa implementação foi utilizada nesta tese.

Certos processos precisam ser realizados nos dados para aplicar o LDA. É necessário criar uma tabela de termos (vocabulário) com todos os termos únicos encontrados no corpus. O corpus também precisa ser ajustado para o formato requerido pelo pacote de análise do R. Após isso, certas estatísticas são computadas e o modelo precisa ser configurado e treinado. Como o LDA utiliza a distribuição de Dirichlet os hiper-parâmetros são dados a priori no modelo. O primeiro parâmetro definido foi o número de tópicos, embora não exista nenhuma regra formal, esse número foi definido baseado no tamanho do conjunto de dados seguindo as recomendações de Abinaya[57]. Os parâmetros e hiper-parâmetros definidos para configuração do modelo foram k=50, α=0.1 e β=0.01, onde k é o número de tópicos previamente definido. Foi definido para o Amostrador de Gibbs executar 5.000 iterações. Apesar de ser um número conservador consegue garantir uma convergência. Para ajustar o modelo, usamos o pacote R "lda" e visualizamos a saída usando o LDAvis[58].

Na modelagem de tópicos é importante encontrar a configuração adequada para o seu modelo de dados, assim o método irá produzir tópicos mais informativos e coerentes. A partir dos tópicos produzidos é possível identificar qual é o assunto que é representado dentro de cada tópico, os tópicos recorrentes no conjunto de dados e encontrar o número de tópicos que defina de forma coerente o corpus.

Todas essas análises foram realizadas com o intuito de interpretar e avaliar os dados selecionados.

4 Resultados

Nesta seção apresentam-se os resultados obtidos em cada etapa metodológica descrita na seção anterior englobando os objetivos específicos.

4.1 Coleta e Seleção dos Dados

As mensagens foram coletadas no período de 28/02/2017 a 31/08/2017 (185 dias) totalizando 141.283.011 mensagens. Após a seleção das mensagens foi possível identificar que a maior parte dos tweets coletados não continham termos de saúde. O total de tweets únicos que apresentou pelo menos uma palavra de saúde foi 1.853.800, portanto, apenas 1,3% dos 141.283.011 de tweets incluía uma palavra de saúde. A Tabela 2 mostra as estatísticas de quantidade de mensagens para cada tipo de categoria de saúde relacionada.

Tabela 2 - Quantidade de mensagens que apresentaram pelo menos um termo das listas de saúde relacionadas pelas categorias.

Categorias	Quantidade de mensagens
Doenças	365.570
Sintomas	685.531
Medicamentos	847.388
Categorias de Medicamentos	22.045
Termos Médicos	96.067

Doenças e Medicamentos são as categorias que possuem o maior número de conceitos e termos nas listas de saúde, enquanto as outras listas apresentam um número comparativamente menor. Consequentemente, Doenças e Medicamentos foram as categorias mais encontradas nas mensagens, todavia Sintomas também trouxe uma quantidade significativa de mensagens mesmo não sendo uma lista tão volumosa.

Em seguida, com a seleção de mensagens com mais de uma palavra de saúde foram encontradas 95.079 mensagens. Todas foram rotuladas com a palavra que foi encontrada e a categoria (“diag”, “sint”, “med”, “tipomed” e “termomed”) que a palavra está inserida. Essas mensagens foram publicadas por 76.900 usuários diferentes.

O Quadro 1 - Exemplos de mensagens com mais de uma palavra das categorias

de saúde encontrada. mostra exemplos de mensagem com mais de uma palavra de saúde. Essa estratégia de selecionar apenas mensagens que contém pelo menos duas palavras de saúde contribui para que apenas mensagens de conteúdo exclusivo relacionado a saúde sejam selecionadas evitando mensagens fora do contexto saúde, ambíguas, e descarta possíveis textos com linguagem figurada[59].

Quadro 1 - Exemplos de mensagens com mais de uma palavra das categorias de saúde encontrada.

Mensagens
"Um mês de cama entre remédios, tosse, febre, diarréias, vômitos, tremores, espir- ros, mais tosse, mais remédios, mal estar, sono infinito..."
"Eu tô aqui morrendo estudando dengue, febre amarela, chikungunya, HPV, HIV, hepatite, sarampo, caxumba, virose... AI, ESSAS PRAGAS DE VÍRUS!"
"Tomei três vacinas, febre amarela, tétano e hepatite... to jogada na cama, muito ruim"
“Ibuprofeno, aspirina, dipirona, e ainda tomei o Loratadina. E estou tomando vitamina C direto, será que vou morrer? Gkskgkks”
“to com dor de garganta, coriza, cabeça pesada etc e não paro de pensar 'af, será se dá pra ir pro treino hj' migas parece q quero é morre"

Abaixo observa-se um gráfico (Figura 7) que apresenta a quantidade de mensagens que foram publicadas em cada mês.

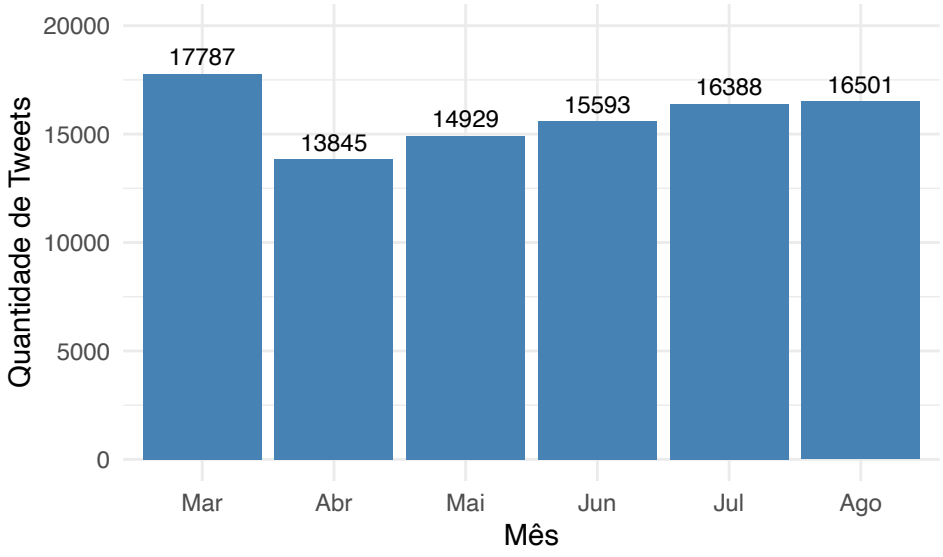


Figura 7 - Gráfico com a quantidade de tweets de saúde publicados em cada mês do período selecionado. Obs.: o mês de fevereiro não foi apresentado na figura pois coletou-se apenas um dia e o número de tweets foi 36.

Nota-se que em nenhum mês o número de mensagens foi abaixo de 10 mil, exceto fevereiro que houve coleta em apenas um dia. O mês que teve um pico com a maior quantidade de tweets foi março, sugerindo que houve um evento específico nesse período.

4.2 Mineração de dados e Análise Quantitativa do Conteúdo

A partir deste recorte foi constituído o corpus de saúde da análise. Técnicas de mineração de dados foram aplicadas nas mensagens para extrair regularidades, padrões e tendências dos dados textuais não estruturados. O intuito foi explorar o conteúdo de forma quantitativa baseando-se na eleição e na classificação de palavras em conceitos e a posterior análise de suas correlações.

Primeiro, foram contabilizados o número de mensagens do corpus que apresentavam *hashtags* (#). Foram encontrados 3.008 *hashtags* únicas em 3.376 mensagens, aproximadamente 3% das mensagens do corpus de saúde. A Figura 8 e Figura 9 apresentam um gráfico ordenado por frequência com as 20 *hashtags* que mais ocorreram nas mensagens e uma nuvem de palavras com as 100 *hashtags* mais frequentes, respetivamente. Através dessas visualizações foi possível identificar determinados assuntos específicos agrupados pelas *hashtags*. Vários assuntos de saúde mais comuns e esperados foram realçados nessas figuras como “#gripe”, “#saude” e “#vitaminas”.

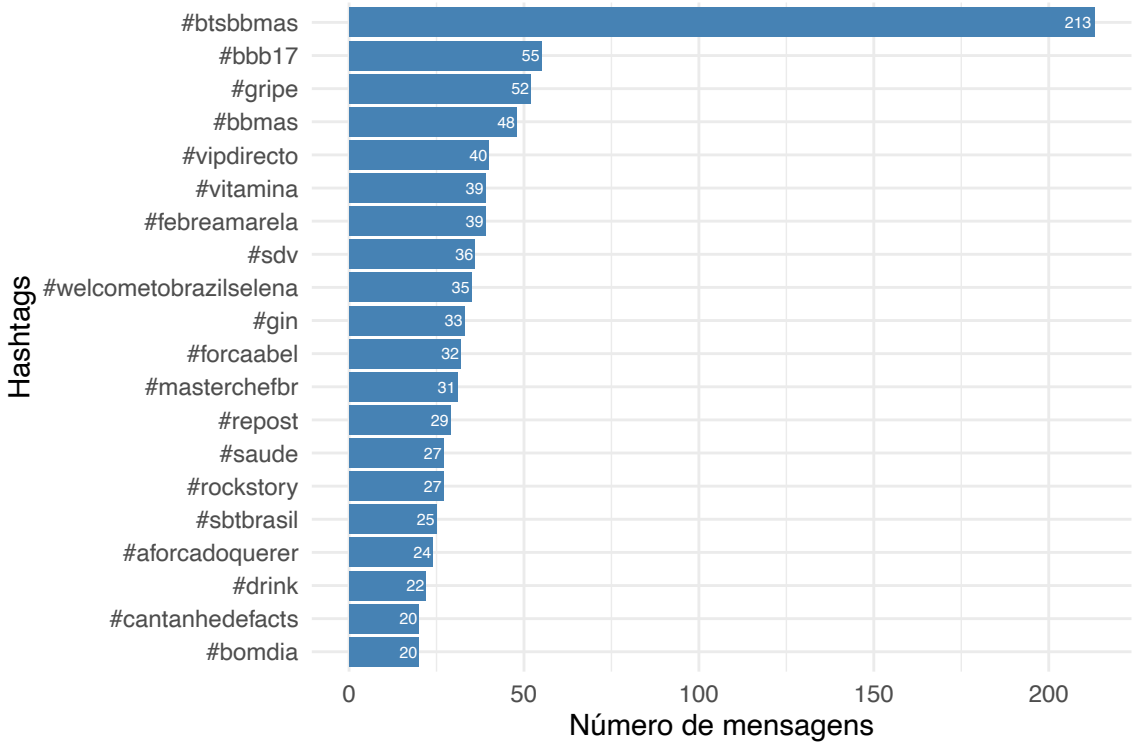


Figura 8 - Ranking das 20 mais frequentes *hashtags* no corpus.

Figura 9 - Nuvem de palavras das 100 *hashtags* mais frequentes no corpus selecionado. O tamanho da fonte das palavras é proporcional à frequência da mesma correspondente no conjunto de dados. Figura com todas as palavras em formato digital interativo disponível em: https://bit.ly/3aerrPp

Na nuvem de palavras a frequência da palavra é usada para medir a importância da palavra em um contexto ou documento. Essa técnica evidencia as palavras importantes no texto (ou em um conjunto de textos). Quanto maior a frequência da palavra maior o tamanho da fonte da palavra que é apresentada. A cor das palavras nesse caso não apresenta nenhum significado para análise.

Um assunto particular como “#febreamarela” também foi observado como uma *hashtag* bem presente no corpus, indicando algum evento que pode ter causado preocupações na população. Outros assuntos como “#vacinacao”, “#dieta”, “#cancer”, ”#depressaonaoedrama”, “#febre”, “#hiv”, “#dengue” e “#zika” são comuns se considerar que todo ano apresentam-se campanhas de vacinação, contra diversos tipos de câncer, depressão e para prevenir o HIV. As hashtag “#SUS” e “#orgulhosus” foram as únicas relacionadas a um serviço público de saúde. O Conselho Municipal de Saúde trouxe para o Brasil o Dia do Orgulho SUS que foi simbolizado pela *hashtag* #orgulhosus e convocou os usuários e profissionais do sistema de saúde a proporem ideias de como melhorar a assistência ao paciente dentro do serviço de saúde. Em particular, essa hashtag mostrou que vários usuários se engajaram para compartilhar ideias de melhorias na assistência que podem ser de grande valia para os serviços de saúde do Brasil.

Muitas *hashtags* referem-se a eventos (esportivos, shows, campanhas de saúde) e a programas de TV (novelas, filmes, séries). Esses exemplos são uma prova do tipo de informação que pode ser obtida em mensagens do Twitter apenas monitorando *hashtags*.

Próximo passo foi analisar a frequência das palavras utilizadas nas mensagens do corpus. As palavras que aparecem com maior frequência foram contadas e representadas de forma ilustrada na Figura 10. Nesta figura temos um gráfico de barras com as 20 palavras que mais ocorreram nas mensagens do corpus e suas respectivas contagens. Por meio desse gráfico ordenado de palavras nota-se que os termos de maior ocorrência no corpus foram os relacionados à saúde. A palavra “dor”, uma das palavras-chaves utilizadas na seleção de mensagens sobre saúde (proveniente da lista de sintomas), foi a que teve um maior destaque entre todas as outras palavras-chaves utilizadas de todas as listas.

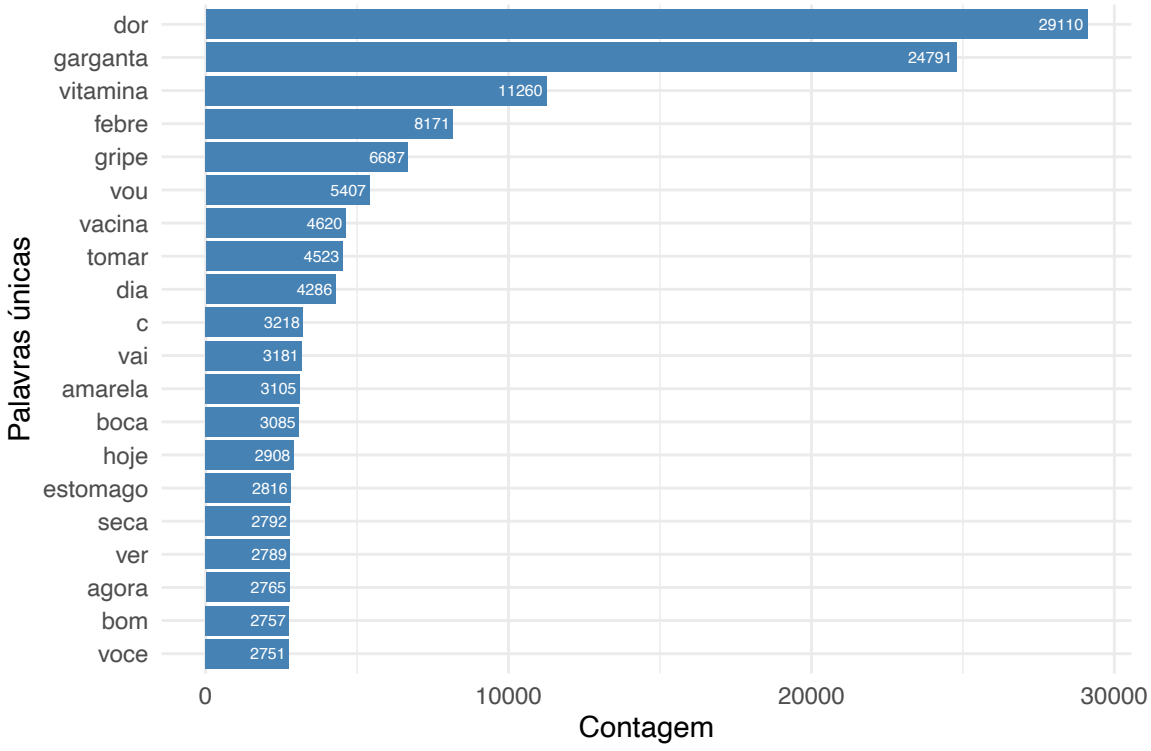


Figura 10 - Gráfico ordenado com as 20 palavras unigramas mais frequentes nas mensagens do corpus.

Por meio da nuvem de palavras (Figura 11) foi possível identificar vários assuntos proeminentes como: “dor”, “garganta”, “vitamina”, “gripe”, “febre”, “vacina” e essas palavras que tiveram maior frequência são palavras consideradas de saúde. Essa visualização também mostra que as palavras relacionadas a “sintomas” se

sobressaíram, o que implica que são frequentes nos textos dos tweets confrontando com palavras referentes a doenças e medicamentos. Várias palavras em destaque apontadas nesta nuvem se convergiram com as palavras encontradas na nuvem de *hashtag*, mostrando a coerência que as *hashtag* podem ter com relação ao assunto tratado em uma determinada mensagem. Outra diferença que pode ser notada entre as duas nuvens foi que na primeira (Figura 9) é possível analisar conceitos particulares já que a hashtag pode conter mais de um termo como por exemplo "#febreamarela", já na segunda nuvem (Figura 11) os assuntos são generalizados, todavia conexões e relacionamentos entre as palavras destacadas podem ser facilmente visualizadas.

Figura 11- Nuvem de palavras (n=100) que mostra o grau de frequência das palavras das mensagens relacionados a saúde do corpus. Figura com todas as palavras em formato digital interativo disponível em: https://bit.ly/33DDCD3

Além da contagem total do número de ocorrência das palavras nas mensagens, a tabela abaixo (Tabela 3) mostra os cinco termos de saúde que tiveram maior ocorrência nas mensagens do corpus, discriminados em suas categorias. As categorias medicamentos e tipos medicamentos foram mescladas nesta análise, assim como as categorias termos médicos com a de doenças.

Tabela 3 - Apresentação dos cinco termos mais frequentes de cada categoria de saúde analisada. A coluna frequência é a quantidade de mensagens em que o termo apareceu

Doenças	Freq.	Sintomas	Freq.	Medicamentos	Freq.
Gripe	8.196	Dor de garganta	23.776	Vitamina	12.285
Febre Amarela	3.061	Febre	8.029	Vacina	5.682
Câncer	1.848	Tosse	4.502	Xarope	805
Pneumonia	1.789	Nervosismo	3.789	Soro	790
TPM	2.119	Vômitos	3.535	Antibióticos	651

Nesta tabela foi possível identificar alguns termos n-gramas. Apesar de poucos os que foram encontrados, esses poucos tiveram destaque por ocorrerem em várias das mensagens selecionadas.

4.3 Análise de Redes Sociais

Depois de realizar as análises linguísticas quantitativas foi explorado a análise de rede de palavras utilizando grafos. Os objetos nesse estudo, ou seja, os nós da rede, são as palavras provenientes do conjunto de mensagens selecionadas. E as arestas são as conexões entre as palavras que representam uma coocorrência entre dois termos que ocorriam ao mesmo tempo em uma mesma unidade (mensagem do Twitter). Com isso é possível analisar a relação de incidência e adjacência entre as palavras. O grafo direcional de todas as palavras dos tweets é mostrado na Figura 12.

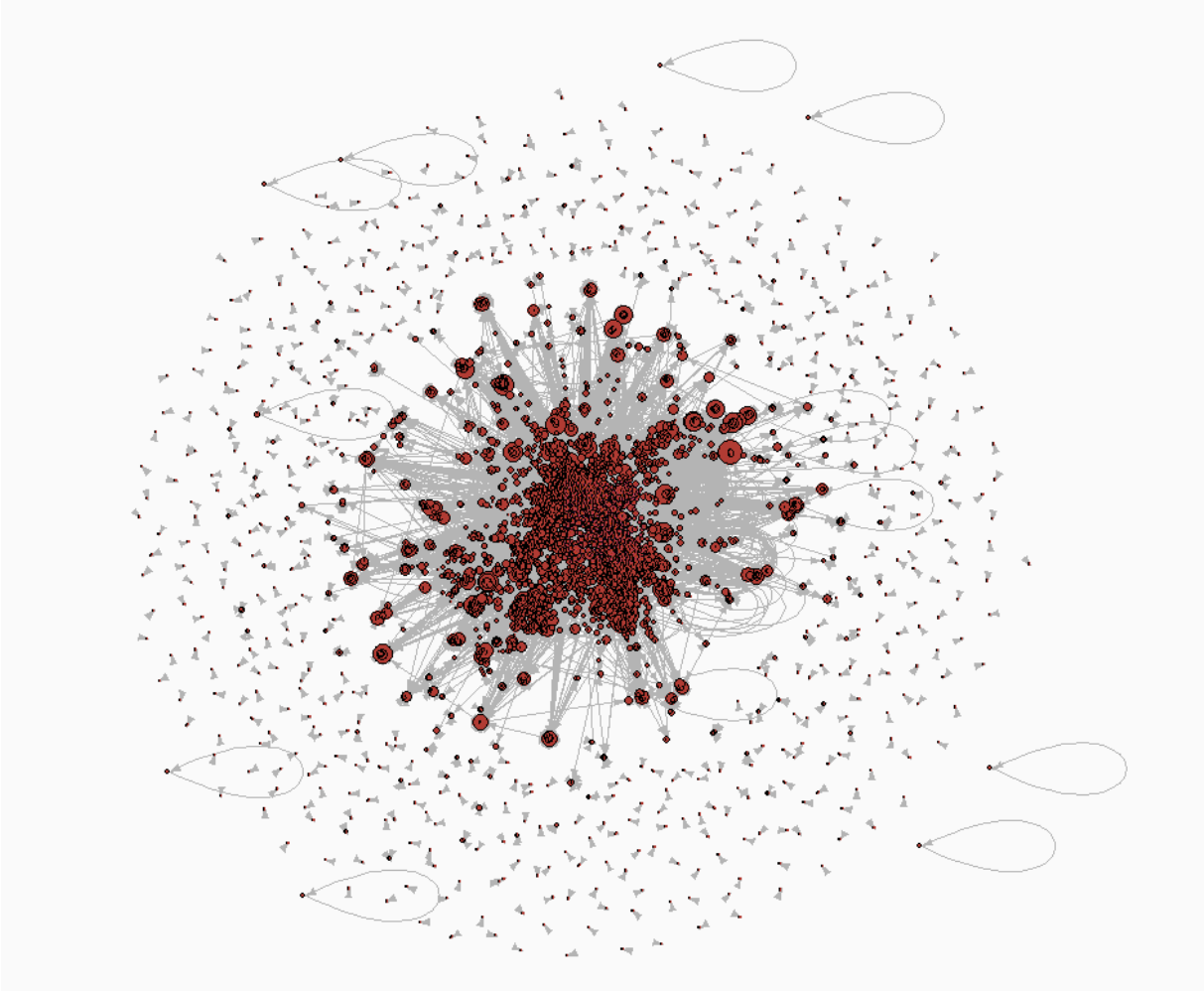


Figura 12 - Rede direcionada de palavras gerada do corpus dos tweets selecionados sobre saúde com 29.015 nós e 101.502 ligações. Esse grafo apresenta também nós com loops, ou seja, possuem palavras ligadas a elas mesmas, e o tamanho do nó representa o seu grau.

Para realizar a análise de grafos, bigramas foram criados a partir das palavras encontradas nas mensagens. O bigrama é uma sequência de duas palavras adjacentes cujo é possível realizar análises de coocorrência. A coocorrência é um grau que descreve com que frequência as duas palavras aparecem juntas.

As medidas de centralidade são métricas que possibilitam identificar e quantificar a importância de um nó ou um grupo de nós em uma rede. Foi calculado a centralidade de grau (degree) dos nós do grafo geral. Como o grafo é direcionado, foi calculado os dois tipos de graus do nós, grau de entrada que é o número de arestas que chegam no vértice e o grau de saída que é o número de arestas que saem do vértice. Além do grau total que é a soma dos graus de entrada e de saída. A Tabela 4 mostra os nós que tiveram maior atividade na rede.

Tabela 4 - Tabela com os valores das medidas de centralidade de grau para as dez palavras que obtiveram os maiores valores de grau total, grau de entrada e grau de saída.

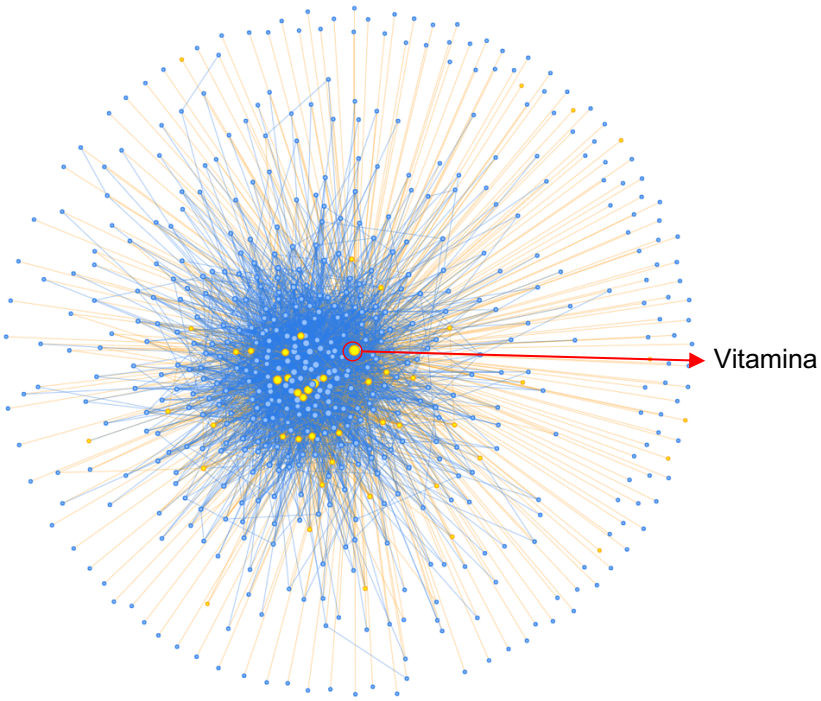
Grau (*all degree*)		**Grau de entrada (*in degree*)**		**Grau de saída (*out degree*)**	
Vitamina	1.596	Dor	1.144	Garganta	1.369
Garganta	1.488	Vitamina	1.004	Gripe	797
Dor	1.344	Acordei	585	Vitamina	592
Gripe	1.244	Febre	567	Febre	558
Febre	1.125	Acho	559	Amarela	542
C	1.000	C	497	C	503
Acordei	724	Gripe	447	Estômago	450
Gente	713	Queria	398	Nervoso	392
TPM	641	Mano	395	Tristeza	381
Tristeza	624	Amanhã	394	Pneumonia	374

É possível observar os nós que tiveram maior prestígio/expansividade através da coluna "in degree", essas palavras receberam mais conexões direcionadas a elas; por outro lado, na coluna "out degree" observou-se as palavras que tiveram maior interesse nas atividades da rede, ou seja, várias ligações saíram desses nós. Praticamente as cinco palavras com maior grau nesta rede são palavras relacionadas a medicamento (vitamina/vitamina C), sintoma (dor, febre, garganta, tristeza) e doença (gripe, febre amarela, pneumonia), o que mostra que as três categorias tiveram destaques na rede. A palavra que se sobressaiu por possuir o maior grau geral, ou seja, possui um maior número de conexões do que as outras foi a "Vitamina". Este tipo de nó é chamado de *hub* por possuir diversas conexões e possibilita distribuir informações para um número maior de nós.

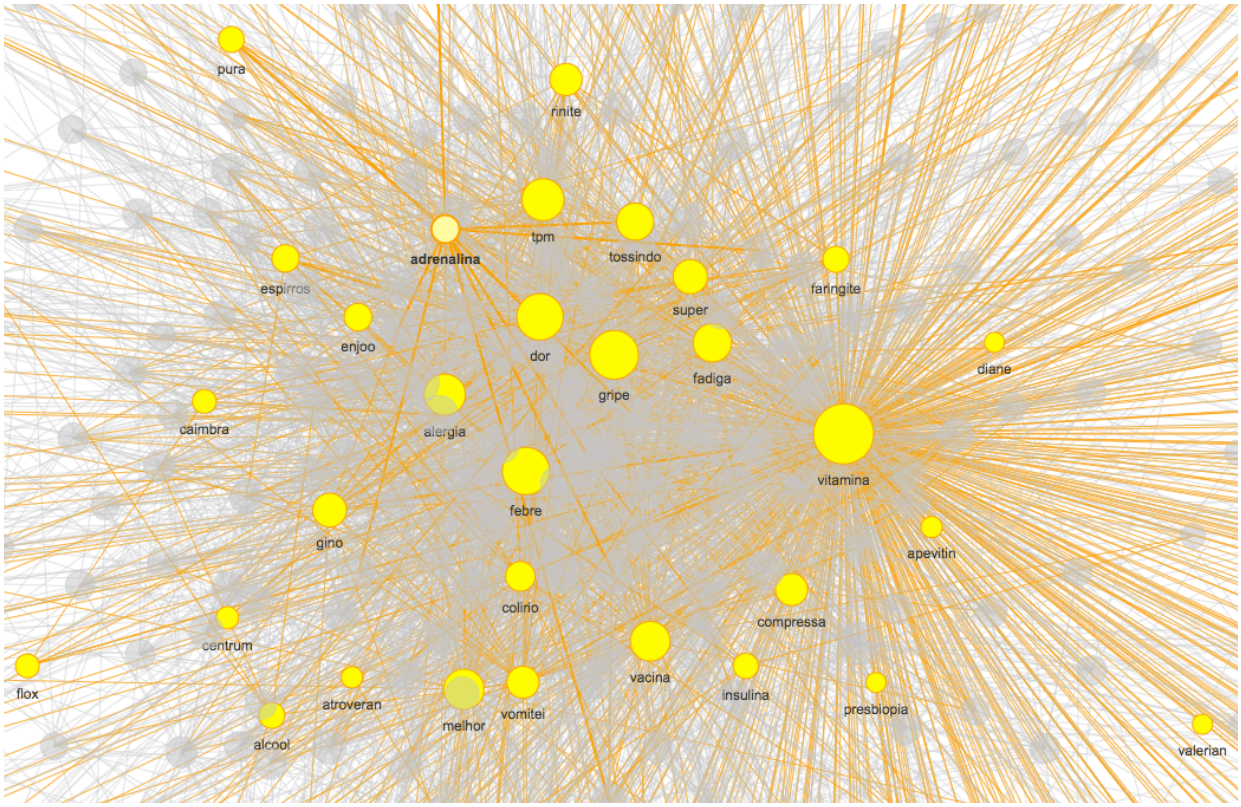
4.3.1 Análise de Subgrafos de Saúde

Para uma investigação meticulosa da rede com a intenção de torná-la visualmente clara e as palavras legíveis, técnicas de particionamento foram aplicadas para extrair subgrafos induzidos a partir dos nós de maior grau. As palavras centrais tendem a coocorrer com um número maior de outras palavras, apresentando uma indicação de importância para conversas do Twitter. Além do mais, examinar subgrafos com um grupo de palavras específicas pode auxiliar na identificação de "frases-chaves" ou conceitos relacionados importantes nas mensagens selecionadas.

Para investigar melhor o nó de maior grau, foi extraído um subgrafo induzido a partir do nó "Vitamina" e seus vizinhos diretos. Para criar uma boa visualização do subgrafo, foi utilizado um pacote do R para visualização e manipulação de rede chamado "visNetwork" que usa a biblioteca javascript "vis.js" (http://visjs.org). Com esse pacote é possível realizar diversas melhorias na estética do subgrafo, como por exemplo, alterar as cores dos nós de acordo com uma classificação. Os nós deste subgrafo possuem palavras de saúde (rotuladas pelas listas controladas de termos de saúde) e aquelas palavras que não estavam nas listas de saúde, portanto ficaram sem classificação. Dessa forma, aproveitando a classificação os nós foram diferenciados em cores, os nós classificados como saúde de amarelo e os não saúde de azul. A Figura 13 mostra o subgrafo induzido a partir do nó "Vitamina".



(A)



(B)

Figura 13 - (A) Subgrafo induzido obtido por um subconjunto de vértices e suas respectivas arestas baseado nos nós vizinhos (adjacentes) do nó que representa a palavra "Vitamina" (ao centro) com 592 nós e 5.557 ligações. (B) Detalhes da ampliação do grafo (A) com apenas os nós (amarelos) que representam as palavras de saúde em evidência. Figura em formato digital interativa disponível em: http://bit.ly/390C6vF

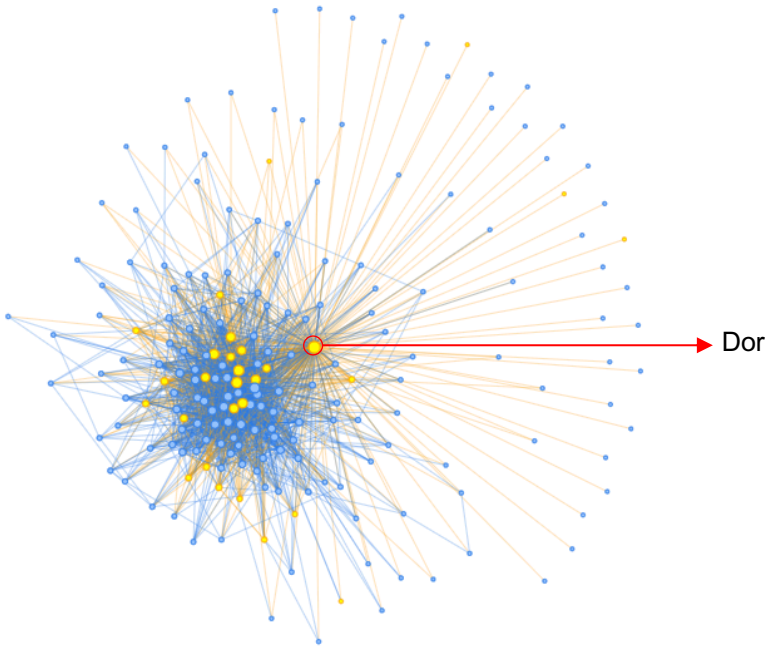
Note que esta visualização gerada é baseada em uma cadeia de Markov[60]. Nesse modelo, comum em processamento de texto, cada escolha de palavra depende apenas da palavra anterior, gerando a rede aleatoriamente seguindo cada palavra. Para tornar a visualização interpretável e analisar os nós de saúde que foram ligados ao nó principal, optamos por mostrar apenas os pares de palavras mais frequentes deixando os nós rotulados como saúde em evidência.

A Figura 13 mostra as conexões que as palavras rotuladas de saúde possuem com o nó principal nesse subgrafo, “Vitamina”. Dentre essas conexões é possível observar vários nós relacionados a sintomas como: enjoo, dor, vomitei, tossindo, espirros, cãibra, tpm, fadiga; assim como ligações com palavras relacionadas a medicamentos, como: centrum (que é um suplemento vitamínico), atroveran, insulina, apevitin (também é um suplemento vitamínico), colírio, vacina, diane, adrenalina; e por fim, os relacionamentos entre doenças: gripe, presbiopia, rinite e faringite. Além desses, nota-se um nó classificado como saúde porém não está relacionado às categorias (sintomas, medicamentos e doenças), refere-se a um tipo de “tratamento”,

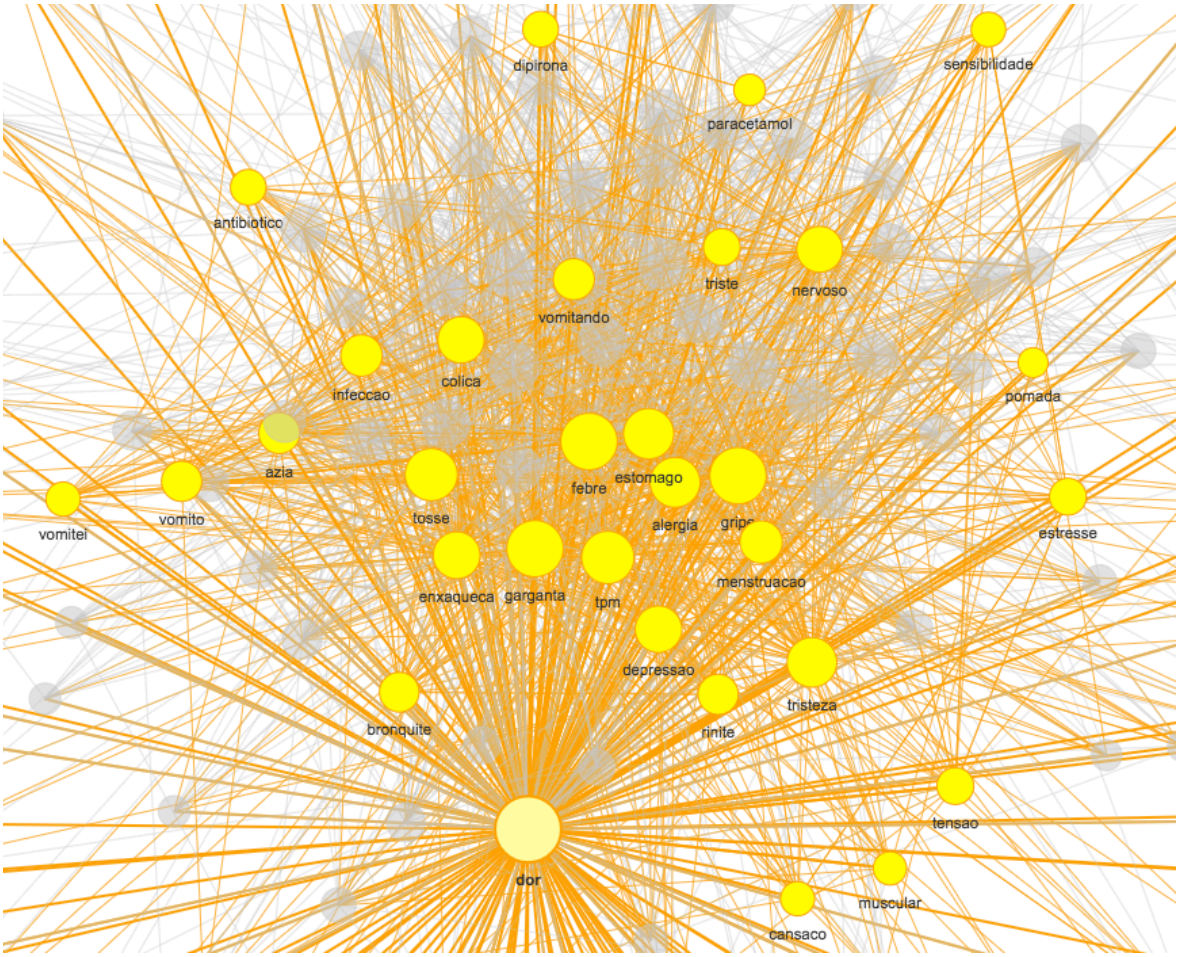
que é o nó “compressa”.

A mesma visualização com subgrafos induzidos baseados em seus vizinhos diretos foi gerada para os nós “Dor”, “Febre” e “Gripe”, nós relacionados a saúde que se sobressaíram com relação ao grau.

A Figura 14 mostra as conexões que o nó “Dor” obteve. “Dor” é um sintoma comum e abstrato, geralmente quando acompanhado de outro sintoma pode identificar uma doença específica. Além disso, a palavra “dor” sozinha não possui um significado semântico significativo, porém acompanhada de uma outra palavra como, por exemplo, partes do corpo, pode gerar um conceito específico. Na Figura 14 (B) é possível identificar algumas ligações relevantes com a palavra dor como: garganta, muscular (palavras que combinadas com a palavra "dor" especificam um sintoma particular de alguma parte do corpo), azia, vômito, tosse, cólica, tristeza (tristeza é considerado um sintoma na CID-10 de código R45.2) e febre (combinações podem levar a melhores significados e relacionar com doenças). Da mesma forma, apresentou diversas ligações com doenças, por exemplo, bronquite, rinite, gripe, depressão, infecção e com determinados medicamentos como, dipirona, paracetamol e antibiótico, que são os medicamentos populares para “dor”, e também relacionado medicamentos com o sintoma "dor" é possível listar certas doenças específicas. Por exemplo: “Dor” e “Garganta” e “Antibiótico” podem indicar uma “Infecção” ou especificamente “Faringite Infecciosa”.



(A)



(B)

Figura 14 - (A) Subgrafo induzido obtido por um subconjunto de vértices e suas respectivas arestas baseado nos nós vizinhos (adjacentes) do nó que representa a palavra "Dor" (ao centro) com 200 nós e 2.573 ligações. (B) Detalhes da ampliação do grafo (A) com apenas os nós (amarelos) que representam as palavras de saúde em evidência. Figura em formato digital interativa disponível em: https://bit.ly/2wur3xJ

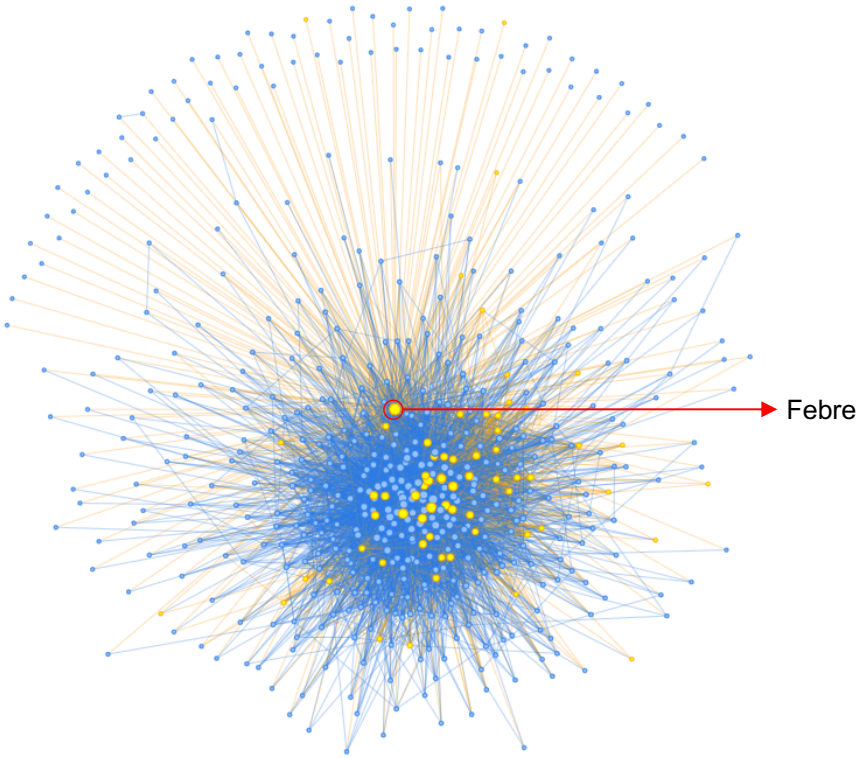
O subgrafo induzido obtido a partir do nó “Febre” (Figura 15) contém uma variedade de doenças relacionadas como: gripe, depressão, câncer, alergia, asma, rinite, catapora, chikungunya, cefaléia, enxaqueca, amigdalite, infecção, constipação, diarréia e torcicolo provendo um indicativo de que esse sintoma “febre” é altamente relacionado pelos usuários a um conjunto de doenças variadas. Vários sintomas como: dor, sono, náusea, calafrios, tontura, espirro, tristeza, apatia e alguns medicamentos como dipirona, antibióticos, xarope, soro e vacina.

Apesar da visualização expor as relações entre as palavras de saúde, há outras relações que foram encontradas no subgrafo da Figura 13 que alteram o tipo do nó de sintoma para doença, por exemplo, esse nó tem ligações com a palavra “amarela” , criando o conceito “Febre amarela”, também há ligações com as palavras “Aftosa” e “Tifóide”, gerando outros conceitos de doenças bacterianas como “Febre tifóide” e até mesmo doenças virais de animais (bovinos, ovinos e suínos), como “Febre aftosa”.

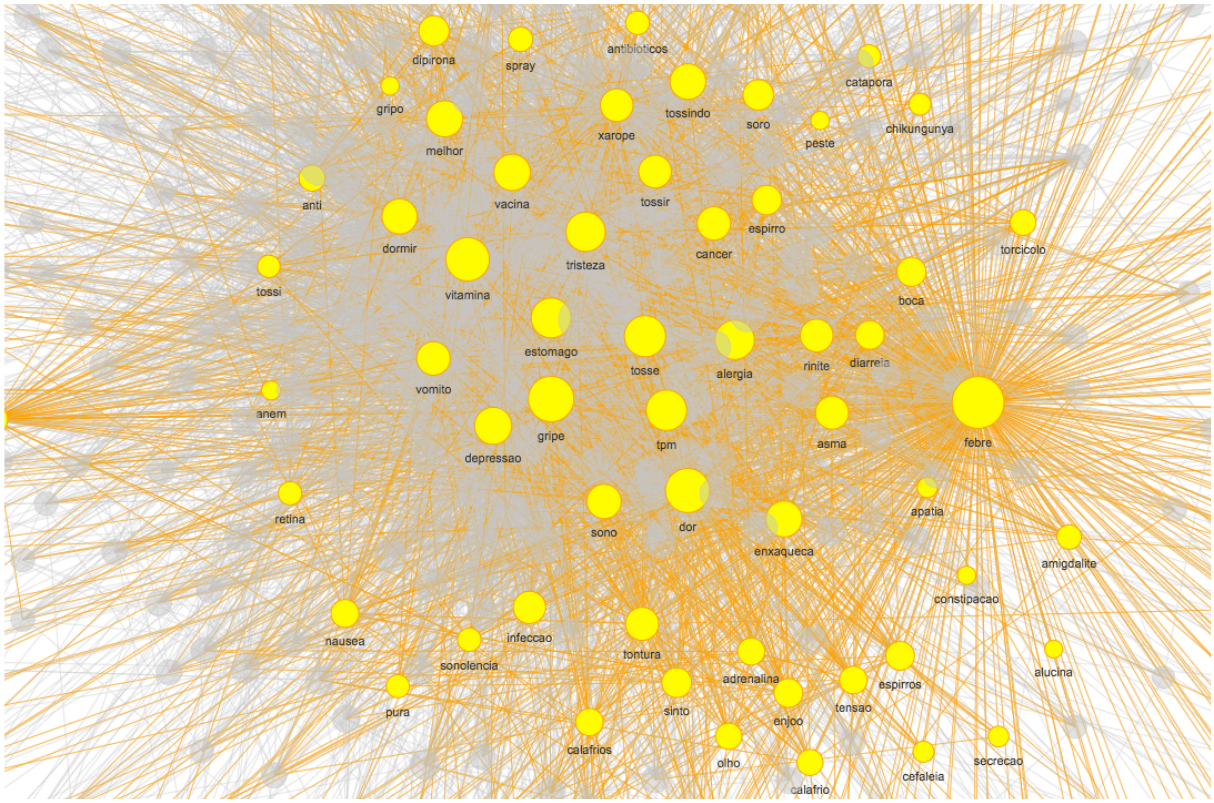
Essas relações foram encontradas analisando a força da ligação que é relacionada a quantidade de vezes que uma palavra se liga a outra. Ou seja, vários outros grupos de relacionamentos entre as palavras podem levar a descobertas de outras doenças. Como o conceito febre amarela já era conhecido no corpus por ter aparecido na nuvem de palavras de *hashtags*, foi realizada uma análise detalhada e minuciosa para o nó “febre”. Coocorrência com palavras não relacionadas à saúde podem também revelar diferentes informações sobre outros contextos nos quais a saúde é discutida.

A centralidade de intermediação (*betweeness*) que identifica e caracteriza os nós com maior vantagem ou poder numa rede, medindo sua importância relativa dentro do grafo, também foi calculada para a rede de palavras. Os dez nós que apresentaram o maior valor de *betweeness* foram: “Vitamina“, “Gripe“, “C“, “Febre“, “Garganta“, “Dor“, “Gente“, “Tristeza”, “Depressão“ e “TPM“.

Com exceção das palavras “Gente” e “C”, todas as outras palavras com maior vantagem na rede estão relacionadas às categorias de saúde (doença, sintomas e medicamentos). Essas palavras são importantes nós de conexão entre os termos e são vértices pontes entre agrupamentos (*clusters*), uma vez que vértices com os maiores valores de *betweenness* tendem a ser acessíveis por caminhos eficientes.



(A)



(B)

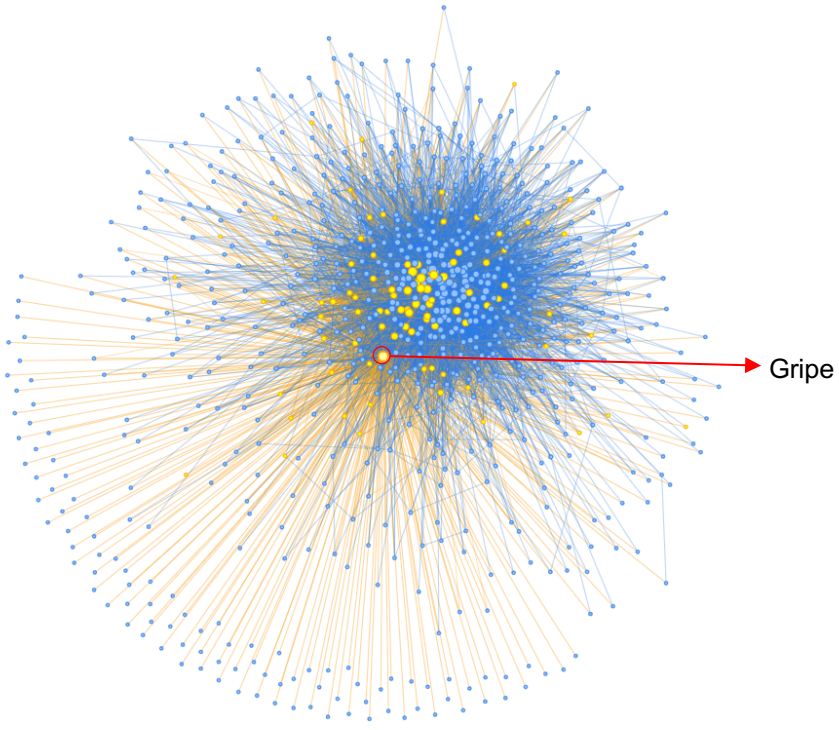
Figura 15 – (A) Subgrafo induzido obtido por um subconjunto de vértices e suas respectivas arestas baseado nos nós vizinhos (adjacentes) do nó que representa a palavra "Febre" (ao centro) com 558 nós e 9.171 ligações. (B) Detalhes da ampliação do grafo (A) com apenas os nós (amarelos) que representam as palavras de saúde em evidência. Figura em formato digital interativa disponível em: https://bit.ly/2Uu9w0D

O subgrafo da Figura 16 mostra as conexões vizinhas do nó “Gripe” que possui um alto valor de *betweeness* sendo considerado um grau ponte entre vários nós, situado nos caminhos geodésicos entre outros nós tendo um grande poder de controlar as informações e o trajeto que elas percorrem. O conceito ”Gripe” é frequente em pesquisas de análise e monitoramento de mensagens de redes sociais, como é uma doença praticamente recorrente nas pessoas ao longo do ano, acredita-se que esse termo sempre irá aparecer nas mensagens se procurado como palavra-chave.

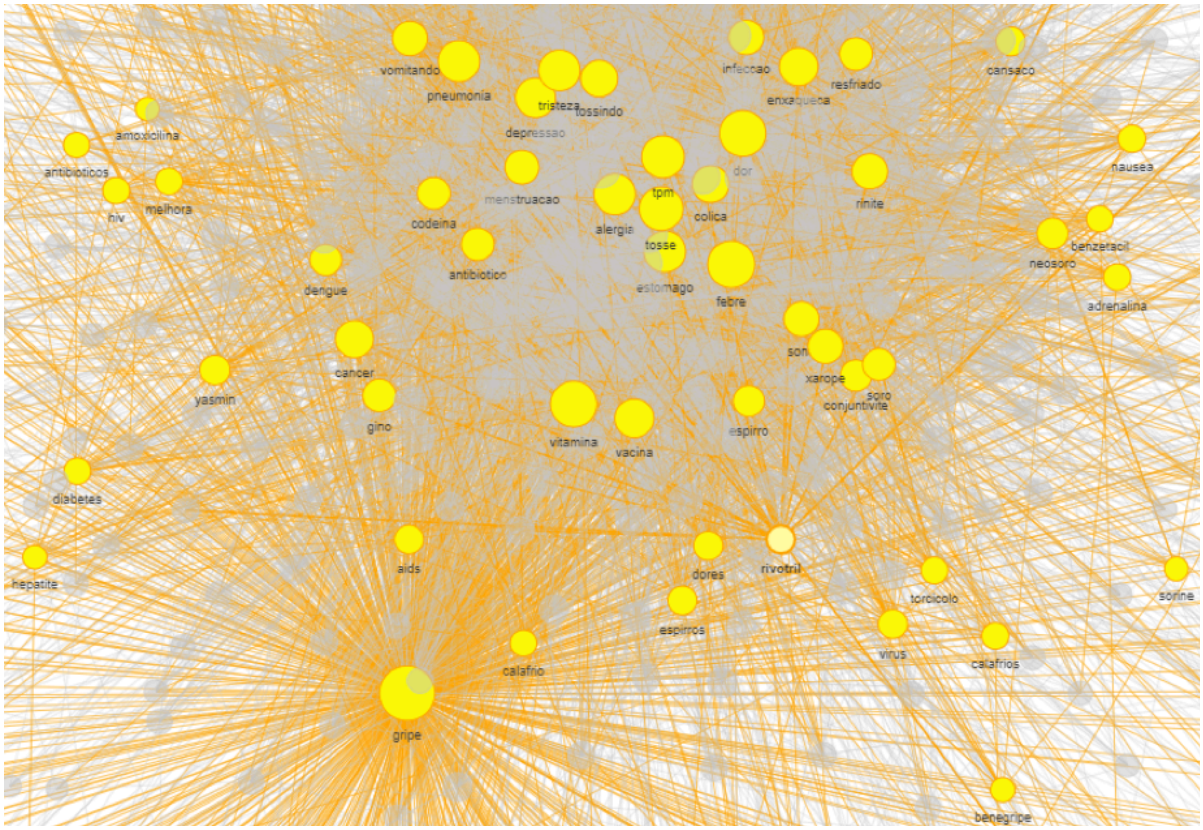
Na Figura 16 observa-se o nó “Gripe” com várias ligações com nós relacionados a sintomas, outras doenças e medicamentos. Uma relação que já foi abordada em um estudo[3] realizado nos EUA foi sobre os usuários publicarem no Twitter mensagens relacionando a doença “Gripe” com medicamentos como “Antibiótico“. Nesse subgrafo essa relação também ocorreu mostrando que a palavra “Gripe” teve uma relação de adjacência com a palavra “Antibiótico” e “Amoxicilina”, ou seja, apareceram numa mesma mensagem. O que supõem que pessoas podem estar tomando esse tipo de medicamento que é utilizado em doenças bacterianas, para

tratar uma doença viral. Além disso, esse nó teve ligação com outros medicamentos que geralmente não são tão comuns para tratamento de gripe, como, “codeína” e “rivotril”. Os sintomas comuns de gripe foram encontrados nas ligações entre os nós, como, “Febre”, “Dores”, “Espirro”, “Calafrio”, “Tossindo”, “Vomitando” entre outros. Vários medicamentos populares para gripe também foram notados na rede como: sorine, neosoro, benzetacil, benegripe, vitamina, vacina e xarope; esse nó foi o que se conectou com a maior quantidade de medicamentos variados comparando com os outros nós analisados. Além de também mostrar alguma relação com doenças como hepatite, diabetes, aids, conjuntivite, rinite, resfriado e enxaqueca.

Para calcular a centralidade de proximidade (*closeness*) é importante notar que em grafos desconexos essa medida não é bem definida. Se existe na rede um nó que não possui nenhum caminho direto ou indireto para outro nó (ou seja, eles pertencem a componentes separados), então a distância entre esses nós é infinito. Se pelo menos um nó é inacessível pelos outros, a distância para todos os outros nós é infinita. Por isso, nesses casos de grafos desconexos a medida de proximidade é limitada ao maior componente de nós (isto é, o maior componente encontrado após aplicar medidas para encontrar agrupamentos no grafo).



(A)



(B)

Figura 16 - (A) Subgrafo induzido obtido por um subconjunto de vértices e suas respectivas arestas baseado nos nós vizinhos (adjacentes) do nó que representa a palavra "Gripe"(ao centro) com 797 nós e 12.063 ligações. (B) Detalhes da ampliação do grafo (A) com apenas os nós que representam as palavras de saúde. Figura em formato digital interativa disponível em: https://bit.ly/3bbJNRo

Mesmo assim foi calculado o coeficiente de agrupamento da rede (transitividade), que mede a probabilidade de que nós adjacentes da rede sejam conectados. Para isso, foi utilizada a função *transitivity* do R e o valor encontrado foi de 0.04929891, o que significa que há uma presença baixa do número de triângulos na rede, e assim, poucos agrupamentos fortes são encontrados.

A forma mais direta de particionar uma rede para encontrar os agrupamentos é calcular os componentes conectados (fracos ou fortes). Para isso, foi utilizado a função “componentes” implementada no R. Por meio dessa função, 571 componentes foram descobertos, sendo um deles o componente “gigante” (*giant connected component*) que, como na maioria das redes, inclui a maioria dos nós. O componente gigante segue a estrutura do grafo original, portanto também apresenta estrutura de comunidade centro-periferia, assim não sendo possível quebrar em comunidades menores.

Depois de extrair e analisar as medidas de centralidade foi criada uma nova

visualização do grafo para poder analisar as relações mais fortes entre as palavras baseadas no peso das ligações. O peso das arestas está relacionado a quantidade de vezes que as palavras (nós) coocorreram num documento (tweet).

Assim, foi gerado um grafo das palavras que tiveram uma quantidade de ligações maior ou igual a 20. Analisar as coocorrências entre as palavras possibilita descobrir tópicos interessantes que os usuários publicaram relacionados à área da saúde.

No grafo da Figura 17 é possível observar certos detalhes da estrutura do texto e das ligações entre as palavras. Nessa figura os nós "garganta", "dor", "gripe", "febre" e "vitamina" formam centros comuns de nós que possuem diversas ligações e ligações fortes entre si.

Um detalhe notório nesta visualização é o nó "vitamina", que possui um alto valor de *betweenes*, mostrando ser um "nó " importante para conexão entre os termos. Esse nó tem uma ligação forte com o nó "C" (formando o conceito "vitamina c") que possui uma ligação forte com o nó "dor" que está em um outro lado no agrupamento de palavras. Além disso, o nó "vitamina" ainda se liga a outros nós que estão vinculados a outros agrupamentos, relacionamento essa palavra a "complexos vitamínicos" e também a vitamina de frutas, relacionando a alimentação.

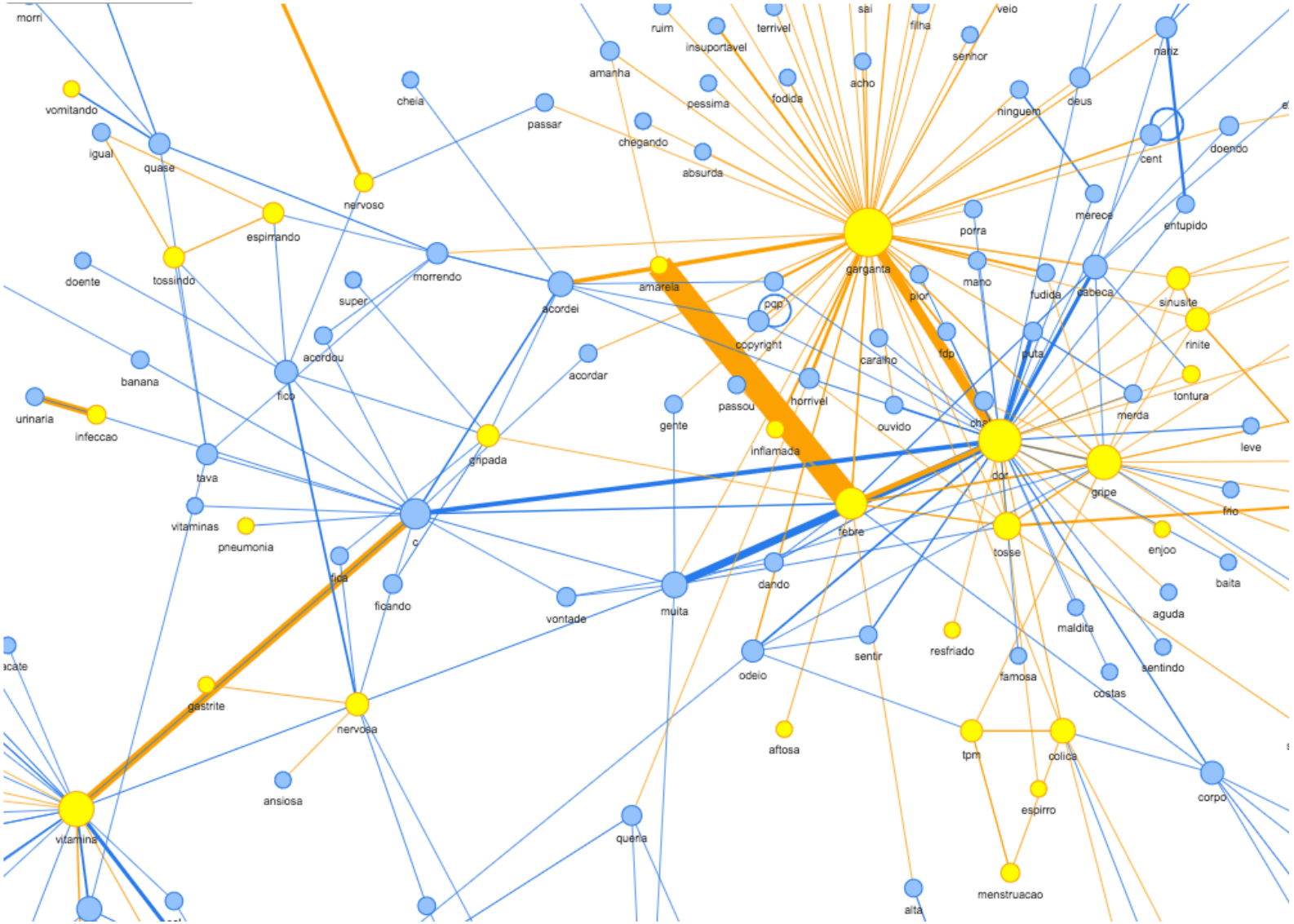


Figura 17 - Grafo ponderado que possui ligações com peso maior ou igual a 20 contendo 452 nós e 501 ligações. Quanto mais grossa a aresta, maior a quantidade de ligações que aquelas palavras possuem. O grau do nó também é representado pelo tamanho do nó. Figura em formato digital interativa disponível em: https://bit.ly/2wlteDW

Também há um grande destaque entre os nós "febre" e "amarela", como uma das ligações mais fortes do grafo, além dessa ligação formar um conceito (a doença febre amarela), o fato dessas palavras terem essa forte conexão representa que este conceito foi relevante nos tweets selecionados. O nó "febre" também se apresentou fortemente conectado aos nós "dor", "gripe" e "tosse", que mostra a frequência com que esses termos aparecem juntos em discursos sobre saúde no Twitter.

No caso desse grafo ponderado o nó que se apresentou com maior grau foi o "garganta". Este nó praticamente faz ponte com os outros nós de destaque na rede, mas principalmente ao nó "dor". O nó "dor" apresentou fortes conexões com palavras que não foram classificadas como saúde, por exemplo, "cabeça", "costas, "ouvido", "corpo" e advérbios de intensidade, como a palavra "muita" e adjetivos como "pior", "odeio", "horrível", "maldita". Mas este nó também teve várias conexões com palavras de saúde, por exemplo, doenças como "gripe", "sinusite", "rinite"; alguns sintomas, "febre", "enjoo", "tontura", "cólica", entre outros.

Alguns triângulos são identificados na Figura 17, como por exemplo, os nós "menstruação", "tpm" e "cólica" que estão diretamente relacionados ao "tema" saúde da mulher, especificamente associados a menstruação e seus sintomas. Outro triângulo contém os nós "espirrando", "tossindo", "igual", sintomas típicos de gripe/alergias relacionados, assim como “dor, “cabeça”, “tontura”.

4.4 Análise Exploratória De Tópicos

Após explorar o conteúdo das mensagens minerando o texto, analisando as conexões entre as palavras e identificando seus relacionamentos, foi empregado a abordagem de modelo de tópico para caracterizar os principais temas de discussão relacionado à saúde nas mensagens do Twitter. Apesar do texto do Twitter ser uma mensagem curta seu conteúdo pode apresentar mais de um assunto.

Foi utilizado o modelo matemático LDA para encontrar a mistura de palavras que está associada com cada tópico nos documentos. A modelagem de tópico também foi executada no software R utilizando os valores dos parâmetros ajustados citados no Capítulo 3 (Métodos). Cada “tópico” tem uma distribuição de probabilidade por palavras, estimada a partir dos dados, e os tópicos são geralmente representados pela apresentação das 10 a 20 palavras mais prováveis no tópico.

Depois de executar o modelo, para obter uma melhor visualização dos 50

tópicos gerados foi aplicado um método de visualização e interpretação de tópicos conhecido como LDAVis[61]. Esse método calcula a frequência dos tópicos, a distância entre os tópicos e projeta os tópicos em um plano de duas dimensões para representar a similaridade entre eles. Além disso possui uma barra de ajuste com um parâmetro que contém uma métrica de relevância (λ) que varia $0 \leq \lambda \leq 1$, fazendo um ajuste de relevância da palavra no tópico. Os valores de λ próximo a 1 fornecem classificações de alta relevância para termos frequentes dentro de um determinado tópico, enquanto valores de λ próximos de zero proporcionam classificações de alta relevância para termos exclusivos dentro de um tópico. Os termos são listados em ordem decrescente de relevância. Em seguida é criado uma visualização que é aberta em um navegador com uma interface interativa. Esta visualização auxilia na interpretação dos tópicos revelando aspectos dos relacionamentos entre os termos e tópicos de maneira compacta e simplificada.

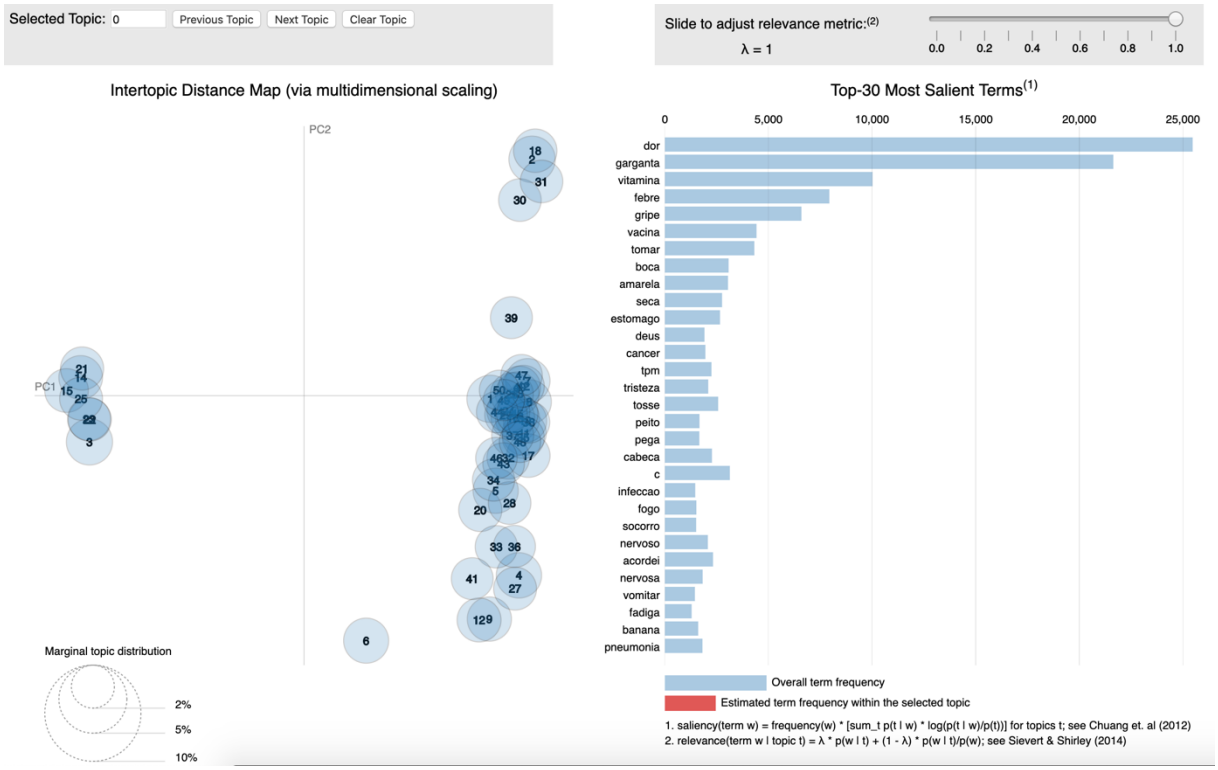


Figura 18 - Interface gerada pelo LDAvis apresentando uma visão global dos tópicos e seus relacionamentos (à esquerda), e o gráfico barras vertical com 30 termos mais salientes nos tópicos (à direita). Disponível em: https://bit.ly/39dJJPJ

Foram examinadas as 10 palavras mais prováveis de cada tópico. Com base nas palavras-chave fornecidas, o significado de cada tópico foi interpretado e rotulado manualmente pela pesquisadora com uma palavra ou frase que descreve o grupo de palavras do tópico.

De todos os tópicos, 43/50 (86%) foram identificados como coerentes, sendo 34/50 (68%) relacionados à saúde e 9/50 (18%) não relacionados à saúde. Os 34 tópicos de saúde foram agrupados em 10 categorias de alto nível: Dores no Corpo, Câncer e Doenças Relacionadas, Saúde Mental, Vitaminas e Suplementos, Saúde da Mulher, Medicamentos e Medicina Alternativa, Vacinas e campanhas, Alergias e Doenças Respiratórias, Saúde da Visão e Doenças Sazonais e Compulsórias.

Tabela 5 – Combinação das top 10 palavras dos tópicos de saúde que melhor descrevem cada categoria.

Categorias	Dores do Corpo	Câncer e doenças relacionadas	Saúde Mental	Vitaminas e Suplementos	Saúde da Mulher
Palavras	Dor	Câncer	Tristeza	Vitamina	TPM
	Garganta	Pulmão	Depressão	C	Menstruação
	Cabeça	Mama	Chorar	B	Cólica
	Ouvido	Pele	Raiva	Cabelo	Enxaqueca
	Dente	Diabetes	Feminismo	Carne	Nervosa
	Costas	Próstata	Violência	Sol	Humor
	Voz	Infantil	Vomitar	Sais	Agitada
	Febre	AVC	Síndrome	Suplemento	Tensão
	Corpo	Diabetes	Pânico	Vitamínico	Infeção
	Inflamada	Hipertensão	Ansiedade	Cálcio	Urinária
Categorias	Medicamentos e Medicina Alternativa	Vacinas e Campanhas	Alergias e Doenças Respiratórias	Saúde da Visão	Doenças Sazonais e Compulsórias
Palavras	Xarope	Vacinação	Neosoro	Retina	Vírus
	Dipirona	Gripe	Sorine	Miopia	Zika
	Paracetamol	Saúde	Nariz	Astigmatismo	Febre
	Chá	Campanha	Alergia	Rotina	Amarela
	Camomila	Braço	Pneumonia	Óculos	Chikungunya
	Boldo	Dolorido	Gripada	Hipermetropia	Sífilis
	Mel	Tétano	Rinite	Grau	Catapora
	Compressa	HPV	Asma	Visão	HIV
	Quente	Hepatite	Bronquite	Frontal	Dengue
	Gelo	Feira	Sinusite	Vista	Caxumba

Um exemplo dos tópicos de saúde organizados por categoria é mostrado na Tabela 5. Esta tabela mostra as top 5 palavras combinadas dos tópicos que levaram a essa classificação. Por meio dessa tabela é possível observar a variedade de tópicos que foram discutidos nas mensagens do corpus no período selecionado. Diferentes tipos diferentes de câncer, assim como várias doenças sazonais foram encontradas. Os tipos de dores mais discutidos no Twitter também foram revelados. Apesar da “Dor de garganta” ter se destacado nas demais análises, outras dores como cabeça e ouvido também se mostraram presente. Os tópicos não relacionados à saúde foram classificados nas seguintes categorias: Drogas (tópicos 1, 39), Religião (tópico 45), Escola (tópico 37), Esporte (tópico 26), Política (tópico 24), Famosos (tópico 13) e Alimentos (tópicos 2, 18).

O modelo também calcula a distância entre os tópicos em um gráfico de dispersão, que aproxima o relacionamento semântico entre os tópicos com base na técnica *multidimensional scaling* (MDS). Essa técnica facilita a visualização do nível de similaridade ou distância entre os objetos investigados, no caso deste estudo, os tópicos encontrados. A distância entre tópicos é calculada usando a divergência de Jensen-Shannon[61].

Através da interface interativa, na parte esquerda (Figura 19) é possível identificar um fenômeno comum de acontecer com os modelos de tópicos não supervisionados, vários *clusters* apresentam um conteúdo semelhante, assim há tópicos que estão sobrepostos. Entre os quadrantes notamos também que os tópicos relacionados a doenças, medicamentos e sintomas estão ligeiramente separados entre si, no primeiro quadrante pode ser observado um aglomerado de tópicos sobre a grande categoria “Vacinas e Campanhas” que pode ser relacionado a medicamentos, já o segundo e terceiro quadrante foram encontrados a maioria dos tópicos sobre sintomas (maioria relacionado a categoria “Dores no Corpo”) e no quarto quadrante foram encontrados a maioria dos tópicos sobre doenças e tópicos não relacionado a saúde.

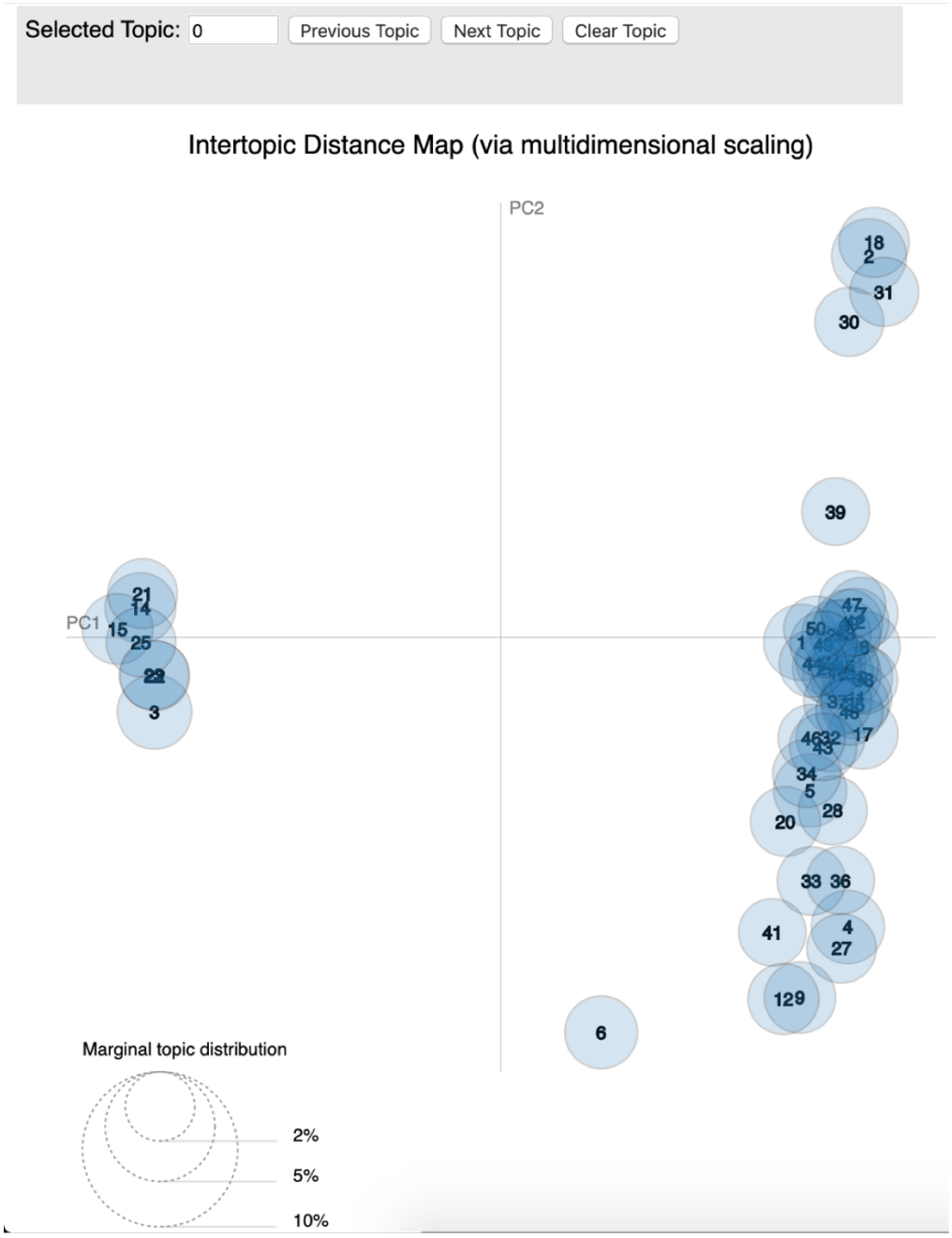


Figura 19 - Mapa de distância entre os tópicos baseado na técnica *multidimensional scaling*.

A parte direita da visualização (Figura 20) oferece a opção de identificar as palavras-chave específicas de determinados tópicos. A especificidade da palavra-chave é calculada como a razão entre a frequência da palavra-chave em um determinado tópico e a frequência geral da palavra-chave no corpus geral[62], dado um parâmetro de ponderação λ. Diminuir o valor do parâmetro de peso revela palavras-chave específicas do tópico; aumentar o valor do parâmetro revela palavras-chave comuns a todo o corpus[61].

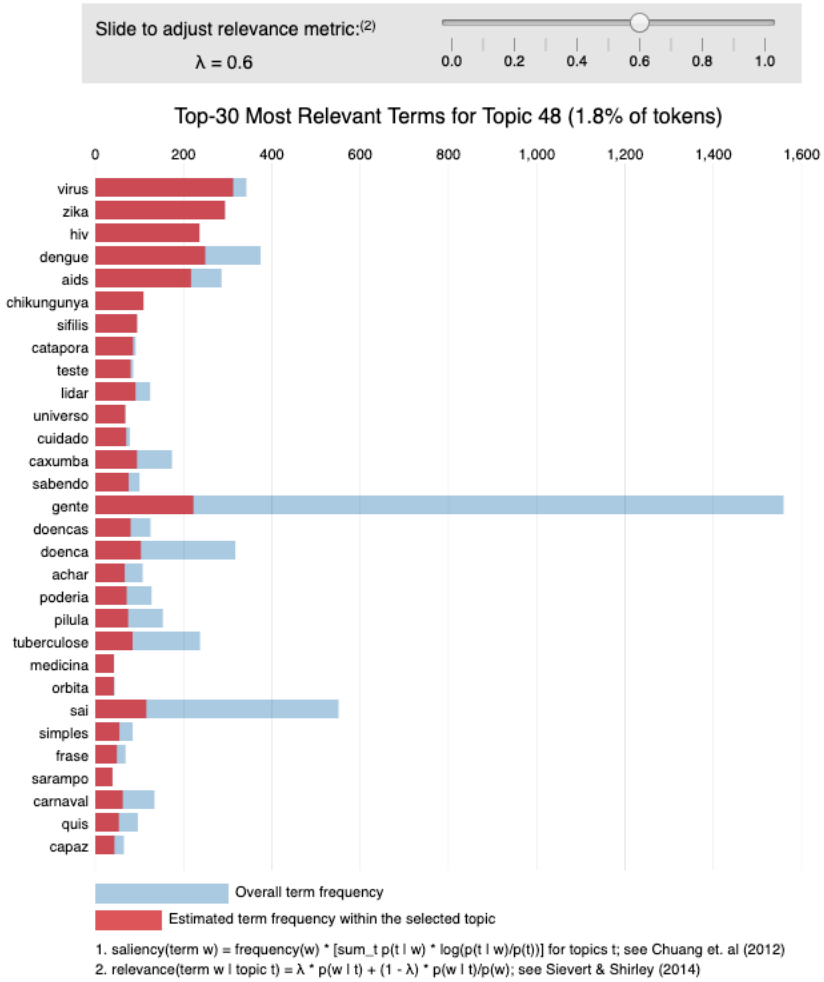


Figura 20 - Top 30 termos mais relevantes para o tópico selecionado onde a frequência estimada do termo é representada pela barra vermelha.

Com base na recomendação de Sievert e Shirley[61], em nesta tese a interpretação dos resultados foi realizada com λ = 0,6, mas apoiada pela observação dos resultados, considerando também diferentes valores de lambda.

5 Discussão

Este capítulo apresenta a discussão dos resultados descritos no capítulo anterior.

5.1 Coleta e Seleção dos Dados

Um dos grandes problemas para realizar uma pesquisa em dados proveniente do Twitter ou de qualquer outra rede social é a seleção de um conjunto de dados específico para o problema que se quer tratar. Apesar do próprio Twitter conter atributos que auxiliam a busca e seleção por um determinado assunto, como as *hashtags*, essas ainda não são suficientes para coletar todas as mensagens que os usuários estão postando naquele momento no Twitter, devido ao fato de que nem todos os usuários fazem uso das *hashtags*.

A maioria dos estudos escolhe um tópico particular, como por exemplo, "gripe" e criam um conjunto de palavras-chaves (termos relacionados ao assunto) para realizarem a busca e seleção por mensagens sobre aquele tópico. Mesmo assim, essa seleção não garante que todas as mensagens que possuem pelo menos uma das palavras-chaves estão relacionadas ao assunto.

No Twitter encontra-se mensagens com ironia, ambiguidade, uso de textos com linguagem figurada, e isso cria uma dificuldade maior em selecionar as mensagens que vão conter o conteúdo importante para análise do assunto. Por exemplo, no estudo de Araujo et al.[2] foi encontrado mensagens com a palavra "câncer" que não dizem respeito à saúde, identificando um fenômeno encontrado no idioma português brasileiro que faz uso de palavras livremente como um adjetivo. Além disso, encontra-se ambiguidade semântica que é quando a mesma palavra pode ser usada para referir-se a diferentes conceitos, ou seja, tem a mesma escrita, porém significados diferentes. Como também no caso da palavra "câncer" que é uma doença, mas também possui um significado que está relacionado ao horóscopo, causando dificuldades na coleta e análise dos dados sobre a doença, como também apresentado no estudo de Araujo et al[2].

Nesta tese foi definida uma estratégia de selecionar apenas mensagens que continham pelo menos duas palavras de saúde. Como o escopo não foi delimitado em somente um assunto de saúde, mas em vários assuntos relacionados a algumas listas controladas de doenças, sintomas e medicamentos, as palavras-chaves para seleção

das mensagens foram os próprios termos encontrados nas listas. Pelo fato do Twitter possibilitar que os usuários publiquem mensagens curtas de até 140 caracteres, ter pelo menos duas palavras relacionadas a saúde oferece uma seguridade de que o conteúdo estará relacionado a saúde. Pode haver algumas perdas com relação às mensagens, porém evita elaborar métodos para eliminar ruídos de mensagens que não apresentam conteúdos relacionados à saúde. No dia 7 de novembro de 2017 foi anunciado pelo Twitter o aumento do limite de caracteres da mensagem de 140 para 280. Essa mudança se deu pelo fato de o Twitter ter refletido sobre as diferenças entre os idiomas e as dificuldades de adequar um pensamento a um tweet que apresenta um espaço tão pequeno para o usuário se expressar. Na época da coleta das mensagens selecionadas neste estudo o número de caracteres ainda era limitado a 140.

A seleção de mensagens de redes sociais apenas por uma palavras-chave é um dos maiores problemas relatados nos trabalhos que buscam filtrar mensagens de um determinado assunto. A maioria dos trabalhos acabam recorrendo a uma validação manual para assegurar que a mensagem é sobre saúde ou não.

Sobre as listas escolhidas para selecionar as mensagens sobre saúde, a escolha do uso da CID-10 se deu basicamente por ser a terminologia oficial utilizada nas consultas médicas no sistema público e privado de saúde no Brasil. Com isso, acredita-se que o paciente tenha, de certa forma, um conhecimento básico sobre essa terminologia.

Existem outros vocabulários controlados de saúde, como por exemplo, o MeSH[63] citado e utilizado em certos dos estudos referenciados na revisão da literatura. Porém, o MeSH é um descritor geralmente aplicado na redação e indexação de artigos científicos, portanto é um vocabulário mais formal e não muito conhecido. A intenção desta pesquisa foi trabalhar com dados de redes sociais (textos informais), dessa forma optou-se por utilizar uma lista de palavras de saúde (doenças e sintomas) comum. Apesar disso, a CID-10 possui um vocabulário estritamente técnico. A CID-10 e a Brasíndice (lista de medicamentos) foram as que mais apresentavam termos complexos e formais comparado as demais listas.

5.2 Mineração de dados e Análise Quantitativa do Conteúdo

O pré-processamento dos dados é uma etapa determinante e significativa para

a realização efetiva da mineração de dados. Essa etapa foi recalculada várias vezes a fim de se obter um bom resultado na organização, limpeza e estruturação dos dados. Ainda assim é possível que haja melhorias nas tarefas de limpeza e transformação que possam contribuir para melhorar os resultados das análises.

Esse processo também colabora para o entendimento dos dados antes de qualquer análise. Entender o conjunto de dados que irá analisar e ter conhecimento das palavras frequentes, através dos gráficos e nuvens de palavras, já direciona o rumo das próximas análises do estudo.

A nuvem de palavras consolida em uma visualização simples as informações importantes do conjunto de dados analisado. Quanto mais uma palavra específica aparece no conjunto de dados maior ela aparecerá na nuvem de palavras. Com isso, essa visualização pode apresentar os assuntos mais pertinentes encontrados naquele conjunto de dados nortear análise específicas nos mesmo. As nuvens de palavras apresentadas nesta tese mostraram de forma pertinente os assuntos mais relevantes no conjunto de dados de saúde analisado, porém não foi possível distinguir por cores as palavras relacionadas a saúde e as não relacionadas a saúde.

5.3 Análise de Redes Sociais

A linguagem escrita é composta de cadeias lineares de palavras, sendo assim, a forma mais natural de se associar palavras é conectar palavras adjacentes. Grafos são basicamente nós e arestas, porém podem gerar uma representação vigorosa onde as relações entre os objetos representados tornam-se informações úteis e relevantes.

Neste estudo foram construídos grafos e subgrafos de palavras encontradas nos tweets relacionados a saúde. Nesses grafos foi possível classificar as palavras em saúde e não saúde, o que auxiliou na interpretação das relações de interesse que eram estritamente relacionadas aos termos de saúde. Em particular, a divisão entre as categorias doença e sintoma pode ser incerta em diversos casos. Por exemplo, pode não ser claro inicialmente se palavra “febre” apenas poderia ser um sintoma, mas “febre amarela” está relacionada a doença.

Trabalhar com um grafo com uma grande quantidade de nós e inúmeras conexões não é uma tarefa trivial. A redução do grafo por indução de subgrafos baseados nas palavras relevantes encontradas pelas medidas de centralidade

auxiliaram na análise qualitativa do grafo. E a redução por peso das arestas reforçou e evidenciou ainda mais certas conexões importantes.

A análise dos subgrafos mostrou quão importante é explorar a conexão entre as palavras para compreender uma narrativa ampla e característica dos assuntos discutidos sobre saúde. Os assuntos encontrados pelos triângulos dão ênfase para certos tópicos específicos de saúde como os citados na Figura 17. Os adjetivos conectados as palavras de saúde revelam algumas características sobre o que os usuários pensam daquele termo, podendo estar relacionados ao sentimento do usuário. A palavra “garganta” por exemplo, apresentou várias ligações com adjetivos como “péssima”, “terrível”, “insuportável”, “ruim”, “horrível” indicando uma tendência de expor esse assunto com uma certa enfase negativas.

Informações adicionais encontradas nos grafos levam a certas interpretações. Por exemplo, a palavra “Vitamina” que apresentou o maior grau geral, ou seja, possui um maior número de conexões, provavelmente não apareceu em uma mensagem isolada, mas sim em diversas mensagens com diferentes contextos que podem ter relação com suplementos alimentares. O uso desses dados para realização de um estudo característico e específico sobre nutrição e a utilização de suplementos vitamínicos podem ser capaz de mostrar como a nutrição é discutida e quais aspectos específicos relacionados a esse tema.

Também foi observado um nó específico relacionado a uma bebida considerada um medicamento natural consumido pelas pessoas para tratamento de alguns sintomas ou doenças, que foi o nó com a palavra “Chá”. Como mostra a figura abaixo (Figura 21), o nó chá apresentou uma forte conexão com a palavra “relaxante”, e também com outras palavras não consideradas da saúde, mas que revelam características relacionadas a saúde como “curador” e “calmante”.

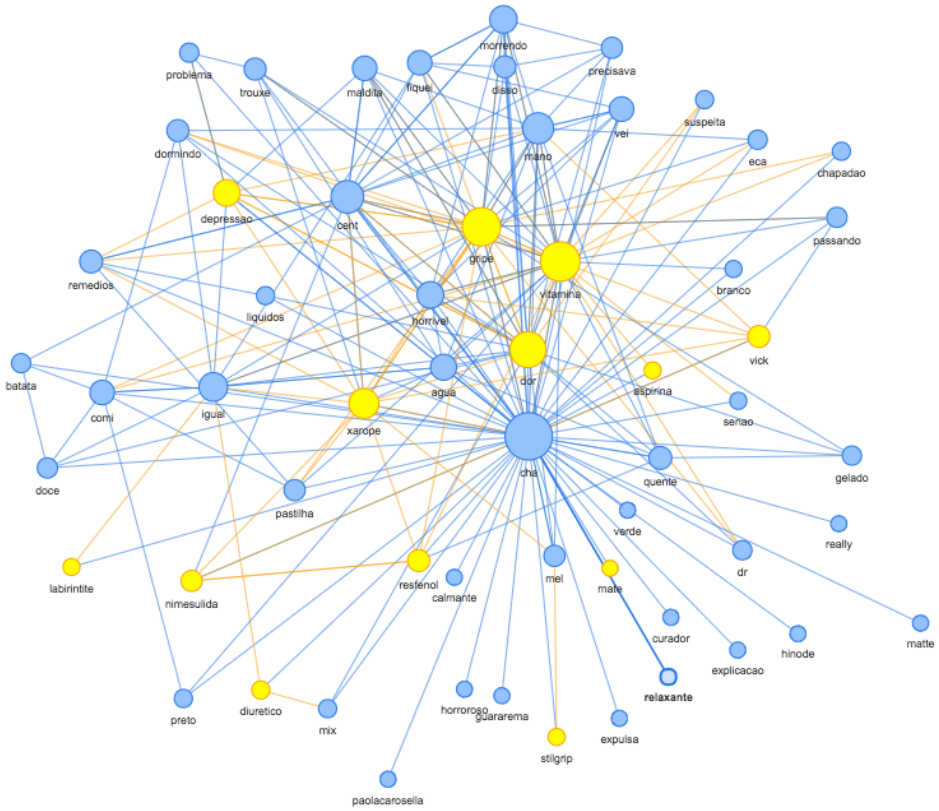


Figura 22 - Subgrafo induzido obtido por um subconjunto de vértices e suas respectivas arestas baseado nos nós vizinhos (adjacentes) do nó que representa a palavra "Chá".

Além disso, as doenças conectadas com esse nó foram “gripe”, “depressão” e “labirintite”. Existem diversas receitas de chás populares como remédios naturais para esses tipos de doenças. Ademais, existem pesquisas científicas[64] que comprovam os poderes terapêuticos do chá e seus diversos benefícios com diferentes doenças. No subgrafo também foi possível observar as conexões com medicamentos e alguns deles como por exemplo, “stilgrip” e “vick” também podem ser encontrados como chás com a função de analgésico e antitérmico.

Coocorrências com palavras não relacionadas à saúde podem fornecer insights sobre outros contextos nos quais a saúde é discutida. Uma das categorias da modelagem de tópicos relacionadas a Drogas (não considerada de saúde) também foi encontrada na análise da rede, porém as palavras relacionadas a esse tópico estavam em nós em partes desconectadas do grafo. Contudo foi possível visualizar todas as palavras conectadas.

5.4 Análise Exploratória dos Tópicos

Na modelagem de tópicos é importante encontrar a configuração adequada para

o seu modelo de dados, assim o método irá produzir tópicos elucidativos e coerentes. O desempenho do método varia de acordo com a quantidade de iterações que é definida, porém a qualidade dos resultados do LDA depende da quantidade e qualidade dos dados. Baseado nos resultados mostrados, a execução com 50 tópicos apresentou uma coerência maior.

Em essência os modelos de tópicos capturam as informações de coocorrência das palavras e essas palavras com alta coocorrência e similaridade são colocadas juntas para compor um tópico. Os resultados da modelagem de tópicos mostram que um conjunto amplo e diversificado de tópicos de saúde é discutido no Twitter. Os tópicos que são considerados de tendência são assuntos e conteúdos que têm um alto volume de postagens na mídia social. Os tópicos encontrados convergiram nos assuntos analisados nas nuvens e rede de palavras gerada. Embora as palavras-chave individuais tenham sido tão boas quanto os modelos de tópicos, os modelos de tópicos podem ajudar na identificação de palavras-chave, principalmente para palavras menos óbvias usadas no Twitter, e podem organizar automaticamente diversas palavras em um pequeno número de tópicos. Novos tópicos com assuntos não observados nas outras análises foram relevados, mostrando o êxito na utilização das técnicas combinadas para extração de conhecimento me dados textuais.

Vários tópicos apresentaram uma informação valiosa do contexto dos discursos gerados, inclusive, houve mais de um tópico relacionado a “Febre Amarela” e “Vacina”, eventos em destaque na época. Entre julho de 2016 e junho de 2017, foi registrado um dos eventos mais expressivos da história da febre amarela no Brasil segundo Informe Especial Febre Amarela no Brasil Nº 01/2017 (http://portalarquivos.saude.gov.br/images/pdf/2017/marco/18/Informe-especial-COES-FA.pdf).

A palavra “Vacina” foi encontrada tanto vinculado a febre amarela quanto a gripe. Vacinas para esses dois tipos de problemas de saúde ainda são bastante populares e a população parece mostrar interesse em compartilhar esse assunto tanto com mensagens de divulgação de locais para vacinar quanto com as experiências ao tomar a vacina ou apenas dizer que tomou a vacina. Mesmo com os rumores de um grupo de pessoas anti-vacina, por meio desta análise notou-se que em casos de surtos (como o da Febre Amarela) as pessoas se conscientizam e buscam orientar outras para também se vacinarem. A palavra “reação” foi uma das palavras com alta frequência em um dos tópicos sobre febre amarela, o que pode nos levar a uma

interpretação que os usuários também discutiram sobre preocupações com as reações que a vacina de febre amarela pode causar.

Com relação as categorias em que os tópicos foram organizados, vários tópicos apresentaram o assunto de forma clara, porém alguns ainda continham uma mistura de palavras relacionada a diversos assuntos tornando-os complexos para rotular.

O Tópico 19, rotulado como “Saúde Mental” abordou uma temática importante pois além de estar relacionado a saúde mental com a palavra mais relevante do tópico sendo “depressão” também abordou assuntos como “feminismo”, “violência” e “mulher” que pode gerar um alerta para uma investigação sobre quais narrativas são compartilhadas nas mensagens das redes sociais sobre esse assunto tão delicado, especialmente para mulheres. Pode ser possível apurar casos e características das pessoas que estão compartilhando este tipo de experiência. Além disso, nesse mesmo tópico palavras como “ansiedade”, “síndrome do pânico”, “distúrbios” e “anorexia” revelam o quanto esses problemas estão presentes na nossa sociedade e que ações complementares são necessárias para melhorar a saúde mental da população no Brasil.

O Tópico 8, rotulado como “Medicamentos e Medicina Alternativa”, levantou praticamente todos os medicamentos compartilhados nas mensagens. O interessante nesse tópico foi o quão eficaz ele foi em agrupar todos os medicamentos em um único tópico sendo possível ter uma ideia geral dos medicamentos mais frequentes e relevantes discutidos nas mensagens. Nele é possível encontrar basicamente remédios para dores em geral (cabeça, estômago, muscular, fazendo uma conexão fiel com os tópicos da categoria “Dores no Corpo”), gripe, alergias, antibióticos, remédios para cólica menstrual e doenças neurológicas, que também podem ser ligados com as doenças relevadas nos outros tópicos. Práticas de abuso de substâncias são comuns, porém difíceis de obter estatísticas quantitativas em curto prazo, as plataformas tradicionais que coletam e geram esses índices estão atrasadas no tempo e em termos de disponibilidade de dados ao público, por exemplo para possível uso em pesquisas. Esses índices estatísticos podem ser complementados com os dados das redes sociais.

O Tópico 10, rotulado como “Câncer e Doenças relacionadas”, foi um tópico coerente sobre diversos tipos de câncer e doenças relacionadas como diabetes, hipertensão, infarto e AVC. Além disso, pode-se notar palavras relacionadas a campanhas de prevenção e combate, e também tratamento dessas enfermidades.

Todo ano campanhas são realizadas e compartilhadas nas redes sociais com o objetivo de atingir grande parte da população para conscientizar sobre as enfermidades, prevenção e tratamento. A doença câncer parece ser a que mais se destaca entre as doenças que possuem campanha de saúde pública online.

Os tópicos da categoria “Saúde da Mulher” definem praticamente todos os sentimentos e sintomas que as mulheres sofrem durante o ciclo menstrual. Essas palavras identificadas nos tópicos podem auxiliar os médicos especialistas a criarem perfis atualizados dos principais sentimentos e sintomas descritos pelas mulheres durante o ciclo menstrual. Além disso, nessa categoria também foram vinculados tópicos relacionados a “infecção urinária” que é uma doença bastante comum em mulheres jovens. Segundo o Blog da Saúde, do Ministério da Saúde, a mulher tem 50 vezes mais chance de ter esse problema do que o homem e aproximadamente 30% das mulheres vão apresentar infecção urinária leve ou grave na vida[65].

O Tópico 16 apresentou como palavra mais relevante “pneumonia”, as outras palavras do tópico parecem estar relacionadas a crenças de como adquirir pneumonia, como por exemplo, “sair de cabelo molhado” ou “pegar friagem”. Essas “causas” podem aumentar o risco de doenças alérgicas e gripe, porém não possui relação com a pneumonia. Essa crença poderia receber uma atenção pelas instituições públicas de saúde compartilhando informações que possam trazer diversos esclarecimentos sobre essa doença, inclusive que existe vacina para prevenção da mesma.

Os tópicos da categoria “Alimentos” não foram considerados de saúde pois não apresentaram nenhuma palavra relacionada à saúde (doença, medicamento ou sintoma), porém, pode ser que esses alimentos citados nesses tópicos tenham alguma conexão com os assuntos de saúde como alimentos que auxiliam para o fortalecimento do sistema imunológico, já que as palavras relevantes no tópico estavam relacionadas a frutas e vitamina.

5.5 Limitações da pesquisa

Este estudo apresentou algumas limitações relacionadas tanto a infraestrutura para armazenar e tratar grandes volumes de dados quanto a seleção e análise dos dados.

Para lidar com um grande volume de dados é necessária uma infraestrutura

apropriada, segura, escalável e reversível. O banco de dados escolhido para armazenar os dados desta pesquisa foi o MongoDB que é um banco de alta performance e adequado para armazenar uma enorme quantidade de dados, porém é necessário um servidor apropriado com as configurações devidas para que o banco funcione da melhor forma e seja possível usufruir todo o seu poder de processamento. Durante todo o período desta pesquisa (2015 a 2019) foram coletadas e armazenadas mais de 500 milhões de mensagens do Twitter, no entanto devido a algumas limitações encontradas no ambiente preparado, o estudo acabou sendo realizado com apenas 141 milhões de mensagens.

Outra limitação está relacionada aos dados. Não foi possível utilizar geolocalização pois poucas mensagens do corpus de saúde apresentavam latitude e longitude, o que implica que poucos usuários publicavam mensagens de dispositivos móveis com marcação geográfica ativada. Além disso, a seleção de mensagens sobre saúde utilizando as listas de saúde por apresentarem um vocabulário mais técnico, o que não caracteriza o que geralmente é utilizado nas redes sociais, pode ter deixado de fora mensagens relevantes sobre saúde que não mencionam as palavras pré-definidas pelas listas.

Existem limitações inerentes ao uso do Twitter e de outras redes sociais para realização de análises de conteúdos sobre saúde. Muitas pessoas não utilizam as redes sociais para compartilharem publicamente seu estado de saúde, portanto o Twitter não é uma amostra representativa da população. No entanto, uma variedade de informações de saúde pode ser detectada, apesar dessas limitações. Ainda assim, acredita-se que as fontes de redes sociais podem complementar as ferramentas de vigilância existentes, com algumas vantagens exclusivas, como acesso quase em tempo real as informações.

6 Conclusão

Nos últimos anos, as redes sociais se tornaram uma plataforma onde os consumidores de saúde estão criando e consumindo conteúdo ativamente. Essas plataformas se tornaram uma base de informações públicas e atuais que contém grande parte das atividades, pensamentos e sentimentos dos usuários. À medida que a Web amadurece, aplicativos personalizados se aproximam dos usuários por serem adaptados às suas necessidades e interesses em saúde. A criação dessas aplicações depende de informações extraídas sobre os usuários e uma das fontes dessas informações são as próprias redes sociais.

A informação eletrônica textual pode ser tratada com auxílio de máquinas graças a área de Descoberta de Conhecimento em Texto que propões várias soluções para isso, com a intenção de diminuir o impacto da sobrecarga de informação.

Saber extrair, manipular e analisar esses dados pode contribuir com vários serviços necessários a toda sociedade, e não apenas para personalização do conteúdo por indivíduo. A participação dos cidadãos expondo suas preocupações e experiências sobre saúde nessas plataformas pode auxiliar no monitoramento de aspectos da saúde pública colaborando para uma vigilância participativa, como por exemplo quais são as doenças sazonais mais discutidas pelos usuários no Twitter, quais os sintomas que possuem conexão com essas doenças etc.

Nesta tese foi apresentada uma investigação de técnicas de mineração de dados e modelagem de tópicos para extrair uma variedade de dados sobre saúde, contextualizá-los e oferecer informações analíticas provenientes de mensagens do Twitter em português brasileiro.

Este estudo por apresentar um objetivo amplo e não profundo, se difere da maioria dos estudos de análise de informações de saúde pública baseados nas redes sociais por não focar em um problema de saúde específico. O propósito era exploratório e foi possível identificar diversos assuntos de saúde com as técnicas utilizadas.

A observação e análise contínua de tendência e disseminação de conteúdos sobre doenças, medicamentos e sintomas podem auxiliar no planejamento de campanhas de saúde pública e indicativos para vigilância. É de interesse das organizações de saúde ter o domínio de técnicas de extração de conhecimento para poder gerir o mesmo e encontrar fenômenos que possam ser monitorados,

examinados e compartilhados ao longo do tempo.

Desta forma, a importância deste estudo se baseia em apresentar a aplicação de técnicas de mineração de dados e modelagem de tópicos na busca de conhecimento “escondido” que pode auxiliar os profissionais e pesquisadores a compreender melhor o conteúdo textual, além de fornecer informações úteis para uma posterior análise aprofundada. Acredita-se que uma abordagem exploratória e orientada a descobertas pode servir como um ponto de partida útil para a mineração de dados saúde.

6.1 Principais contribuições

A metodologia utilizada apresentou as melhores práticas para amostragem e análise de dados textuais sobre saúde extraídos de redes sociais, mostrando a importância de cada técnica para compreender todas as descobertas. Todavia essa metodologia não é restrita apenas ao assunto saúde, pode também ser aplicada para outros assuntos.

Com relação a análise de redes sociais, utilizando redes complexas no qual as palavras eram representadas pelos nós e as arestas representadas pelas relações de adjacência, foi possível explorar a topologia da rede gerada pelas mensagens do Twitter. Através dos cálculos das medidas de centralidade vários aspectos relevantes com relação aos termos de maior grau, os termos que eram pontes para conexão com outros termos e relações de significância foram revelados. Padrões interessantes foram observados. Todavia, outras abordagens ainda podem ser exploradas.

Os resultados forneceram evidências para confirmar os impactos positivos de iniciativas e eventos de conscientização que foram promovidos por organizações e profissionais de saúde na rede social Twitter. Mostrando o potencial da utilização de técnicas de mineração de dados para identificação da repercussão de ações de promoção e recuperação da saúde.

Foi possível identificar doenças, sintomas e medicamentos mais discutidos no Twitter em um determinado período por meio das técnicas aplicadas e assegurar que informações de saúde são compartilhadas nessa plataforma. Relacionamentos entre os três conceitos de saúde também foram identificados e analisados com base na força de suas conexões. Os interesses e preocupações da população sobre temas variados, como por exemplo “Febre Amarela”, também foram identificados podendo de alguma forma auxiliar gestores de saúde a promoverem ações oportunas que

aumentem a eficiência de resposta do sistema para com a população.

No que diz respeito a identificação de tópicos no corpus, a modelagem de tópicos se mostrou útil pela capacidade de analisar grandes quantidades de dados de texto não estruturados e expor os principais tópicos relacionado no conjunto de dados. Encontrando tópicos que não foram identificados facilmente pelas outras abordagens, mostrando, portanto, que as diferentes abordagens utilizadas se complementaram na descoberta de conhecimento nos textos analisados. Também foi possível identificar diversos interesses e preocupações da população sobre assuntos não tão populares de saúde.

Esta pesquisa pode ainda auxiliar na criação de uma plataforma de colaboração em pesquisa para saúde pública e seus resultados podem contribuir na formação do conhecimento sobre como as redes sociais podem assistir estudos e análises de saúde pública.

6.2 Trabalhos Futuros

A partir dos resultados obtidos, notamos que determinados pontos merecem atenção especial. Como proposta de futuros trabalhos algumas ideias podem ser consideradas. Dentre diversas possibilidades de trabalhos futuros, temos:

- Desenvolver um classificador para identificar com precisão mensagens sobre saúde e não relacionadas a saúde em português brasileiro;
- Construir uma interface (estática e dinâmica) para exportação dos dados de relacionamentos encontrados no corpus quanto aos medicamentos, doenças e sintomas;
- Descrever o relacionamento de mensagens sobre doenças, medicamentos, sintomas e os tópicos por geolocalização e por período de tempo;
- Utilizar técnicas de reconhecimento de entidades nomeadas para identificar os termos relacionados a pessoas, organizações, cidades e etc. E técnicas de identificação de classes gramaticais conhecida como “Part-of-Speech (POS) tagging” que identificam as classes gramaticais das palavras baseando-se na sintaxe e morfologia. Isso poderia possibilitar análises detalhadas nos grafos de palavras;
- Construir um cenário no qual as análises possam ser realizadas com dados em tempo real, fornecendo informações dos tópicos de saúde e relacionamentos

entre as palavras praticamente no mesmo momento em que se é discutido nas redes sociais.

Referências

1. Newman R, Chang V, Walters RJ, Wills GB. Web 2.0 – The past and the future. International Journal of Information Management [periódicos na Internet]. 2016 [acesso em 04 jun 2018];36(4):591-598
2. Araujo GD, Teixeira FO, Mancine F, Guimarães MP, Pisa IT. Sentiment Analysis of Twitter's Health Messages in Brazilian Portuguese. J Health Inform [periódicos na Internet]. 2018 [acesso em 04 jun 2018];10(1):17-24. Disponível em http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/566/326
3. Paul MJ, Dredze M. You Are What You Tweet: Analyzing Twitter for Public Health. International AAAI Conference on Web and Social Media [Internet]. 2011 Jul 5; EUA. [acesso em 04 jun 2018]. Disponível em https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880
4. Chew C, Eysenbach G. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS ONE [periódicos na Internet] 2010 [acesso em 04 jun 2018]; 5: e14118. Disponível em https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0014118
5. Araujo, GD, Sousa FS, Teixeira FO, Mancini F, De Domenico, EBL, Guimarães MP, Pisa IT. Análise de sentimentos sobre temas de saúde em mídia social. J Health Inform [periódicos na Internet] 2012 [acesso em 04 jun 2018]; 4(3). Disponível em http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/195/164
6. Carvalho, G. A saúde pública no Brasil. Estud. av. [periódicos na Internet] 2013 [acesso em 04 jun 2018]27(78):7-26. Disponível em http://dx.doi.org/10.1590/S0103-40142013000200002
7. Oliveira CM, Cruz MM. Sistema de Vigilância em Saúde no Brasil: avanços e desafios. Saúde em debate [periódicos na Internet]. 2015 [acesso em 04 jun 2018];39(104):255-267. Disponível em http://dx.doi.org/10.1590/0103-110420151040385
8. Blog da Saúde – Ministério da Saúde. Aplicativo vai ajudar na vigilância participativa e prover informações de saúde para usuários. [Internet] 2016 [acesso em 04 jun 2018]. Disponível em http://www.blog.saude.gov.br/index.php?option=com_content&view=article&id=50792&catid=566&Itemid=50155

9. Blei, DM. Probabilistic topic models. J Commun. ACM [periódicos na Internet]. 2012 [acesso em 04 jun 2018]; 55(4);77-84. Disponível em https://doi.org/10.1145/2133806.2133826
10. Dorow B. A Graph Model for Words and their Meanings. Stuttgart. Tese [Doutorado] – Institut fur Maschinelle Sprachverarbeitung der Universitat; 2006.
11. Teixeira MG et al. Vigilância em Saúde no SUS - construção, efeitos e perspectivas. Ciênc. saúde colet [periódicos na Internet]. 2018 [acesso em 1 jul 2018];23(6). Disponível em https://doi.org/10.1590/1413-81232018236.09032018
12. Pense SUS [homepage na internet]. Vigilância em saúde. [acesso em 04 jun 2018]. Disponível em: https://pensesus.fiocruz.br/vigilancia-em-saude
13. Sanders-Jackson A, Brown CG, Prochaska JJ. Applying linguistic methods to understanding smoking-related conversations on Twitter. Tobacco Control [periódicos na Internet]. 2015 [acesso em 4 jun 2018];24:136-138. Disponível em https://tobaccocontrol.bmj.com/content/24/2/136.
14. Paul MJ, Dredze M. Discovering Health Topics in Social Media Using Topic Models. PLos ONE [periódicos na Internet]. 2014 [acesso em 04 jun 2018];9(8):e103408. Disponível em https://doi.org/10.1371/journal.pone.0103408
15. Santos JC, Matos S. Analysing Twitter and web queries for flu trend prediction. Theor Biol Med Model [periódicos na Internet]. 2014 [acesso em 04 jun 2018]11, S6. Disponível em https://doi.org/10.1186/1742-4682-11-S1-S6
16. Beykikhoshk A, Arandjelović O, Phung D et al. Using Twitter to Learn about the Autism Community. Soc. Netw. Anal. Min. [periódicos na Internet]. 2015 [acesso em 04 jun 2018];5: 22. Disponível em https://doi.org/10.1007/s13278-015-0261-5
17. Ji X, Chun SA, Wei Z, Geller J. Twitter sentiment classification for measuring public health concerns. Soc. Netw. Anal. Min. [periódicos na Internet]. 2015 [acesso em 04 jun 2018];5: 13. Disponível em https://doi.org/10.1007/s13278-015-0253-5
18. Tangherlini RT et al. “Mommy Blogs” and the Vaccination Exemption Narrative: Results From A Machine-Learning Approach for Story Aggregation on Parenting Social Media Sites. JMIR Public Health Surveill [periódicos na Internet]. 2016 [acesso em 04 de jun 2018];2(2):e166. Disponível em https://doi.org/10.2196/publichealth.6586
19. Chen L, Tozammel H, K.S.M. Butler P, Ramakrishnan N, Prakash BA. Syndromic Surveillance of Flu on Twitter Using Weakly Supervised Temporal Topic Models.

Data Min Knowl Disc [periódicos na Internet]. 2016 [acesso em 04 jun 2018];30: 681. Disponível em https://doi.org/10.1007/s10618-015-0434-x.

20. Miller WR, Groves D, Knopf A, Otte JL, Silverman RD. Word Adjacency Graph Modeling: Separating Signal From Noise in Big Data. Western Journal of Nursing Research [periódicos na Internet] 2017 [acesso em 4 jun 2018];39(1), 166–185. Disponível em https://doi.org/10.1177/0193945916670363
21. Mackey TK, Kalyanam J. Detection of illicit online sales of fentanyls via Twitter. F1000Research [periódicos na Internet] 2017 [acesso em 4 jun 2018];6:1937. Disponível em https://doi.org/10.12688/f1000research.12914.1
22. Mackey et al. Solution to Detect, Classify, and Report Illicit Online Marketing and Sales of Controlled Substances via Twitter: Using Machine Learning and Web Forensics to Combat Digital Opioid Access. J Med Internet Res. [periódicos na Internet] 2018 [acesso em4 jun 2018];20(4):e10029. Disponível em http://doi.org/10.2196/10029
23. Klein GH, Guidi NP, Tezza R. Big Data e mídias sociais: monitoramento das redes como ferramenta de gestão. Saude soc. [periódicos na Internet]. 2017 [acesso em 4 jun 2018];26(1):208-217. Disponível em http://dx.doi.org/10.1590/s0104-12902017164943
24. Kagashe I, Yan Z, Suheryani I. Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data. J Med Internet Res [periódicos na Internet] 2017 [acesso em 4 jun 2018];19(9):e315. Disponível em https://www.jmir.org/2017/9/e315
25. Stefanidis A et al. Zika in Twitter: Temporal Variations of Locations, Actors, and Concepts. JMIR Public Health Surveill [periódicos na Internet]. 2017 [acesso em 4 jun 2018];3(2):e22). Disponível em 10.2196/publichealth.6925
26. Pruss D et al. Zika discourse in the Americas: A multilingual topic analysis of Twitter. PLoS ONE [periódicos na Internet] 2019 [acesso em 20 nov 2019] 14(5): e0216922. Disponível em https://doi.org/10.1371/journal.pone.0216922
27. Bian et al. Using Social Media Data to Understand the Impact of Promotional Information on Laypeople's Discussions: A Case Study of Lynch Syndrome. J Med Internet Res [periódicos na Internet] 2017 [aceesso em 4 jun 2018];19(12):e414. Disponível em https://doi.org/10.2196/jmir.9266
28. Muralidhara S, Paul MJ. #Healthy Selfies: Exploration of Health Topics on Instagram. JMIR Public Health Surveill [periódicos na Internet] 2018 [acesso em

10 nov 2019];4(2):e10150. Disponível em https://publichealth.jmir.org/2018/2/e10150

29. Grover P, Kar KA, Davies G. “Technology enabled Health”–Insights from twitter analytics with a socio-technical perspective. International Journal of Information Management [periódicos na Internet] 2018 [acesso em 20 nov 2019];43:85–97. Disponível em https://doi.org/10.1016/j.ijinfomgt.2018.07.003
30. Hoffman et al. It’s not all about autism: The emerging landscape of anti-vaccination sentiment on Facebook. Vaccine [periódicos na Internet] 2019 [acesso em 20 non 2019];37 2216-2223. Disponível em https://doi.org/10.1016/j.vaccine.2019.03.003
31. Fayyad U, Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. KDD-96 Proceedings. [periódicos na Internet]1996 [acesso em 4 jun 2018]. Disponível em https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf
32. Silva LA, Peres SM, Boscarioli C. Introdução à Mineração de Dados: Com Aplicações em R. Elsevier; 2016.
33. Shearer C. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing. 2000;5(4)13-22.
34. Loh S, Amaral LA, Wives LK, Oliveira JPM. Descoberta de Conhecimento em Textos através da Análise de Seqüências Temporais. In: Workshop em Algoritmos e Aplicações de Mineração de Dados, WAAMD, II; SBBD, 2006, Florianópolis, ISBN 85-7669-088-8. Florianópolis: Sociedade Brasileira de Computação, 2006. p. 49-56.
35. Appel AP. Métodos para o pré-processamento e mineração de grande volume de dados espaciais e redes complexas. São Carlos. Tese [Doutorado] – Instituto de Ciências Matemáticas e de Computação – USP; 2010.
36. Instituto Brasileiro de Pesquisa e Análise de Dados [homepage na internet]. Iniciação a Redes: Conceitos Essenciais e Principais Ideias. 2015. [acesso em 4 jun 2018]. Disponível em https://www.ibpad.com.br/publicacoes/whitepapers/iniciacao-a-redes-conceitos-essenciais-e-principais-ideias/
37. Ferreira GC. Redes sociais de informação em organizações num contexto da sociedade contemporânea. São Paulo. Dissertação [Mestrado] – Escola de Comunicações e Artes – USP; 2012.

38. Lemieux V, Ouimet M. Análise Estrutural das Redes Sociais. 2. ed. Lisboa: Instituto Piaget; 2014.
39. Borgatti SP, Everett MG. Models of core/periphery structures. Social Networks [periódicos na Internet] 2000 [acesso em 4 jun 2018];21(4)375-395. Disponível em https://doi.org/10.1016/S0378-8733(99)00019-2
40. Amancio DR. Classificação de textos com redes complexas. São Carlos. Tese [Doutorado em Ciências] – Instituto de Física de São Carlos – USP; 2013.
41. Faleiros TP. Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais. São Carlos. Tese [Doutorado] – Instituto de Ciências Matemáticas e de Computação - USP; 2016.
42. Grün B, Hornik K. topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software [periódicos na Internet] 2011 [acesso em 4 jun 2018];40(13)1-30. Disponível em https://www.jstatsoft.org/article/view/v040i13
43. Deerwester et al. Indexing by latent semantic analysis. J. Amer. Soc. Info. Sci. [periódicos na Internet] 1990 [acesso em 4 jun 2018];41(6)391–407. Disponível em https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
44. Hofmann T. Probabilistic Latent Semantic Analysis. Proc. 15th Conf. Uncertainty in AI [periódicos na Internet] 1999 [acesso em 4 jun 2018]289-296. Disponível em https://dl.acm.org/citation.cfm?id=2073829
45. Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review [periódicos na Internet] 1997 [acesso em 4 jun 2018];104(2)211–240. Disponível em https://doi.org/10.1037/0033-295X.104.2.211
46. Graber JB, Blei DM. Multilingual topic models for inaligned text. Proc. 25th Conf. Uncertainty in AI [periódicos na Internet]. 2009 [acesso em 4 jun 2018]. 75-82. Disponível em https://dl.acm.org/citation.cfm?id=1795114.1795124
47. Lakatos E. Fundamentos de metodologia científica. 8th ed. São Paulo: Grupo Gen - Atlas; 2017.
48. Leavitt N. Will NoSQL Databases Live Up to Their Promise?. Computer [periódicos na Internet] 2010 [acesso em 4 jun 2018];43(2)12-14. Disponível em http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5410700&isnumber=5410692

49. Ministério da Saúde - Datasus [homepage na Internet]. Organização Mundial da Saúde. CID-10 - Classificação estatística internacional de doenças e problemas relacionados à saúde. Brasília: Ministério da Saúde - Datasus; 2008 [acesso em 4 jun 2018]. Disponível em http://www.datasus.gov.br/cid10/V2008/cid10.htm
50. Nubila D, Ventura HB, Buchalla CM. O papel das Classificações da OMS - CID e CIF nas definições de deficiência e incapacidade. Rev. bras. epidemiol. [periódicos na Internet]. 2008 [acesso em 4 jun 2018];11(2):324-335. Disponível em http://dx.doi.org/10.1590/S1415-790X2008000200014
51. Alvares RV. Algoritmos de Stemming e o Estudo de Proteomas. Rio de Janeiro. Tese [Doutorado] - Universidade Federal do Rio de Janeiro; 2014.
52. Culotta A. Detectiong influenza outbreaks by analyzing Twitter messages. arXiv [periódicos na Internet] 2010 [acesso eem 4 jun 2018]. Disponível em https://arxiv.org/abs/1007.4748
53. Jimeno-Yepes A et al. Identifying Diseases, Drugs, and Symptoms in Twitter. MEDINFO 2015: eHealth-enabled Health [periódicos na Internet]. 2015 [acesso em 4 jun 2018];216. Disponível em http://ebooks.iospress.nl/publication/40288
54. Ziviani N, Ribeiro-Neto B. Text Operations. In: Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval. Nova York: ACM Press- Addison-Wesley; 1999. p. 163–90.
55. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management [periódicos na Internet] 1988 [acesso em 4 jun 2018];24:513--523. Disponível em https://doi.org/10.1016/0306-4573(88)90021-0
56. Heimerl F, Lohmann S, Lange S, Ertl T. Word Cloud Explorer: Text Analytics Based on Word Clouds. 47th Hawaii International Conference on System Sciences; 2014 Jan 6-9; Waikoloa, HI, EUA: IEEE; 2014. Disponível em https://doi.org/10.1109/HICSS.2014.231
57. Abinaya G, Winstre GS. Event Identification in Social Media through Latent Dirichlet Allocation and Named Entity Recognition. IEEE International Conf. on Computer Communication and Systems; 2014 Feb 20-21; Chennai, India. ICCCS; 2014.
58. Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces [periódicos na Internet] 2014 [acesso em 4 jun 2018];63-70. Disponível em https://www.aclweb.org/anthology/W14-3110/

59. Ramos ML. Fenomenologia da obra literária. Belo Horizonte: Editora UFMG, 2011.
60. Robinson D, Silge J. Text Mining with R: A tidy approach. [livro na Internet]. O'Reilly Media; 2017. [acesso em 4 jun 2018]. Disponível em https://www.tidytextmining.com/index.html
61. Sievert C, Kenneth ES. LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces;2014 Jun 27; Baltimore, Maryland, USA. Association for Computational Linguistics; 2014. Disponível em https://www.aclweb.org/anthology/W14-3110/
62. Chuang J, Ramage D, Manning C, Heer J. Interpretation and trust: designing model-driven visualizations for text analysis. CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems;2012 Mai 5-10;Texas, USA;2012. Disponível em https://dl.acm.org/doi/10.1145/2207676.2207738
63. MeSH [homepage na Internet]. Medical Subject Headings. Bethesda:US National Library of Medicine. 2005 [acesso em 4 jun 2018]. Disponível em www.nlm.nih.gov/mesh
64. Mousa HA. Prevention and Treatment of Influenza, Influenza-Like Illness, and Common Cold by Herbal, Complementary, and Natural Therapies. *J Evid Based Complementary Altern Med*. [periódicos na internet] 2017 [acesso em 10 dez 2019];22(1):166–174. Disponível em https://doi.org/10.1177/2156587216641831
65. Blog da Saúde – Infecção Urinária: Tratamento, sintomas e fatores de risco. [Internet] 2016 [acesso em 10 dez 2019]. Disponível em http://www.blog.saude.gov.br/index.php/promocao-da-saude/51735-infeccao-urinaria-tratamento-e-fatores-de-risco

Anexo A – Aprovação do Comitê de Ética em Pesquisa – UNIFESP – HSP

paraidentificar relacionamentos entre os achados. Os objetivos específicos deste projeto são:
Objetivo 1: Identificar entidades nomeadas (pessoas, locais, organizações etc.) e categorizar temas de saúde associados a doenças, medicamentos e sintomas em mensagens do Twitter;
Objetivo 2: Analisar o sentimento (positivo, negativo ou neutro) de opiniões publicadas em mensagens do Twitter e a repercussão temporal e de localização sobre temas de saúde; Objetivo 3: Propor um modelo de tópico para mensagens
do Twitter sobre temas de saúde usando categorias de saúde, entidades nomeadas, análise de sentimento e de repercussão.

Avaliação dos Riscos e Benefícios:
Conforme parecer n: 1.135.186 de 02 de Julho de 2015

Comentários e Considerações sobre a Pesquisa:
Conforme parecer n: 1.135.186 de 02 de Julho de 2015

Considerações sobre os Termos de apresentação obrigatória:
Conforme parecer n: 1.135.186 de 02 de Julho de 2015

Recomendações:
Conforme parecer n: 1.135.186 de 02 de Julho de 2015

Conclusões ou Pendências e Lista de Inadequações:
As pendências anteriormente apontadas foram atendidas (ver abaixo).Estudo aprovado.

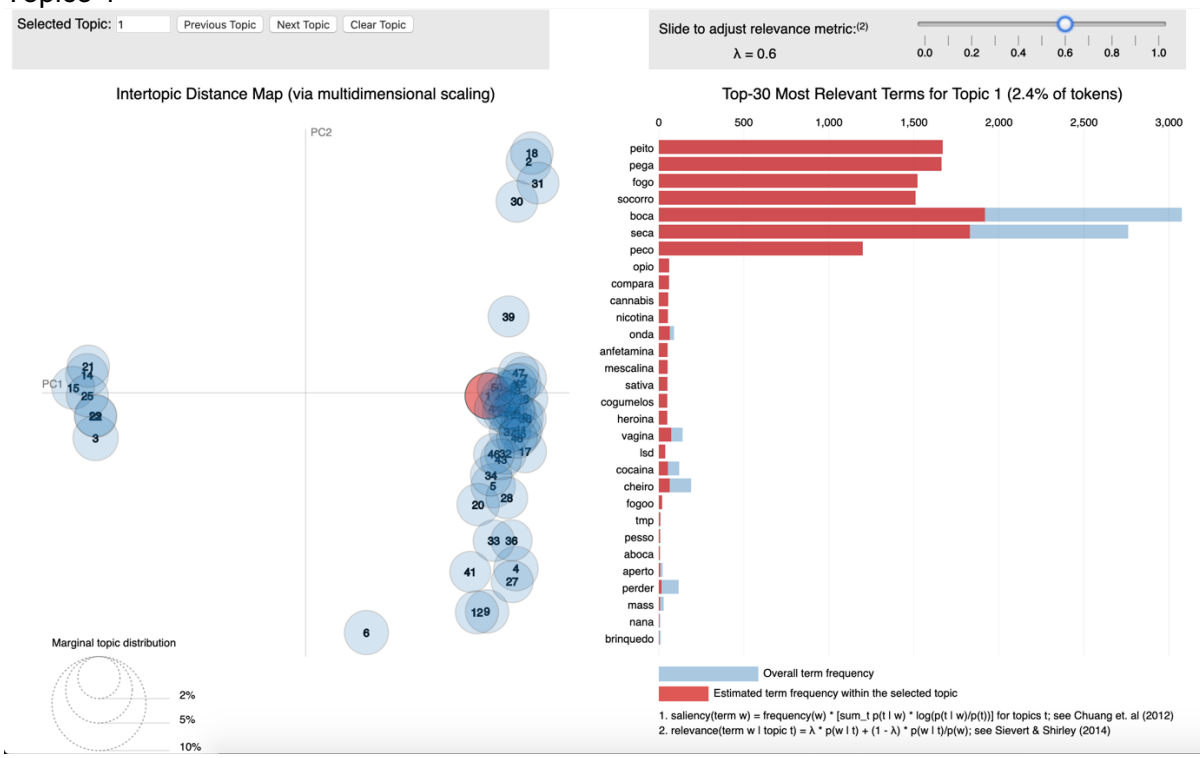
1-Pede-se que a pesquisadora readeque a parte dos Riscos e Benefícios. Um primeiro ponto importante: Risco, nesse caso, diz respeito aos riscos potenciais que existem para os seres humanos que participarão da pesquisa. Eles sempre estão presentes, ainda que sejam mínimos. No documento apresentado, a pesquisadora infere que Risco diz respeito a conflito de interesses, o que não é o caso.
RESPOSTA:as devidas alterações foram realizadas na Plataforma Brasil na seção "Detalhamento do Estudo" na parte de Riscos conforme solicitado no parecer do CEP.
PENDÊNCIA ATENDIDA.

Situação do Parecer:
Aprovado

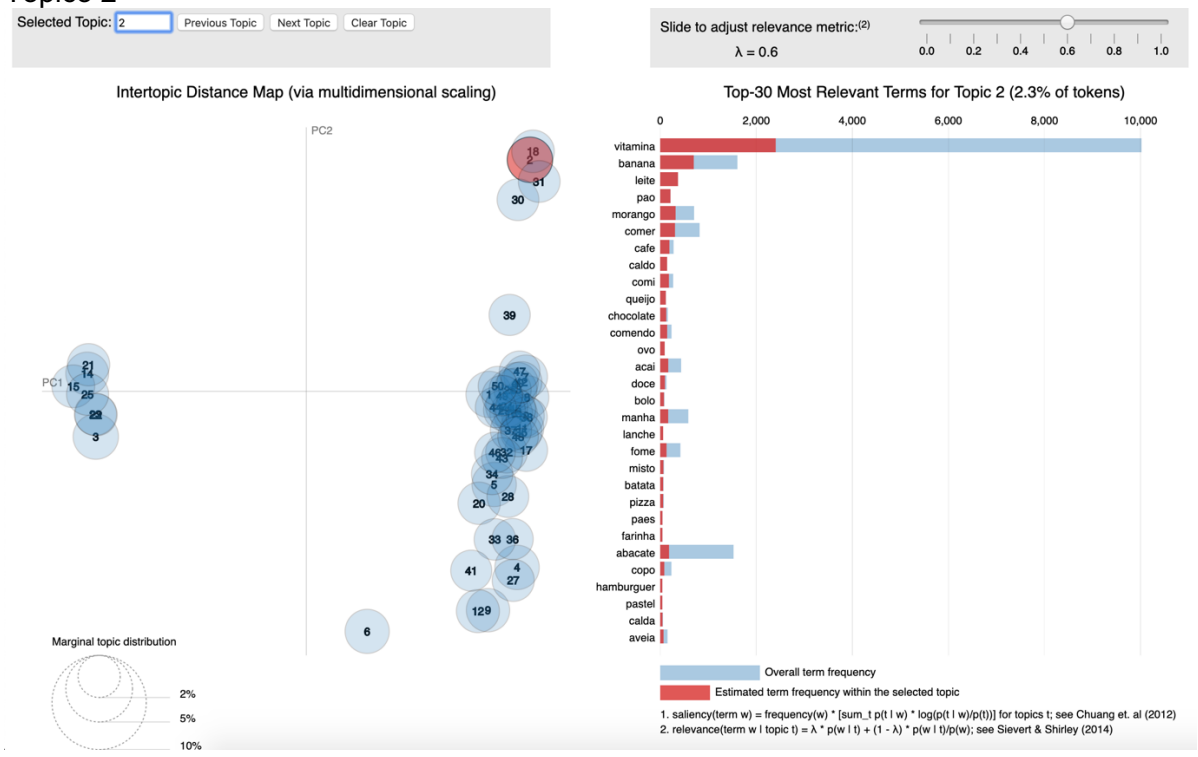
Necessita Apreciação da CONEP:
Não

APÊNDICE A – Ilustração de todos os 50 tópicos gerados

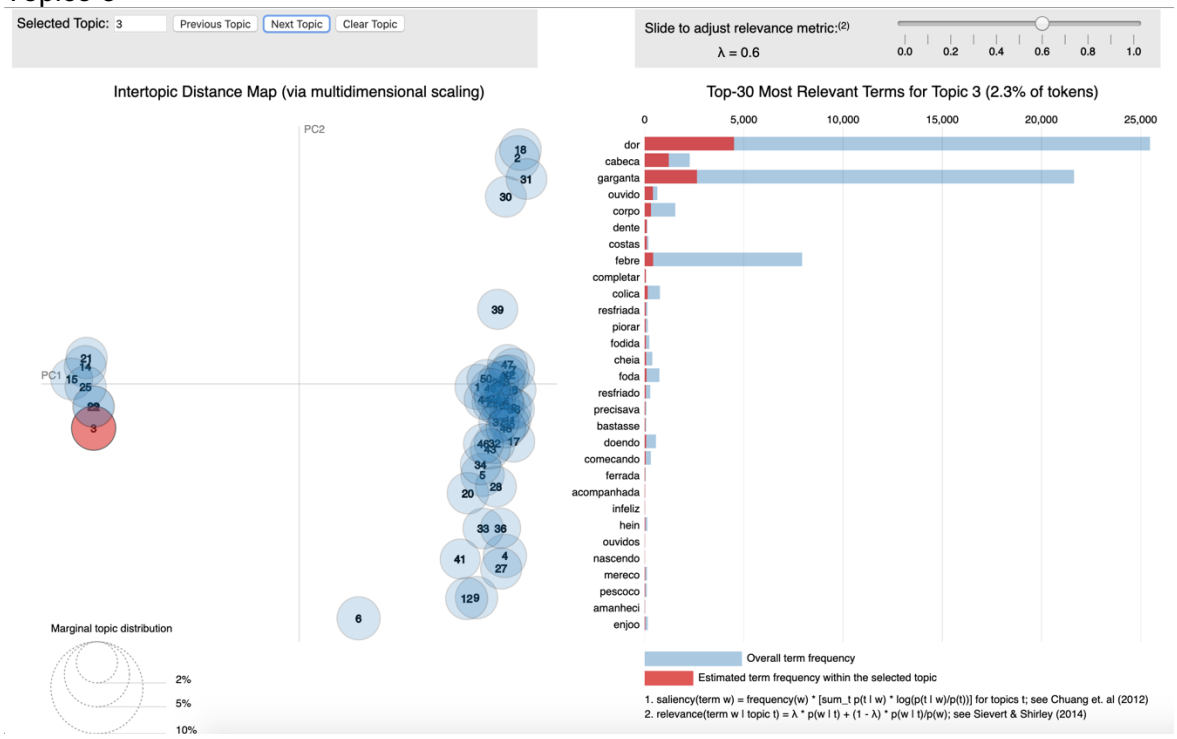
Tópico 1



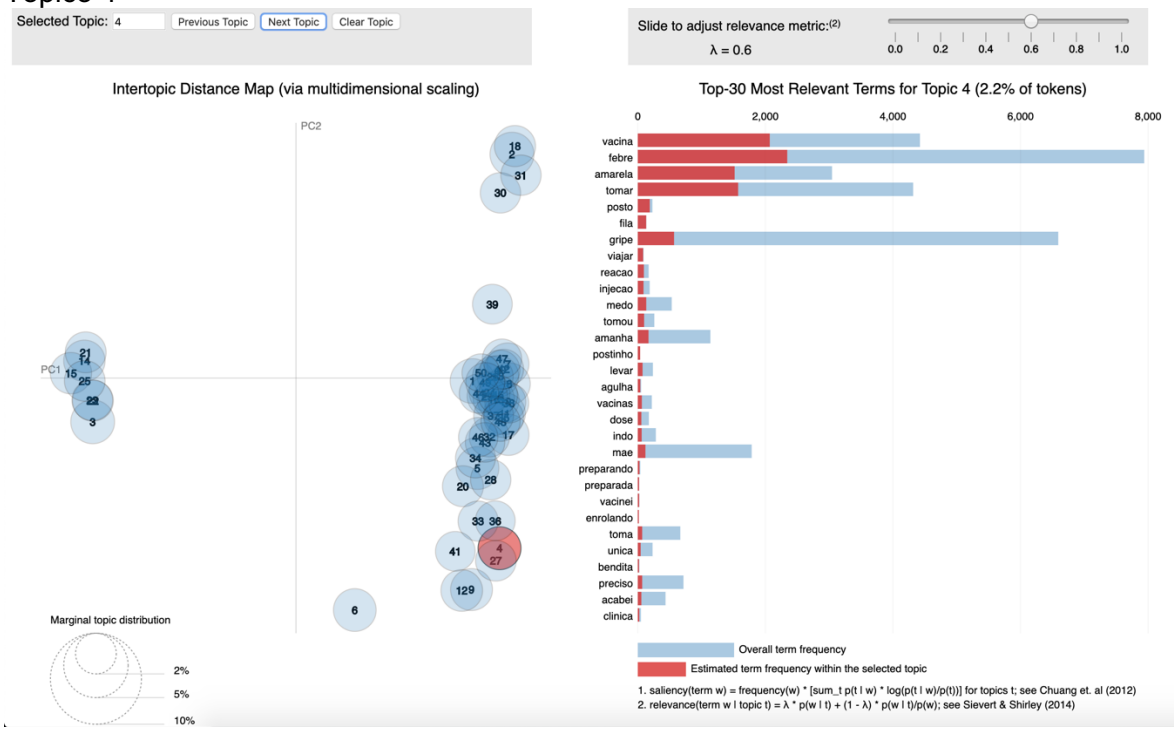
Tópico 2



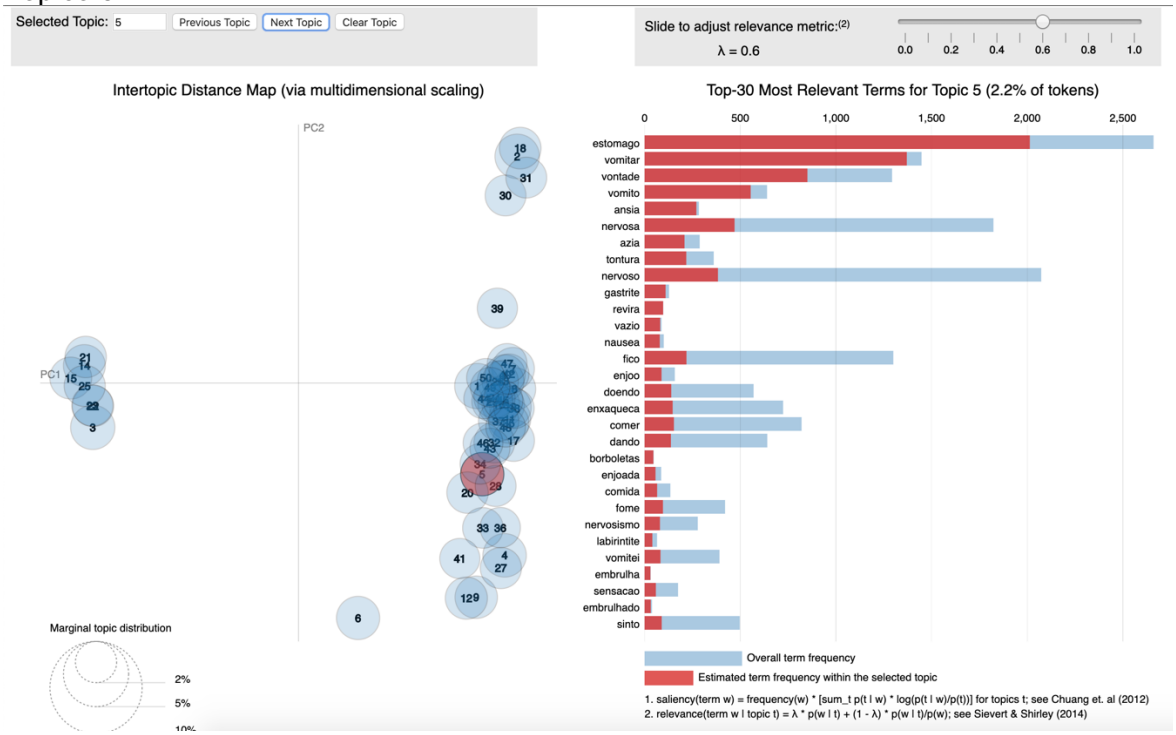
Tópico 3



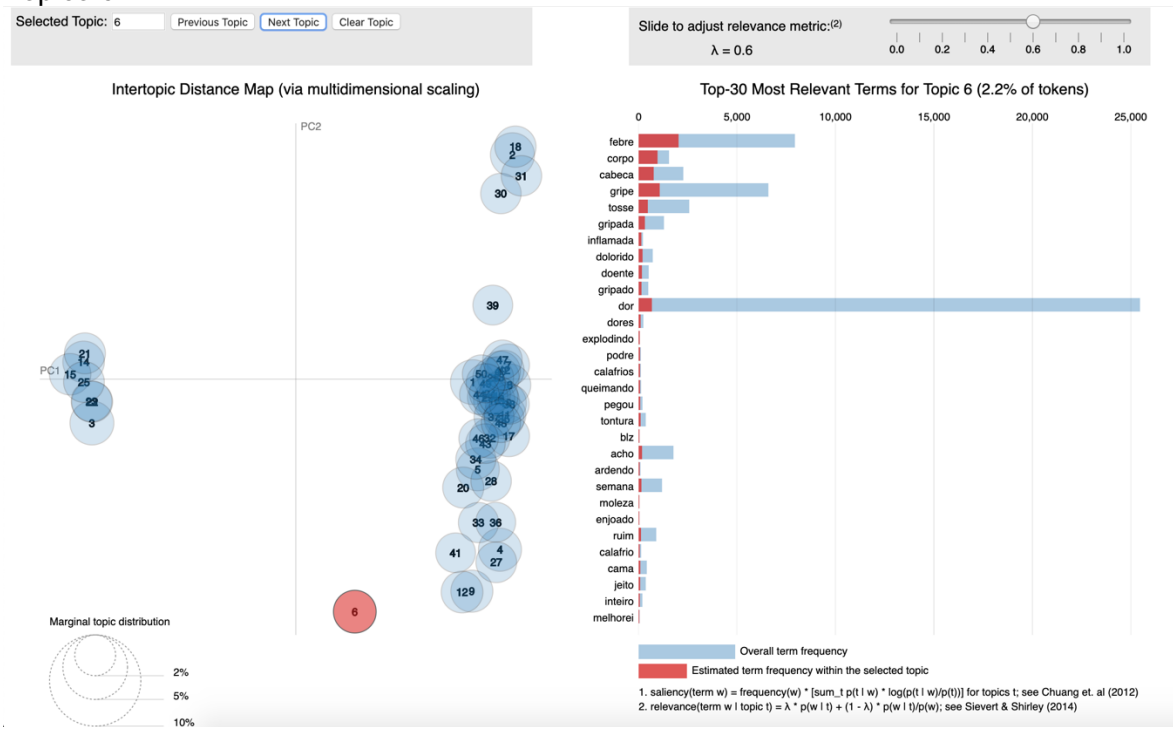
Tópico 4



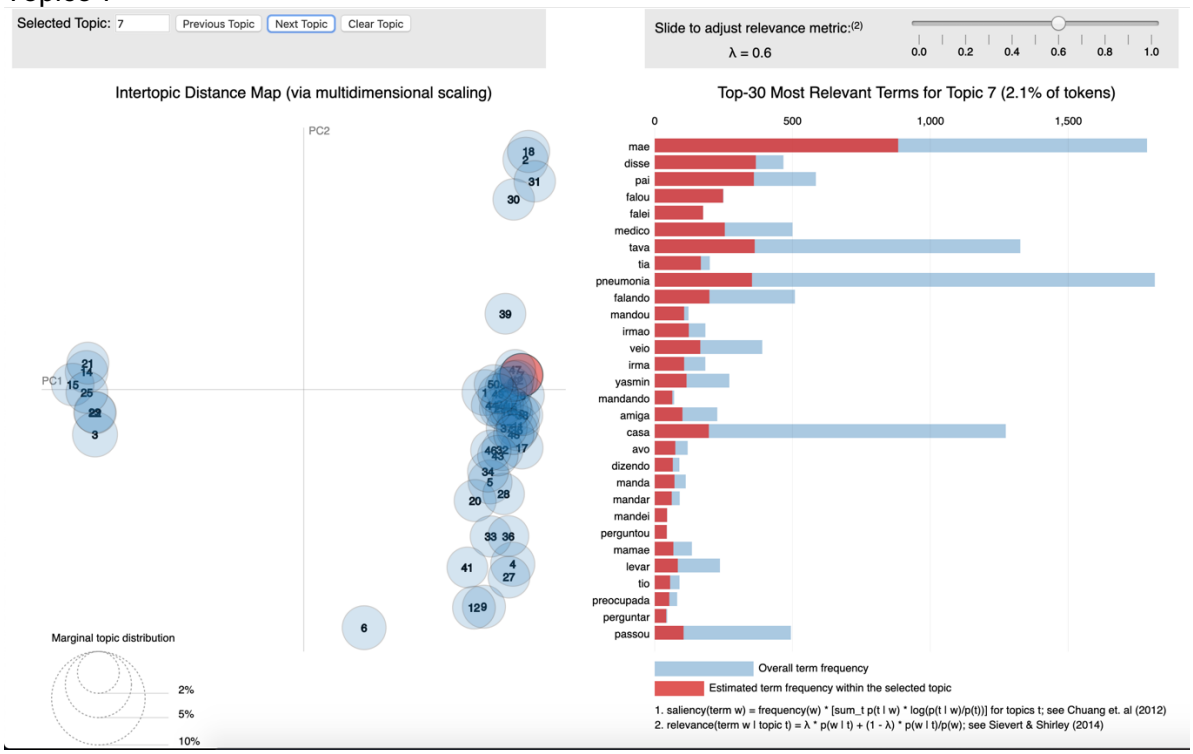
Tópico 5



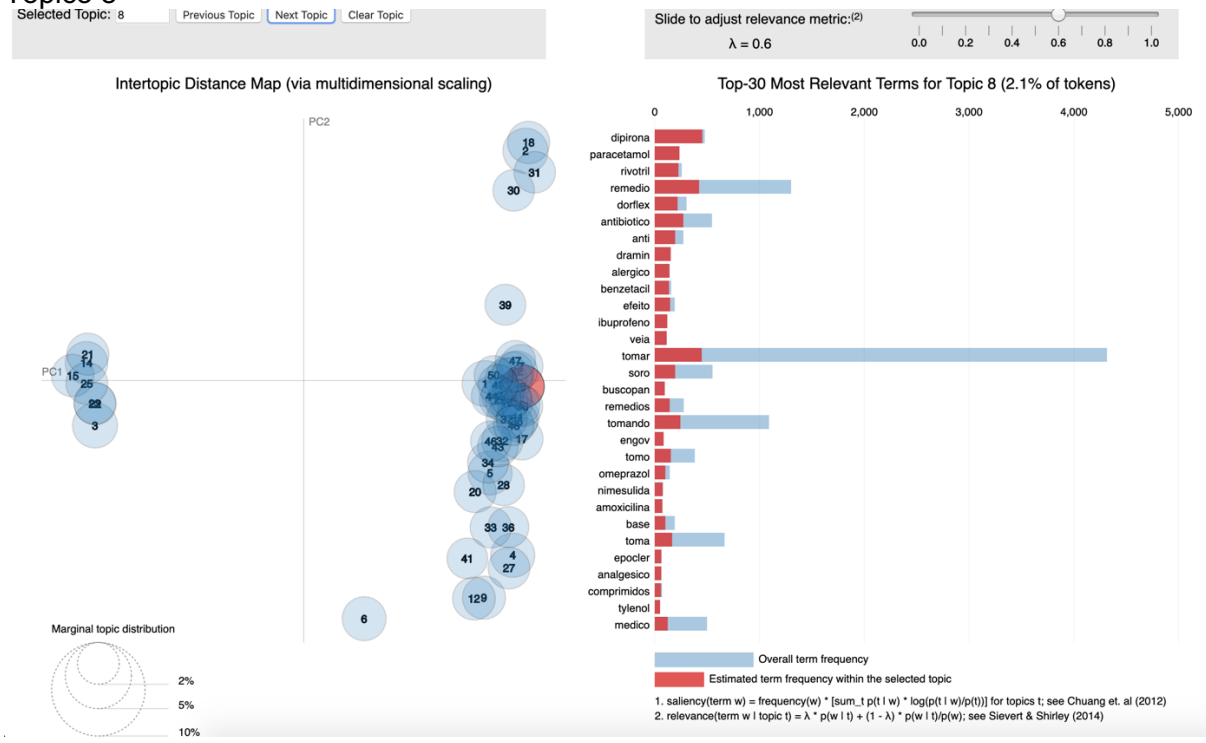
Tópico 6



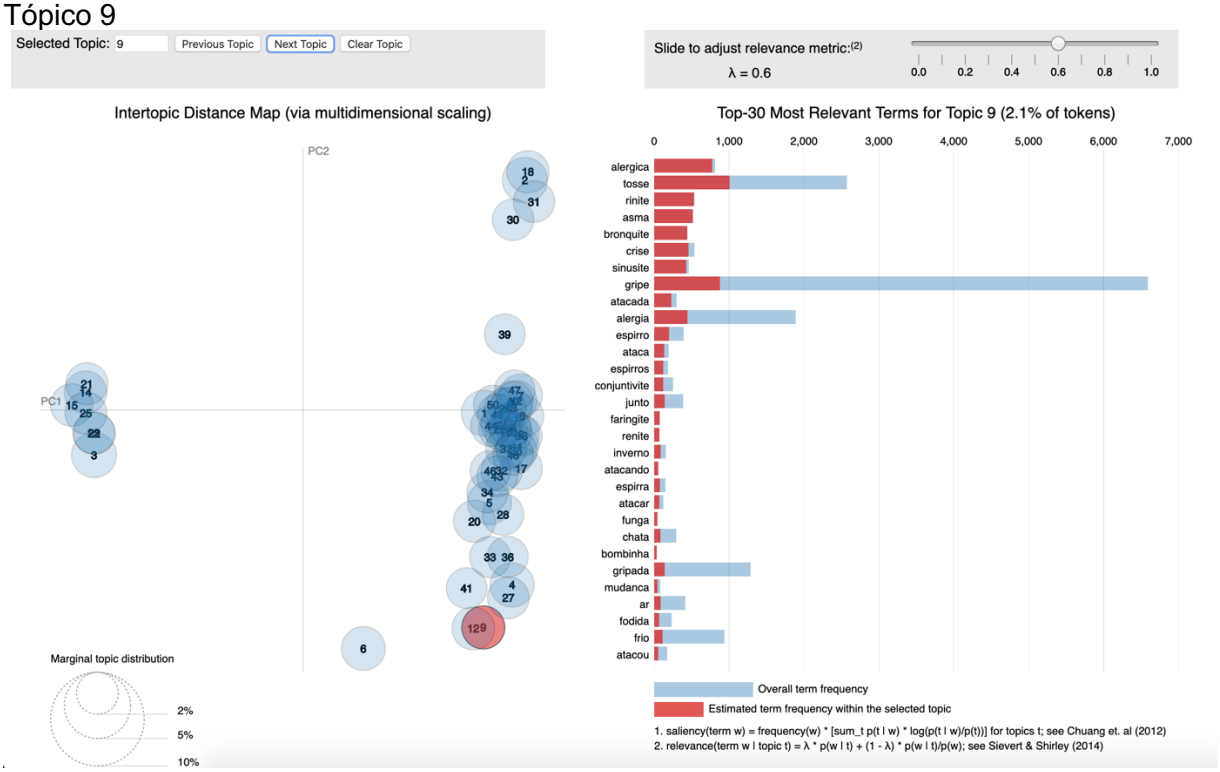
Tópico 7



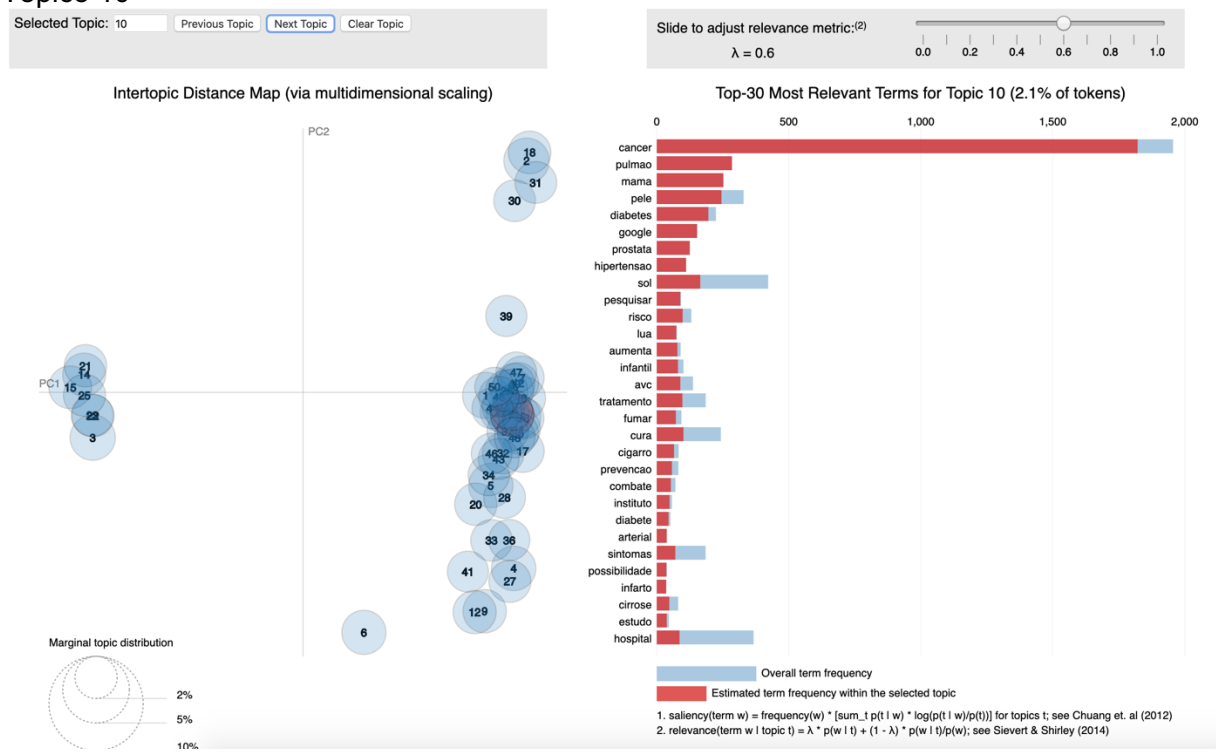
Tópico 8



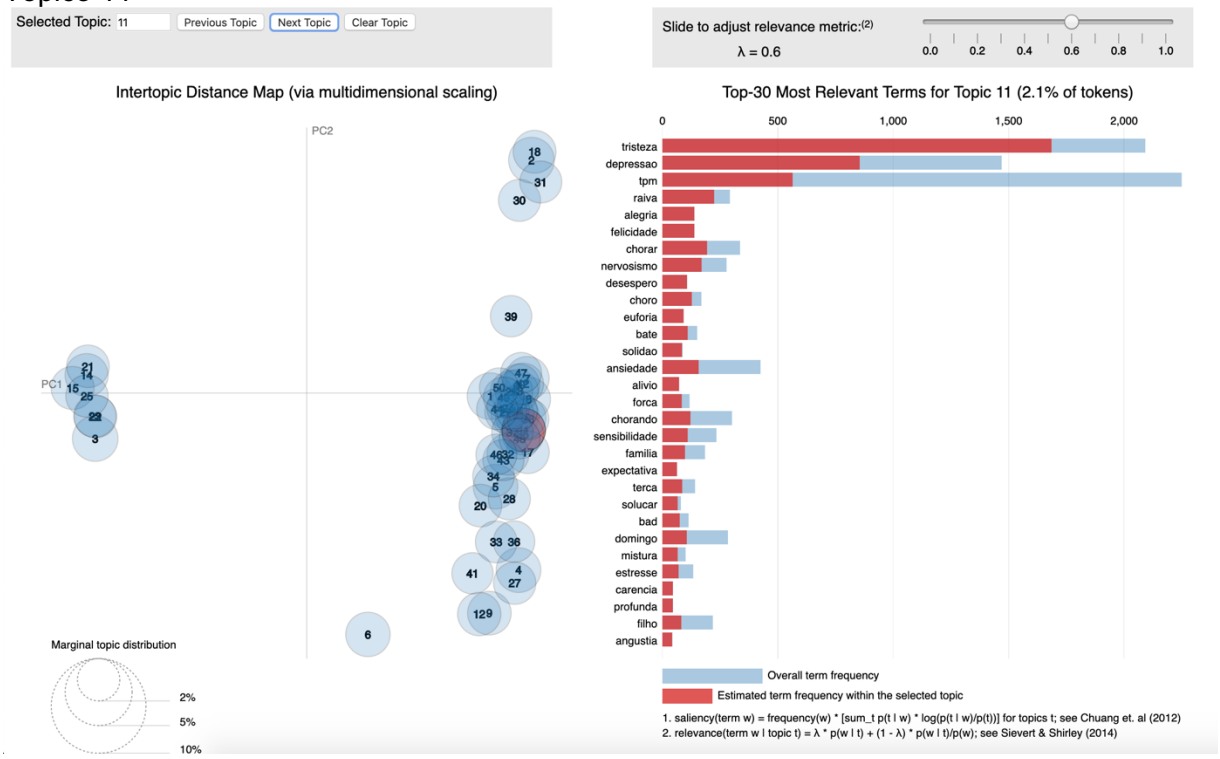
Tópico 9



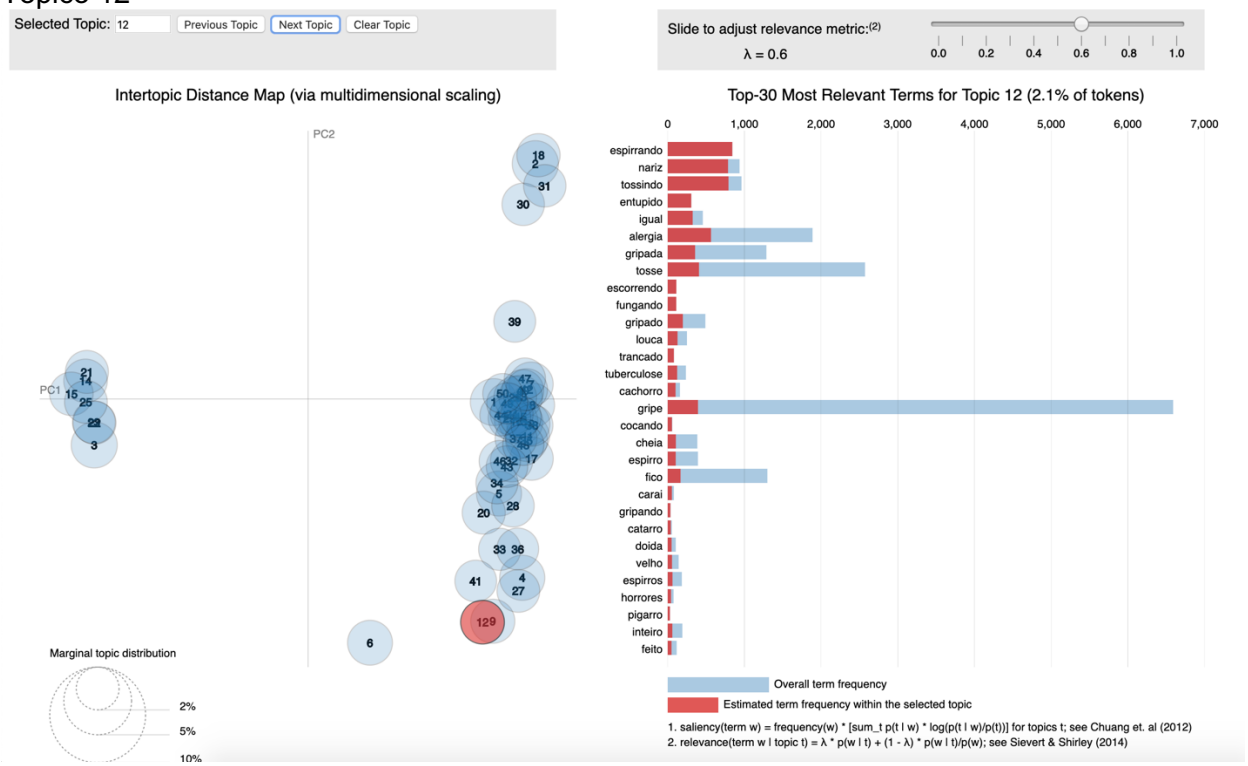
Tópico 10



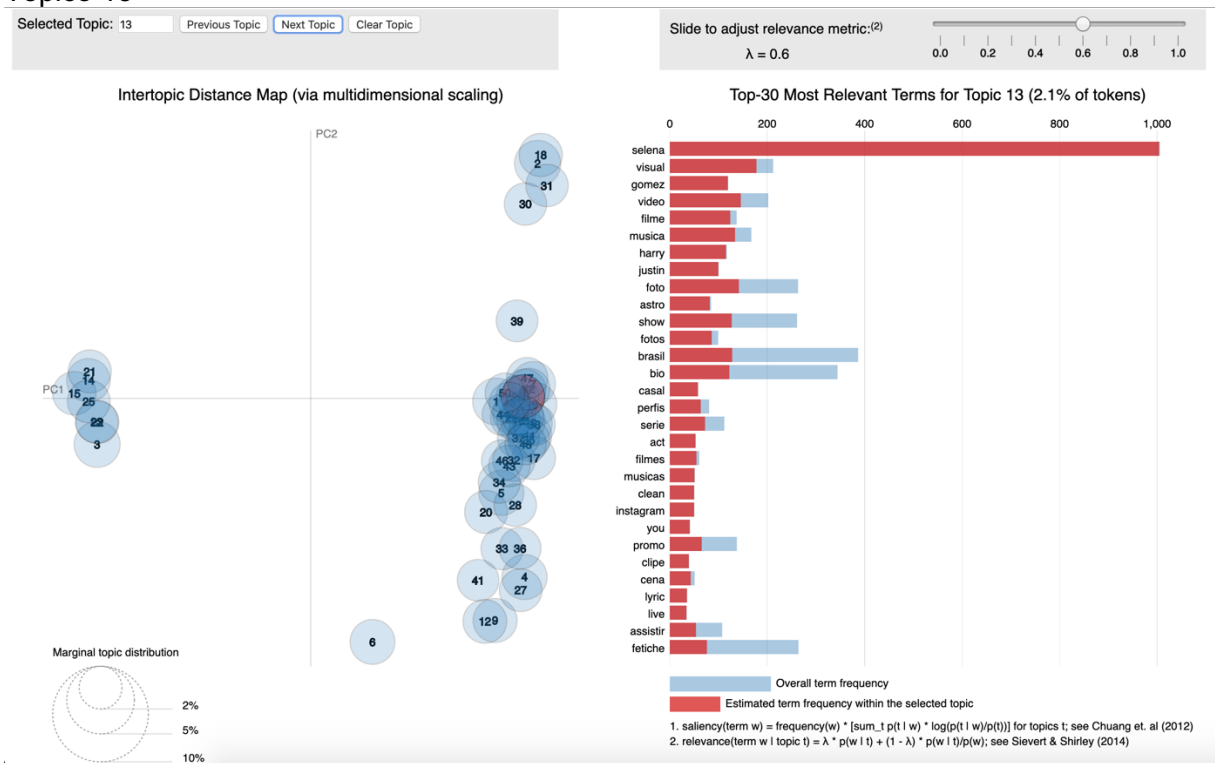
Tópico 11



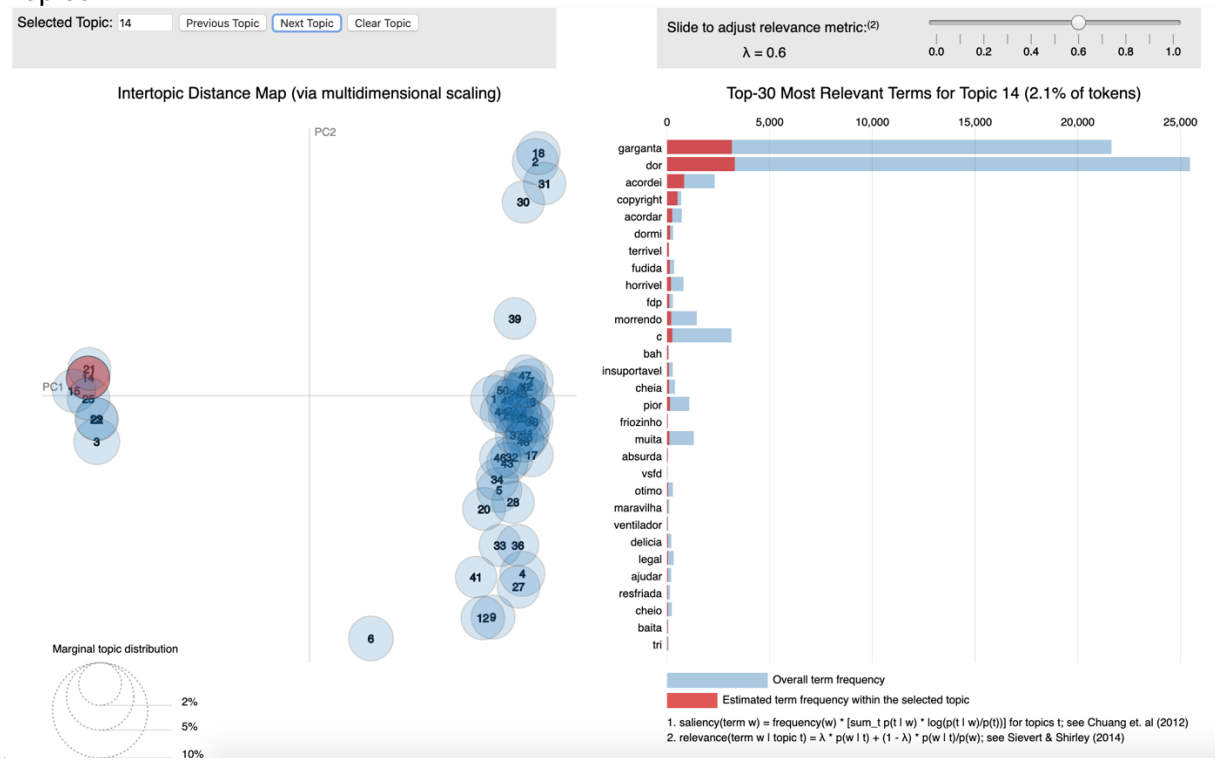
Tópico 12



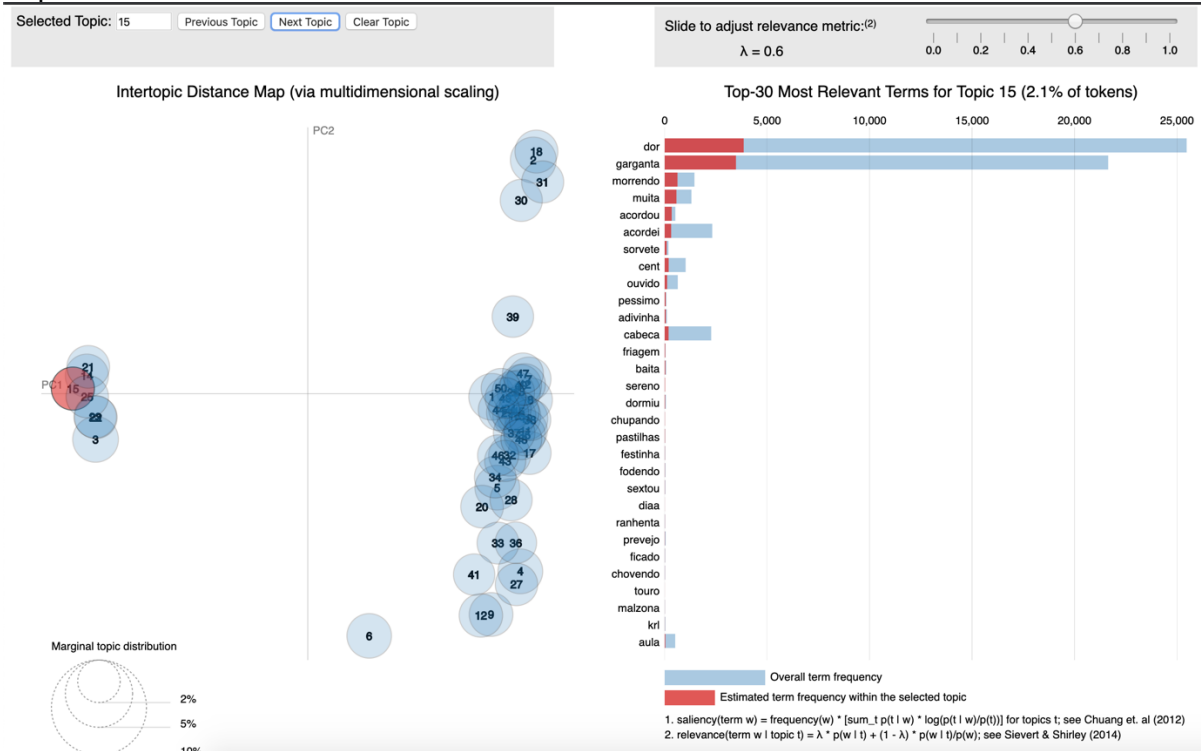
Tópico 13



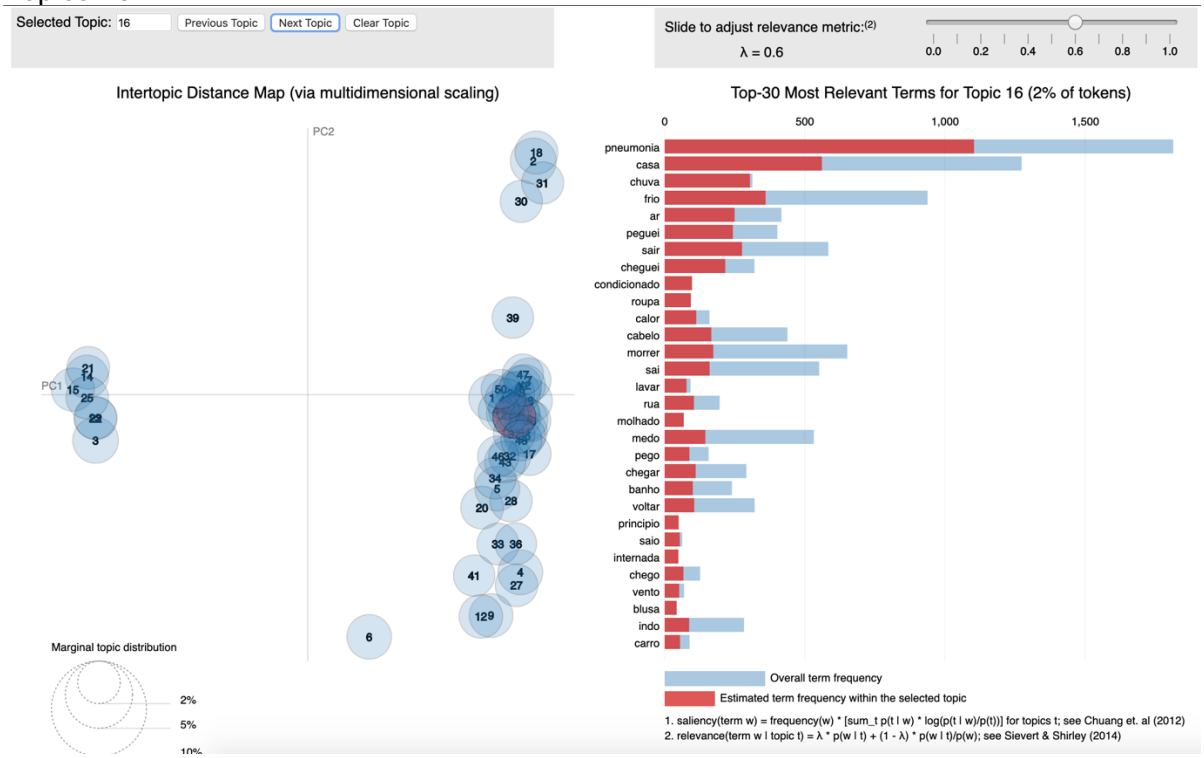
Tópico 14



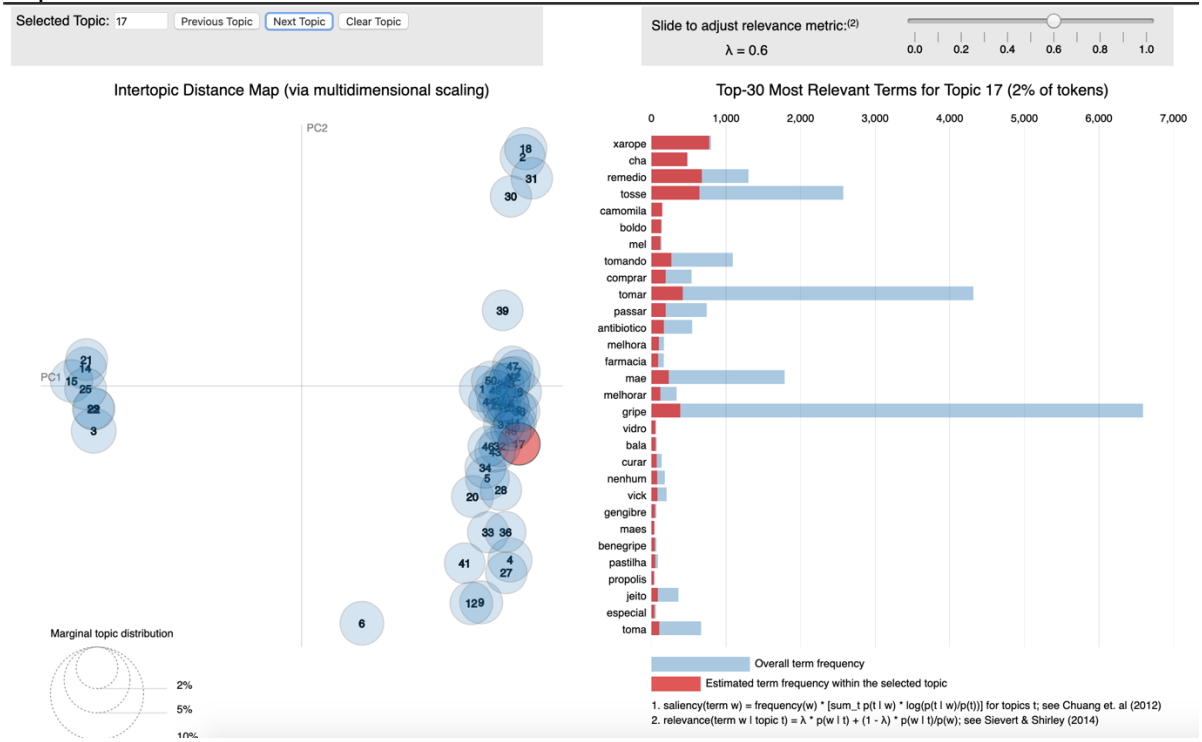
Tópico 15



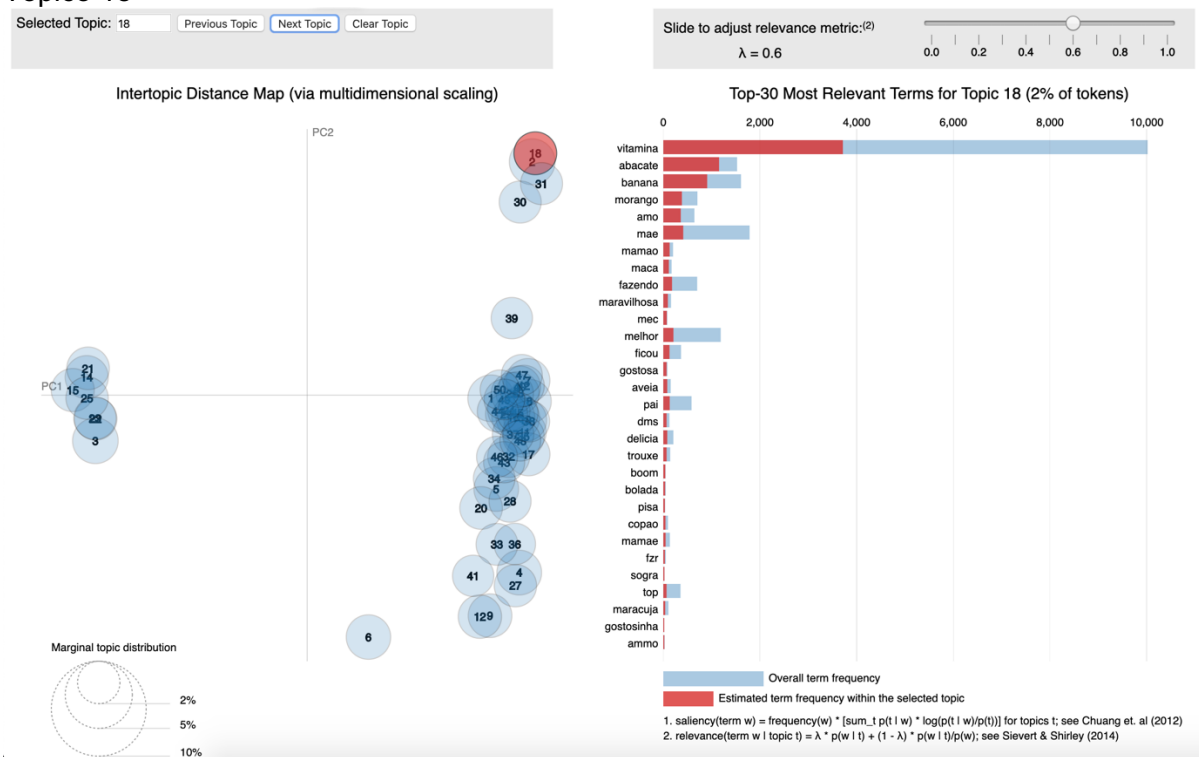
Tópico 16



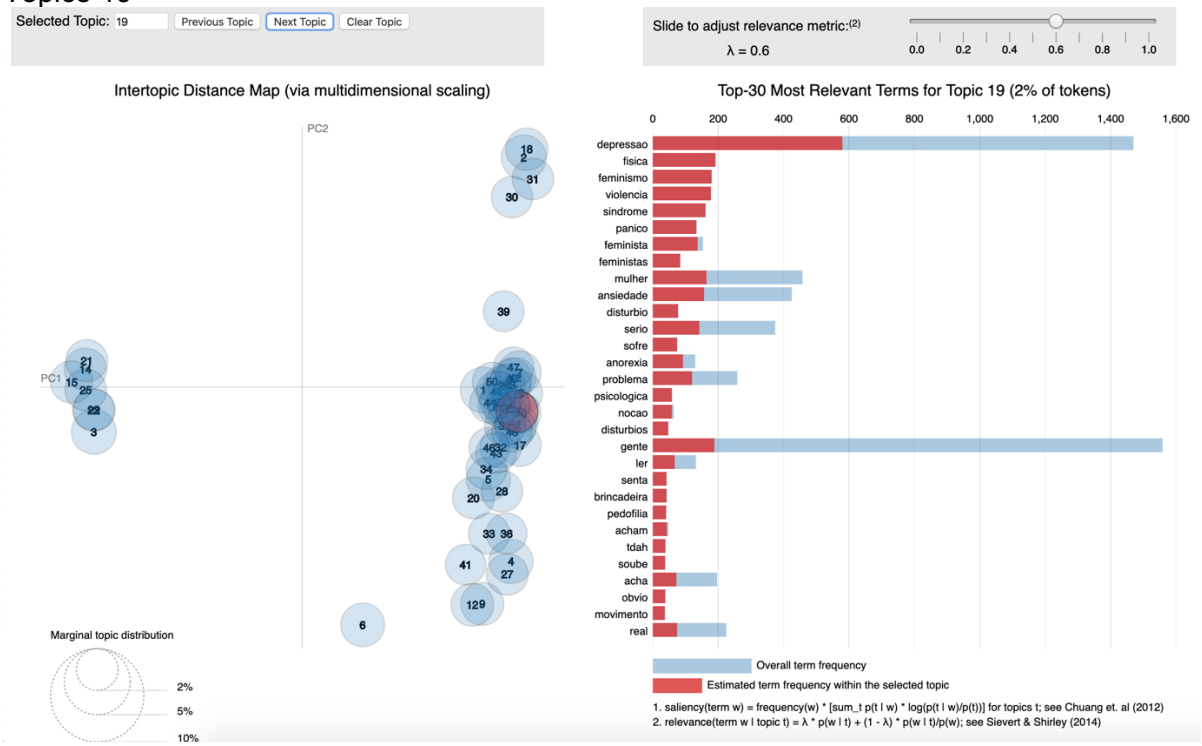
Tópico 17



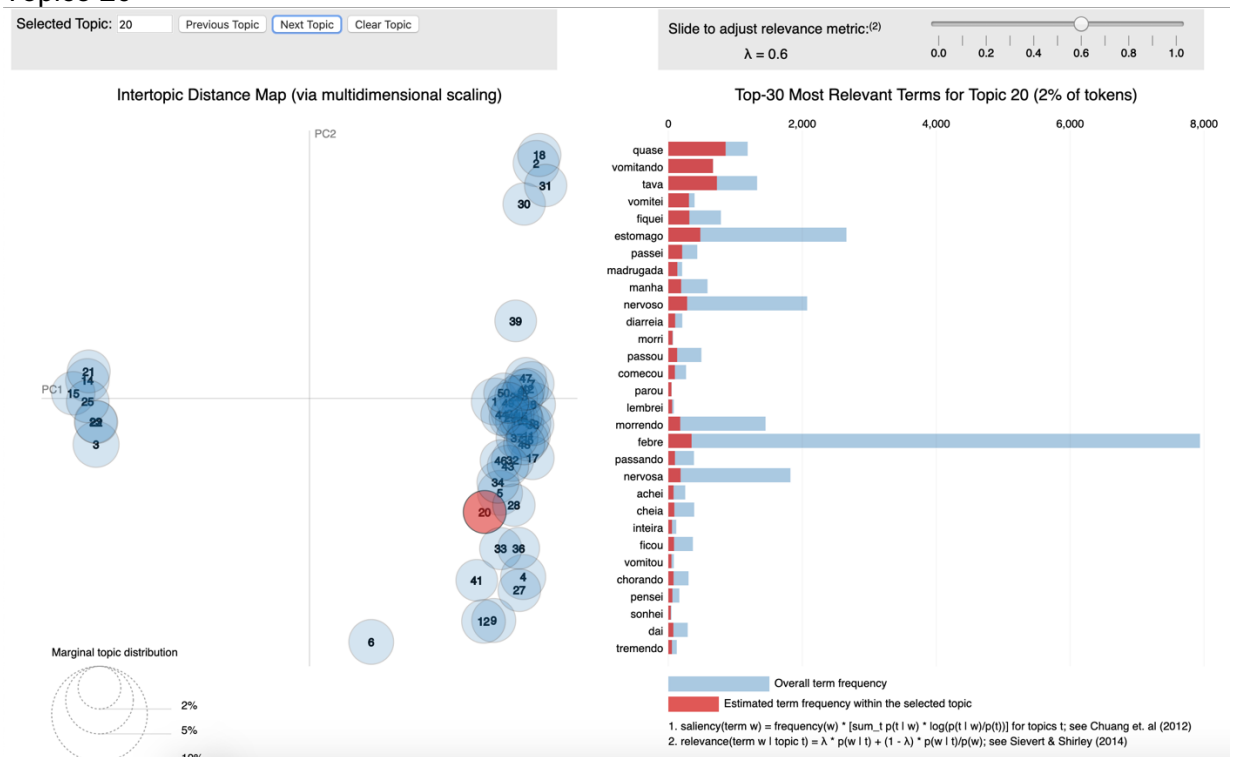
Tópico 18



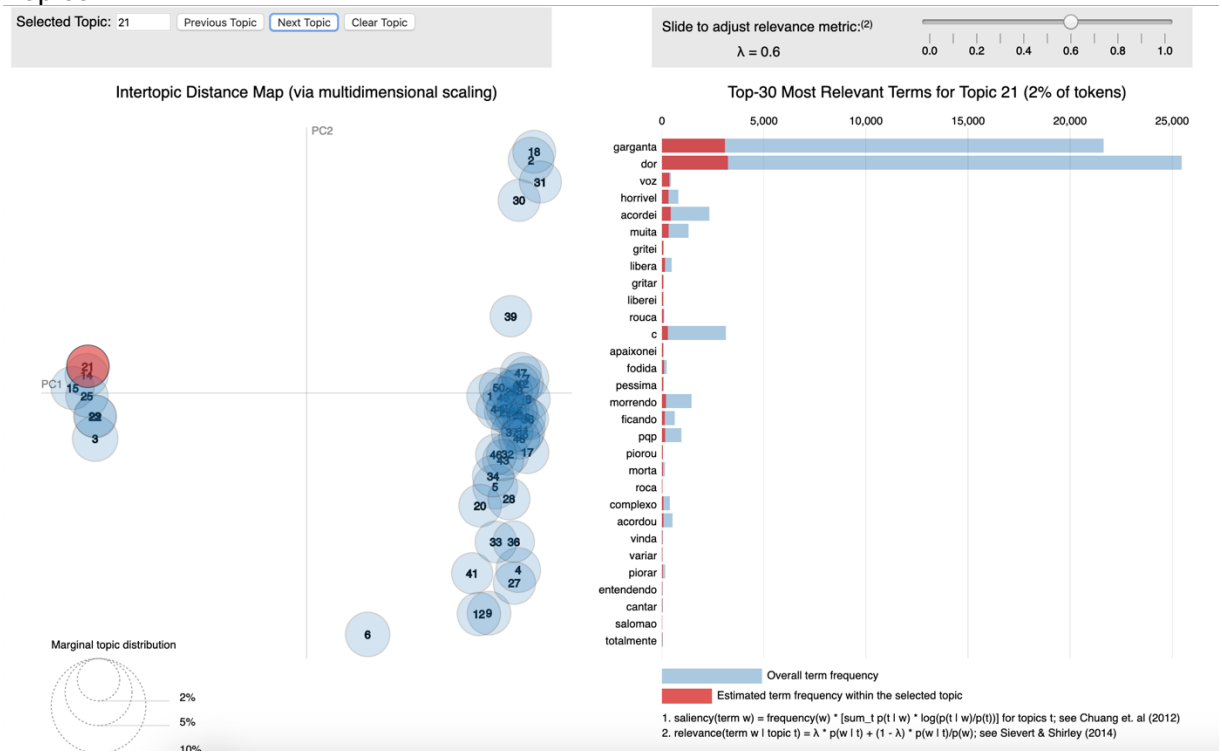
Tópico 19



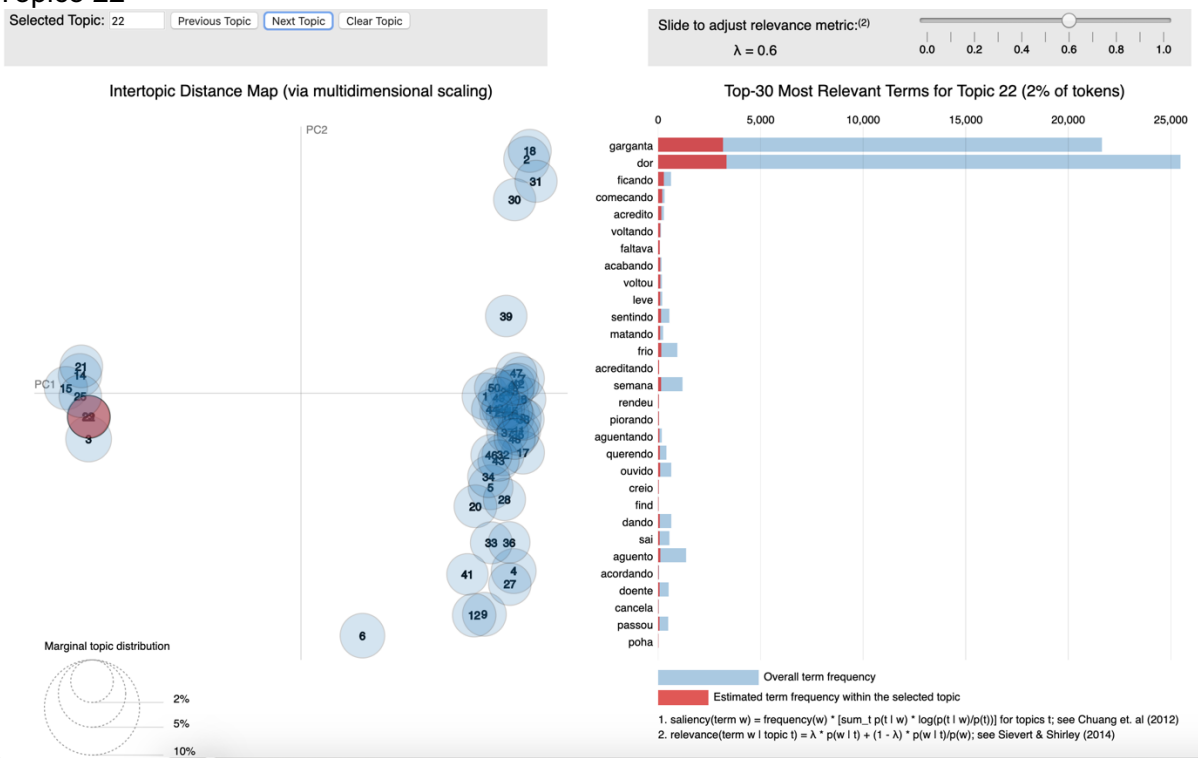
Tópico 20



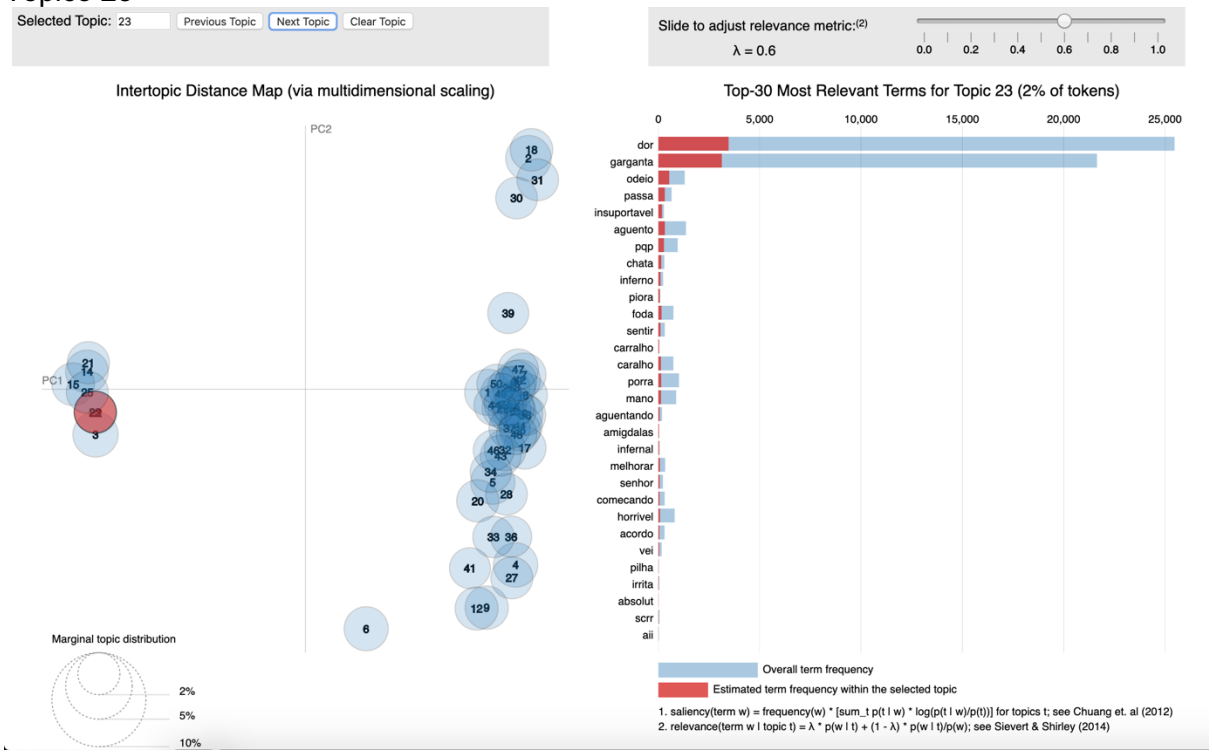
Tópico 21



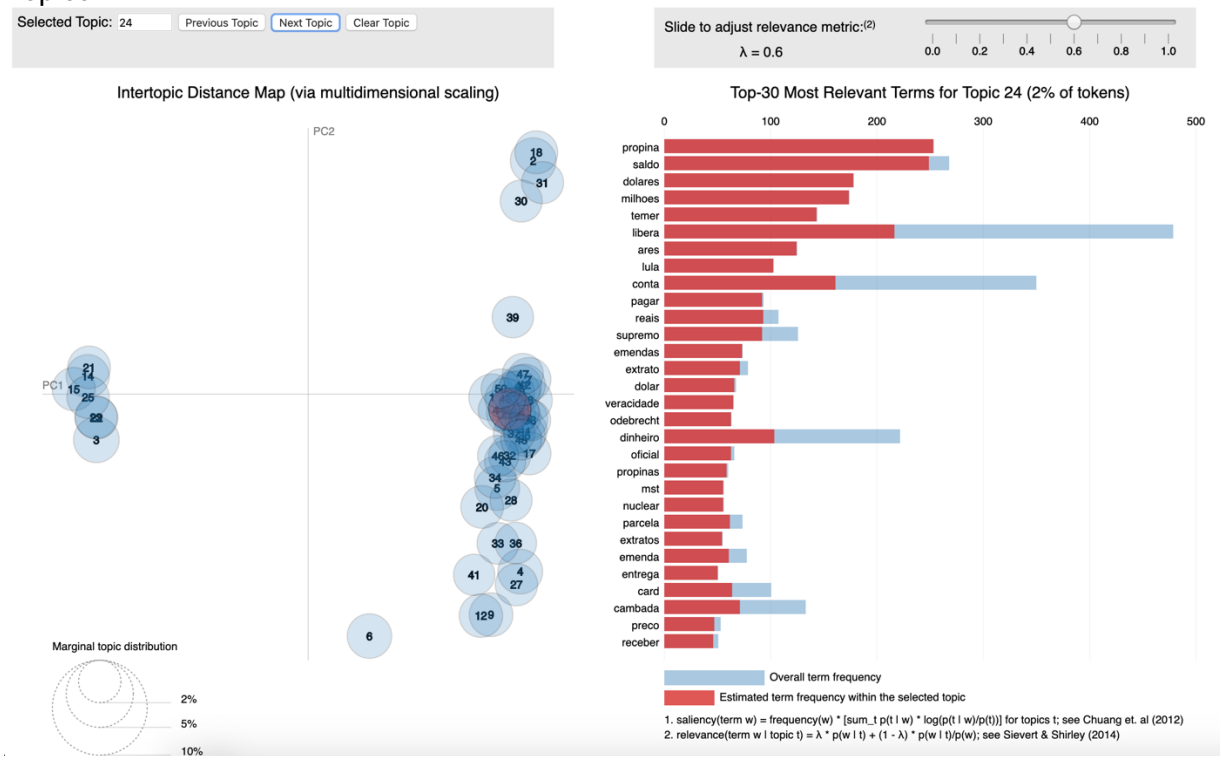
Tópico 22



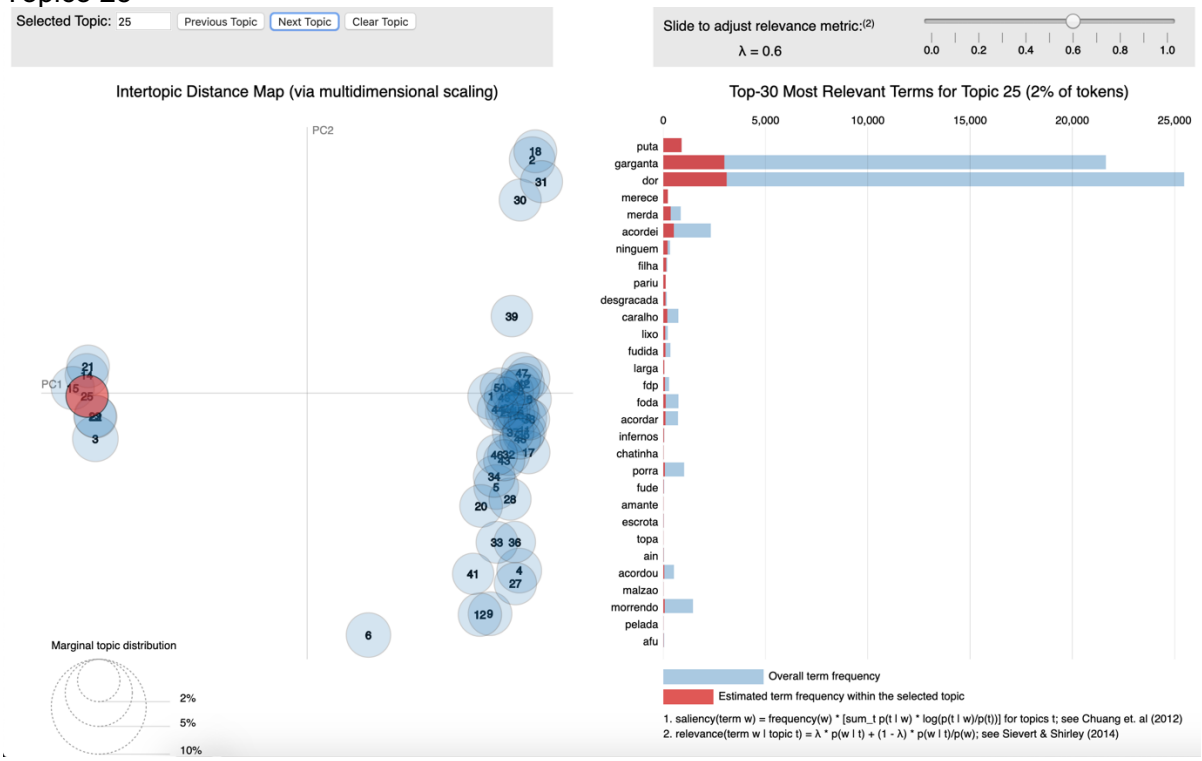
Tópico 23



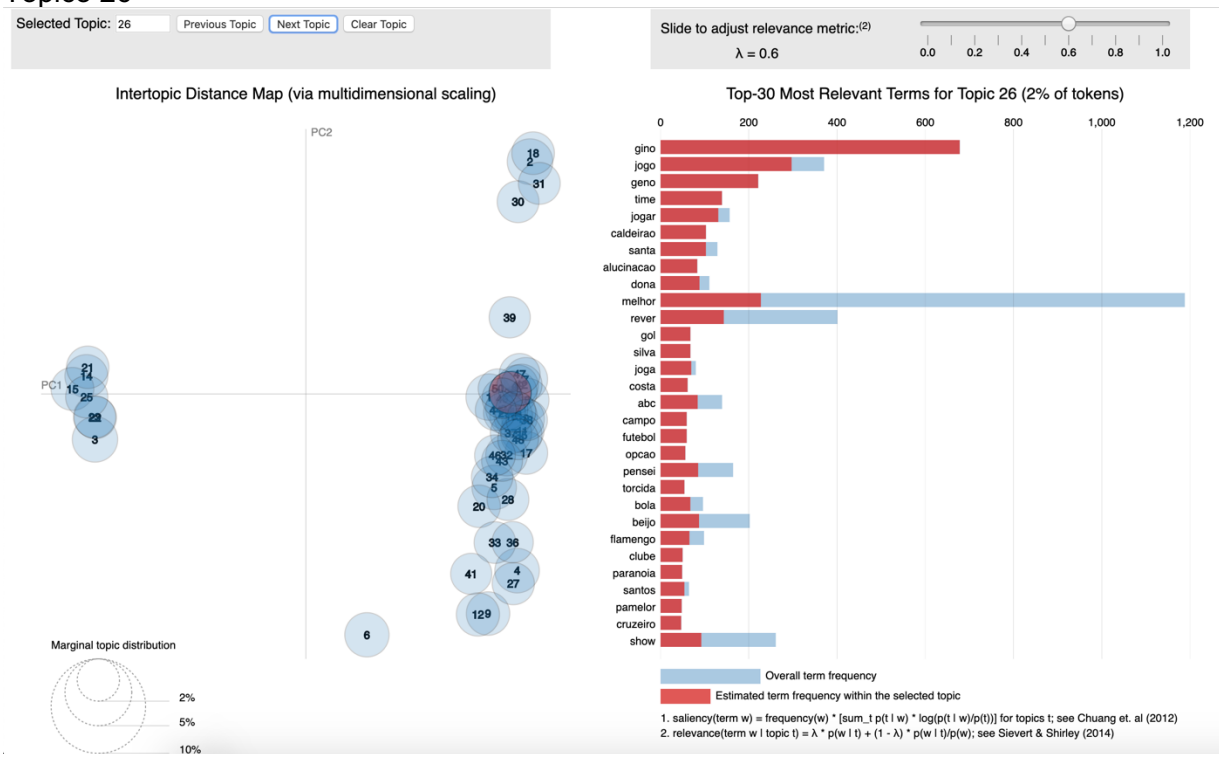
Tópico 24



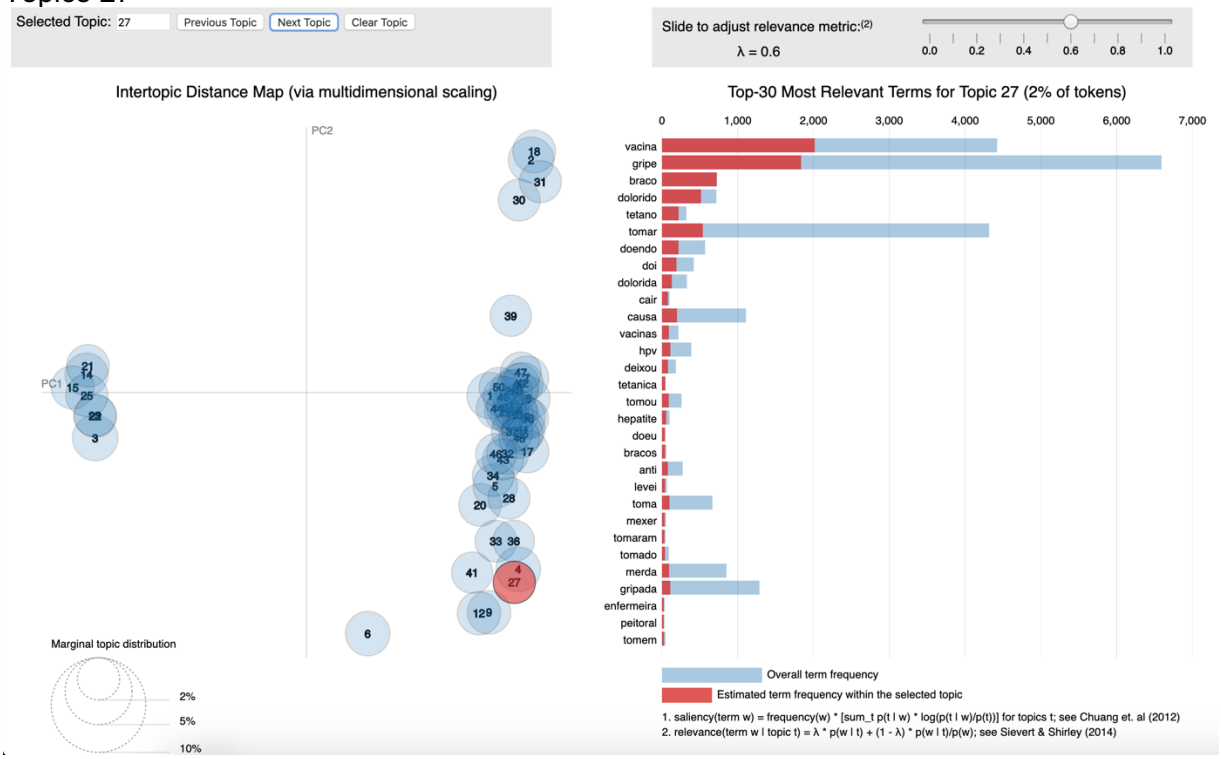
Tópico 25



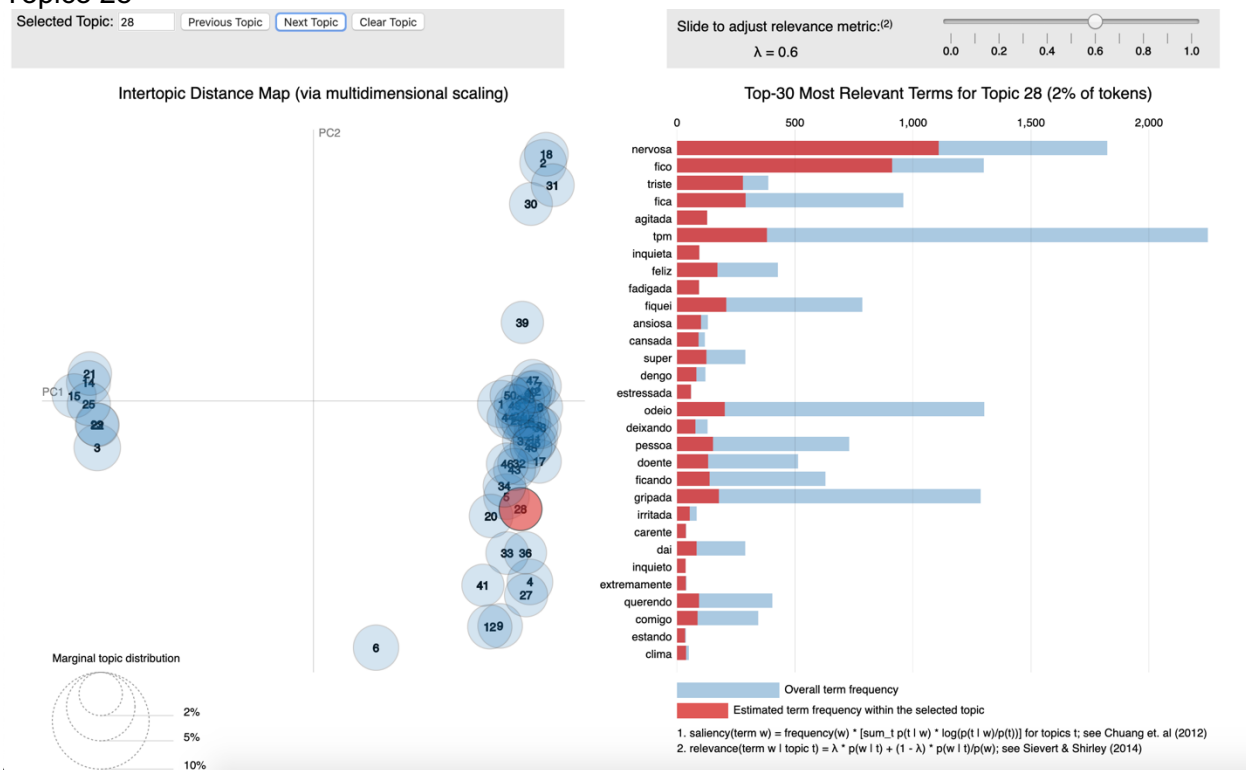
Tópico 26



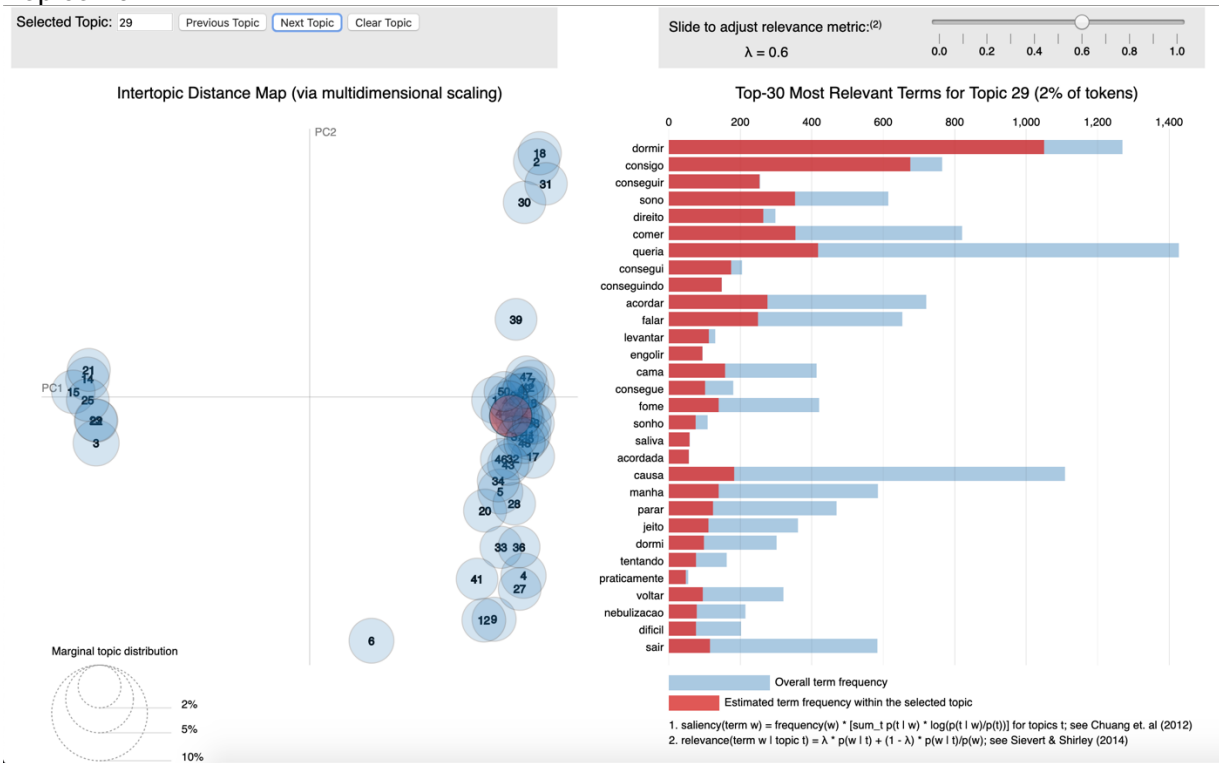
Tópico 27



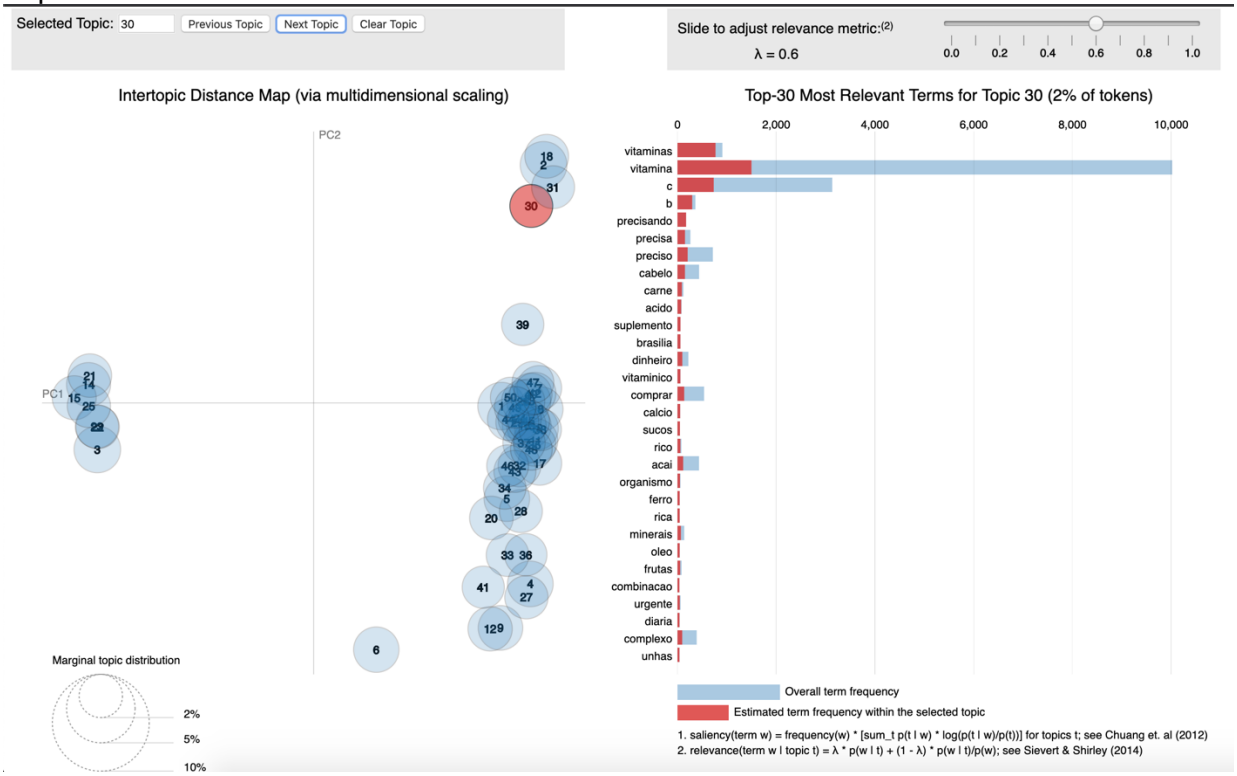
Tópico 28



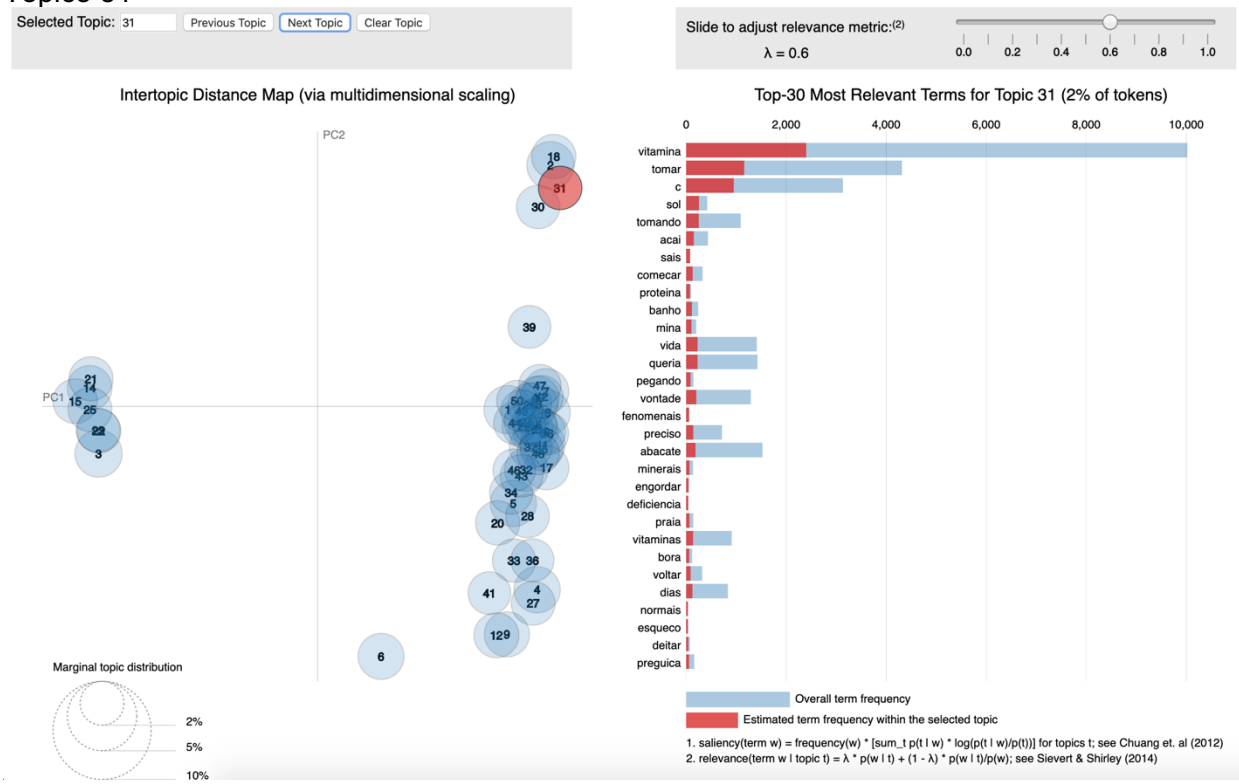
Tópico 29



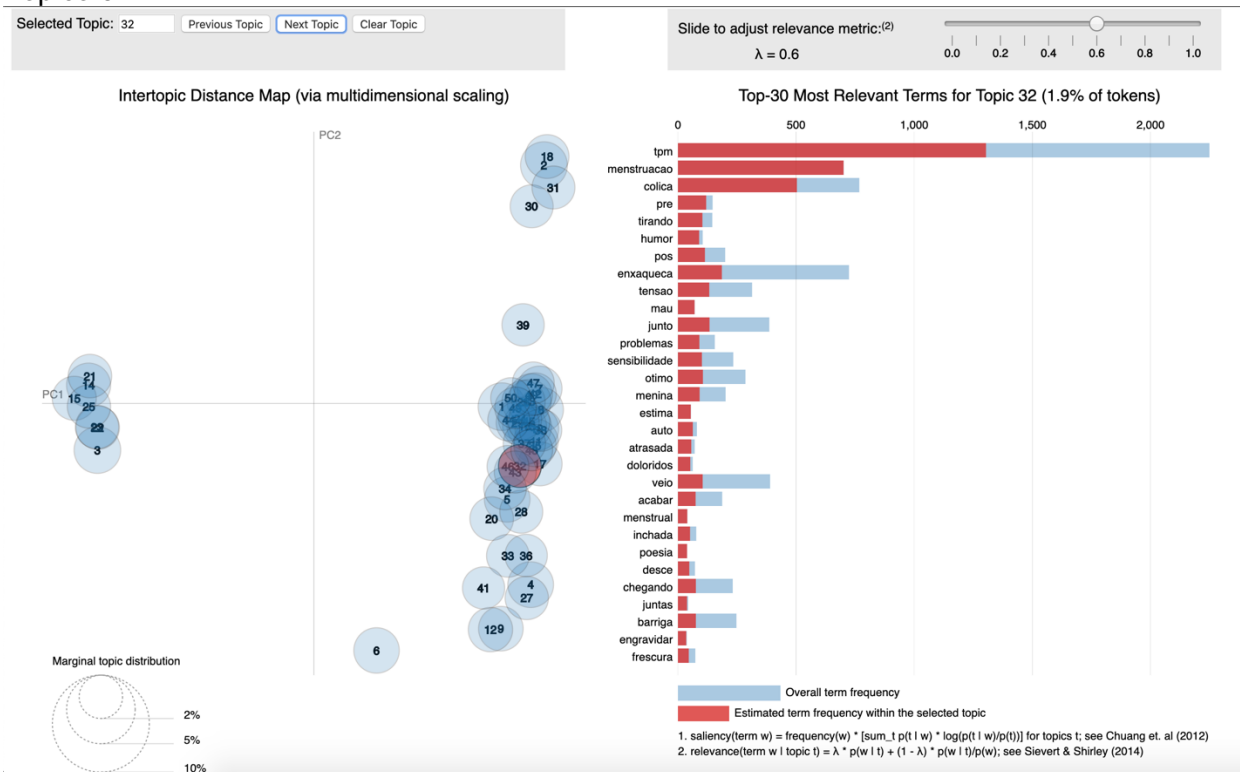
Tópico 30



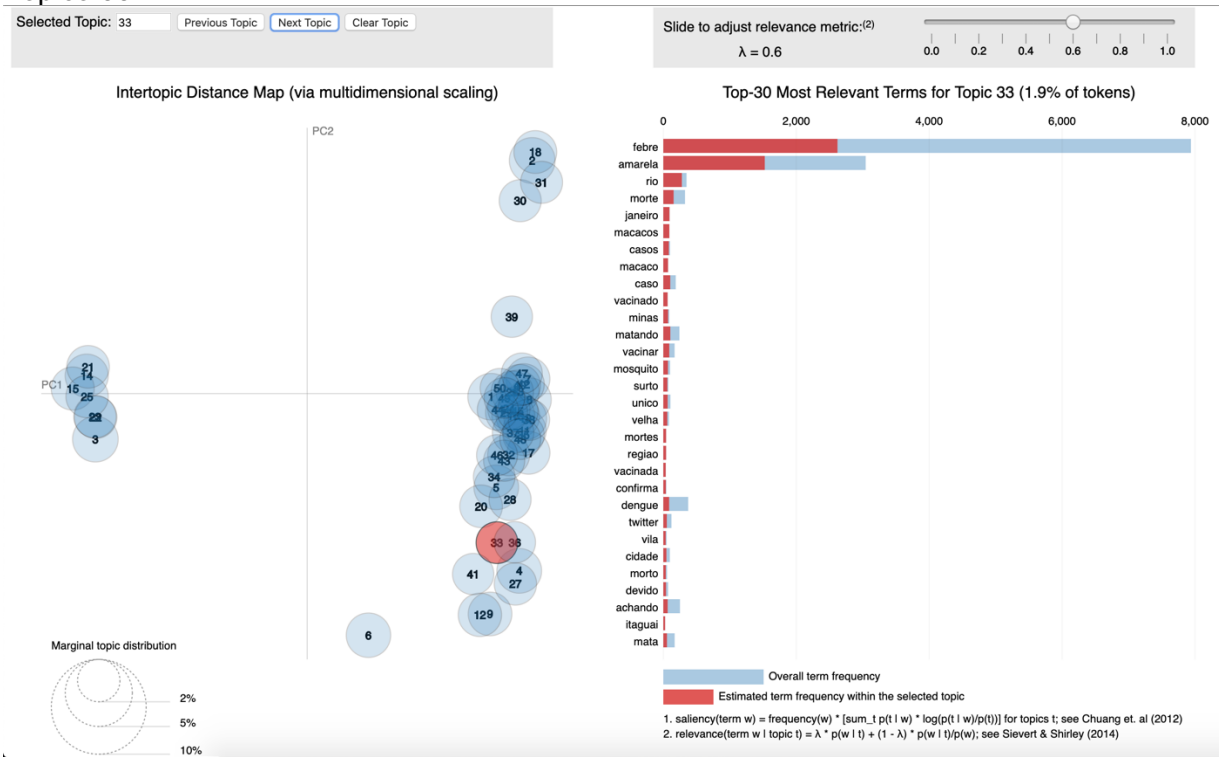
Tópico 31



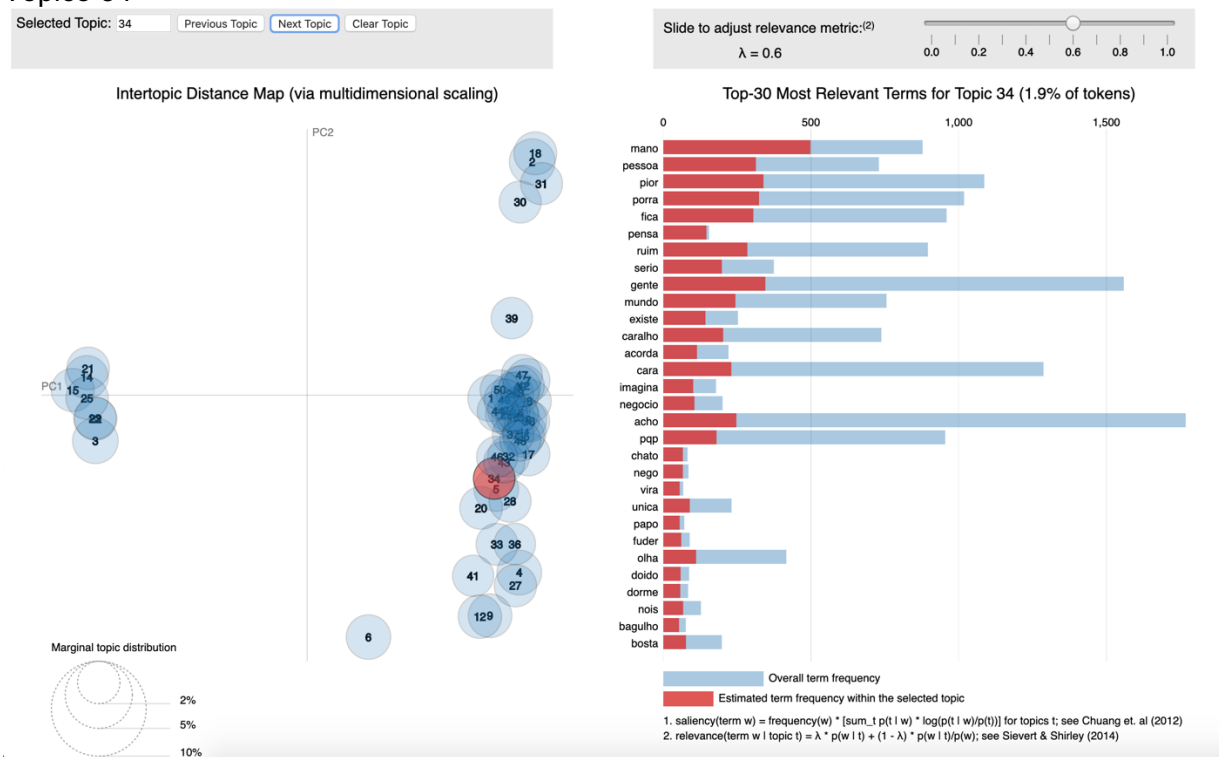
Tópico 32



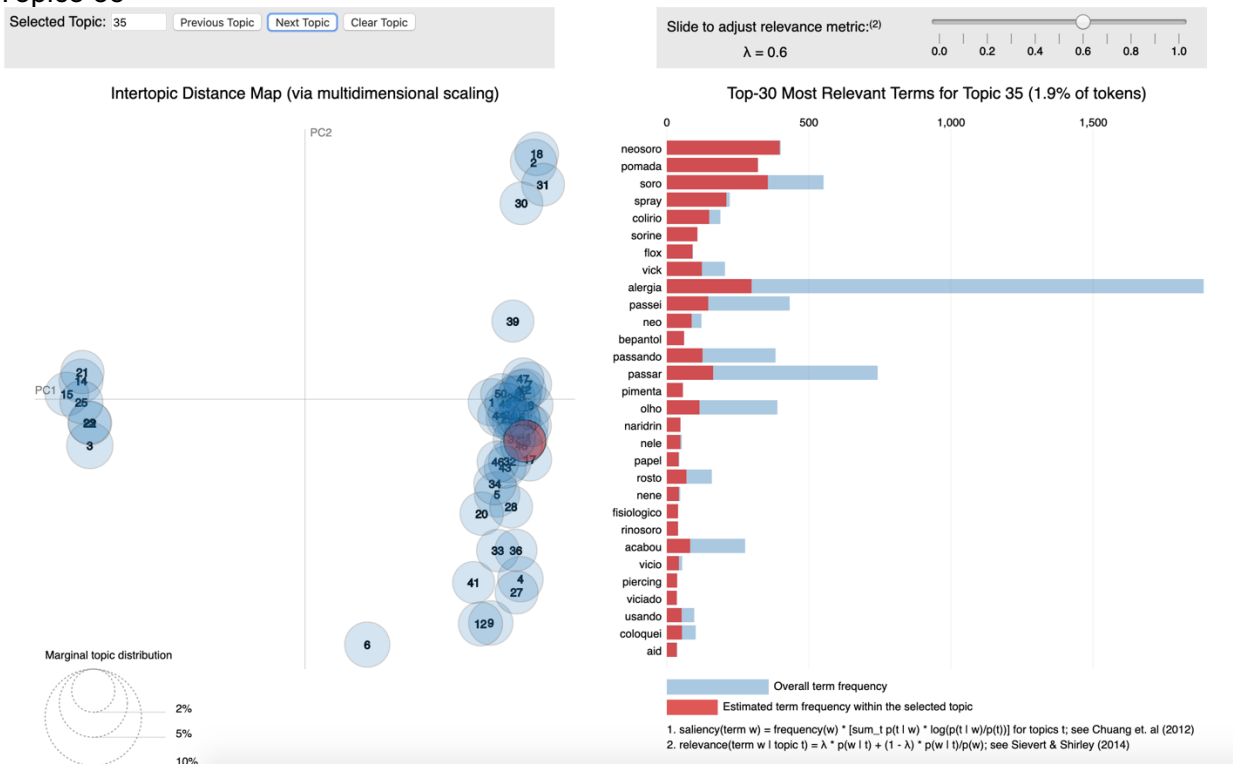
Tópico 33



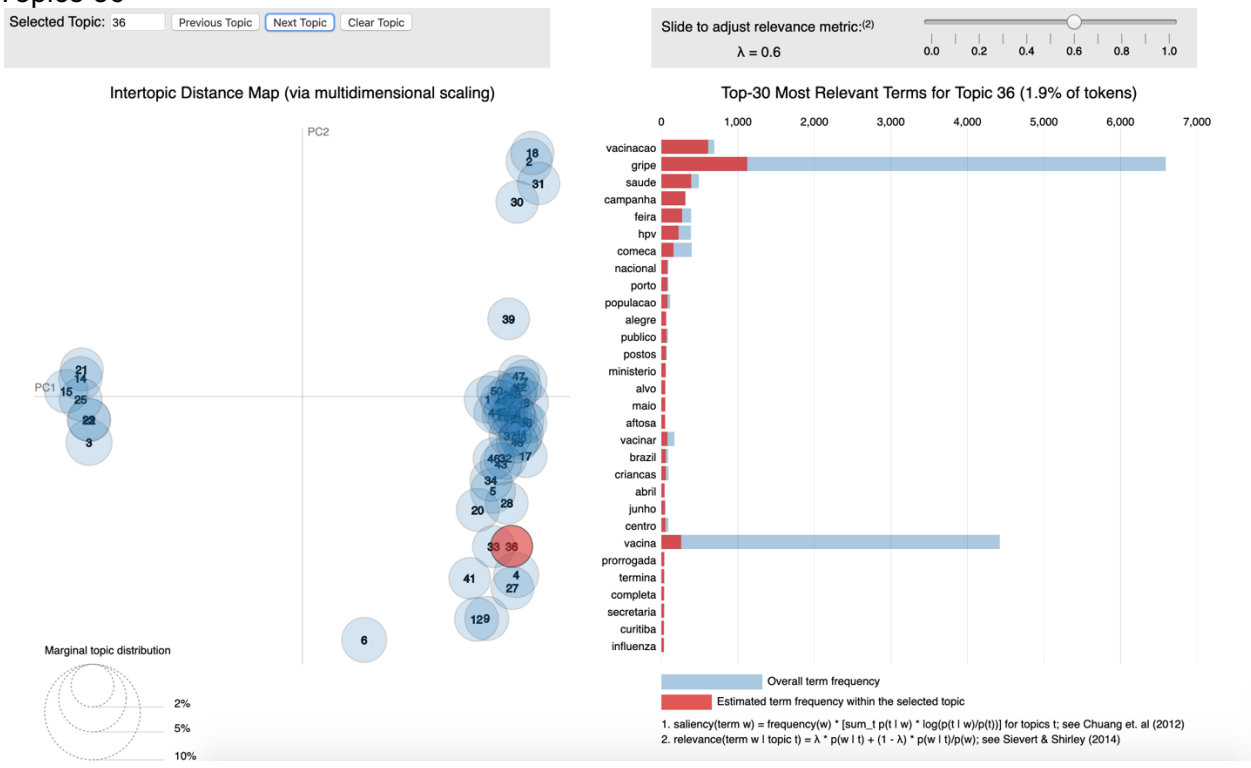
Tópico 34



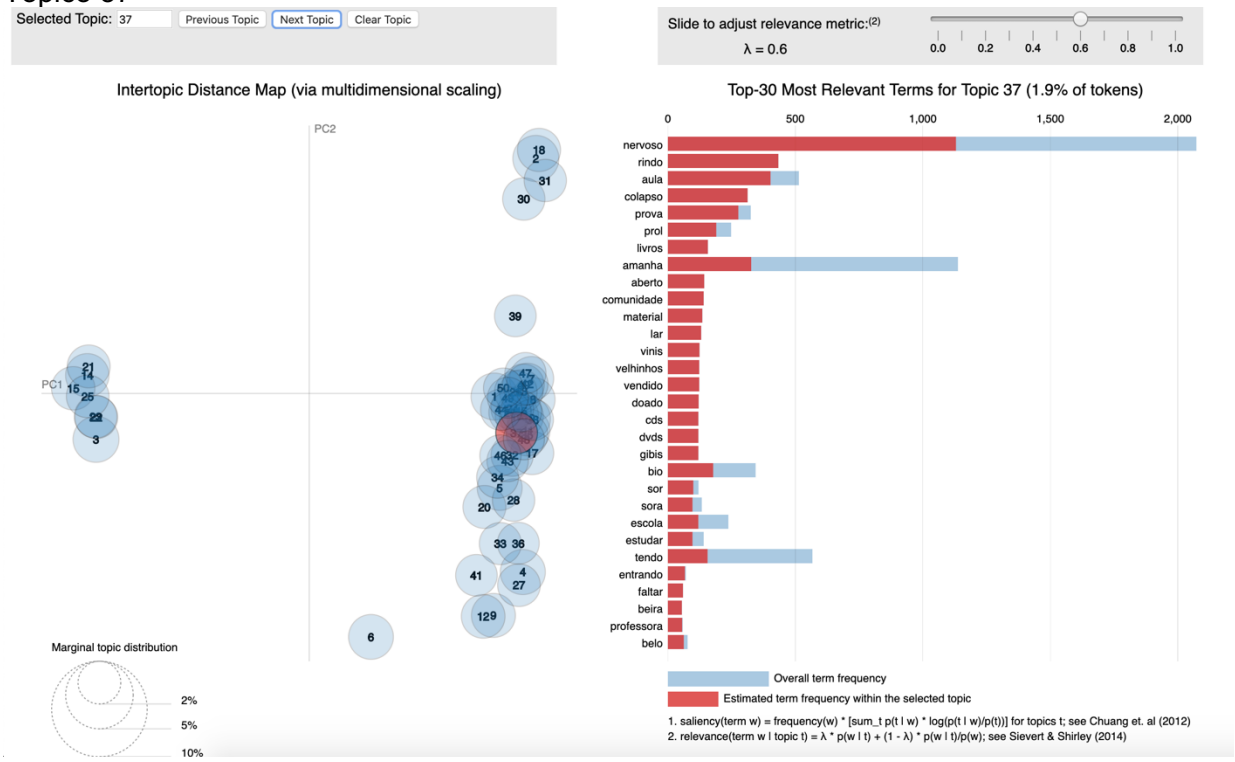
Tópico 35



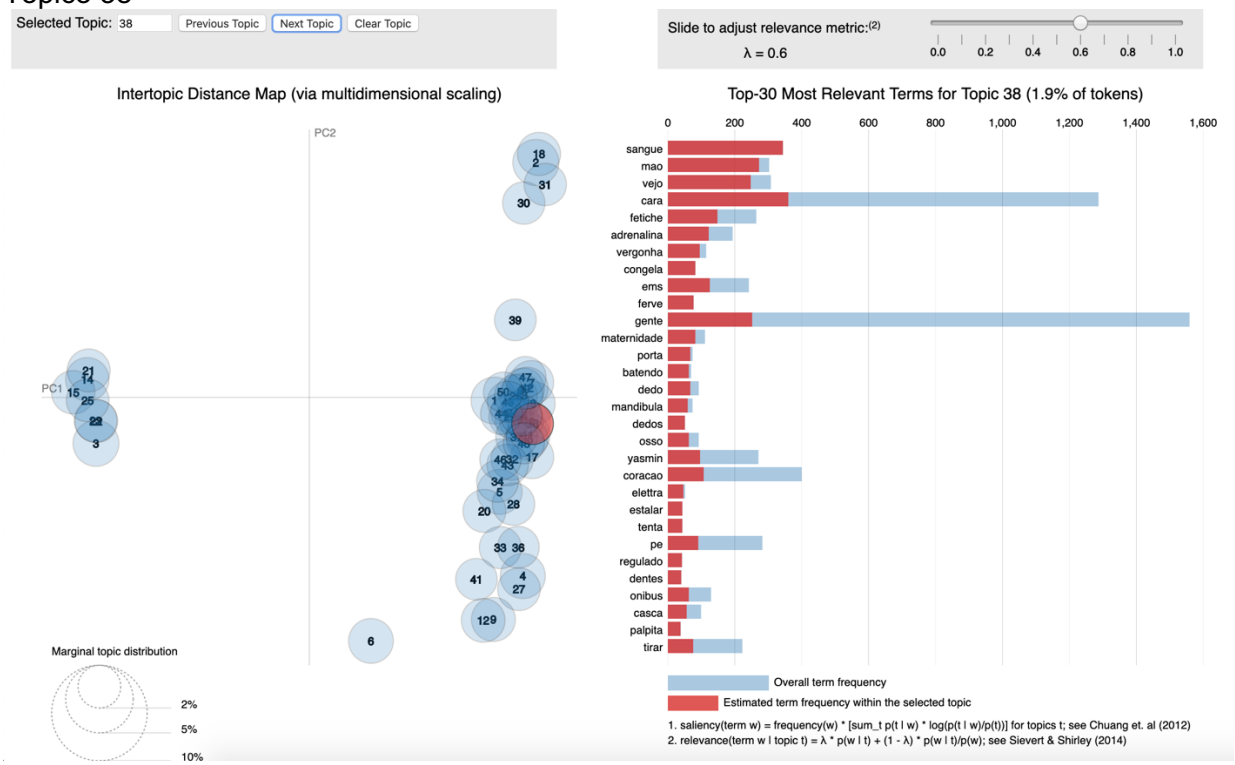
Tópico 36



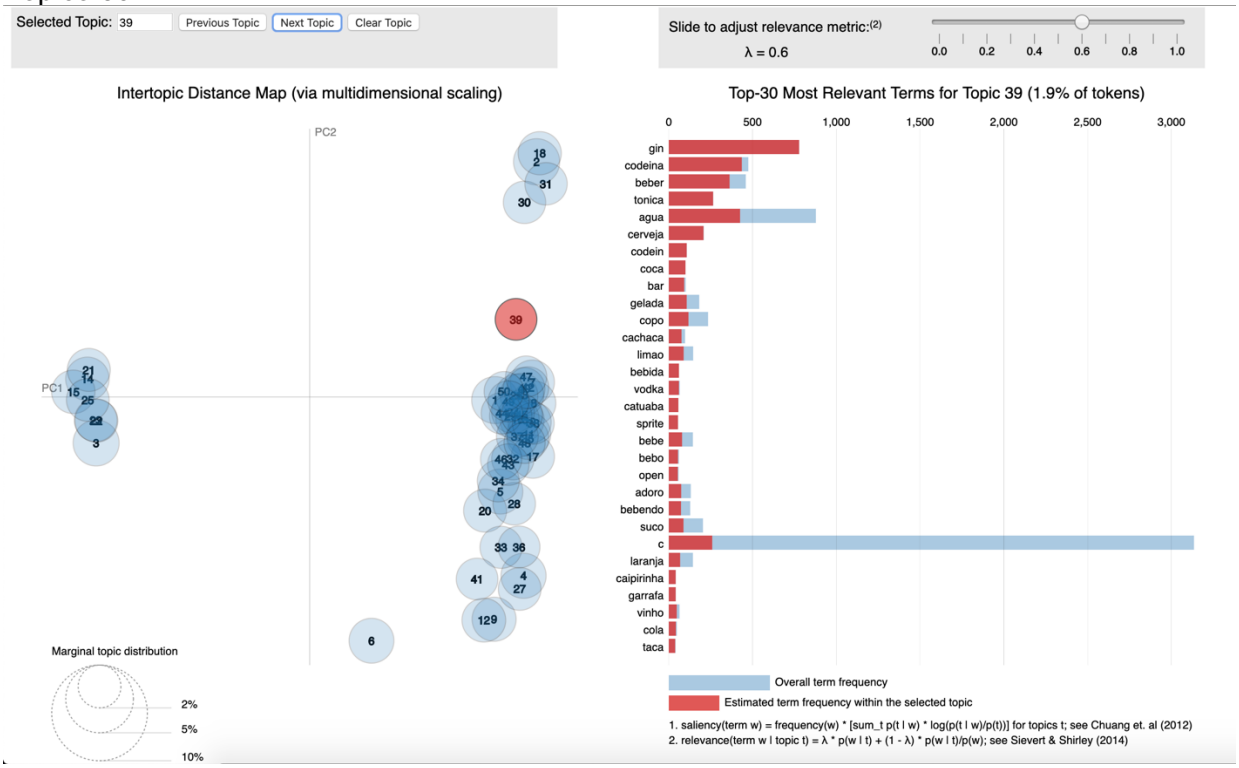
Tópico 37



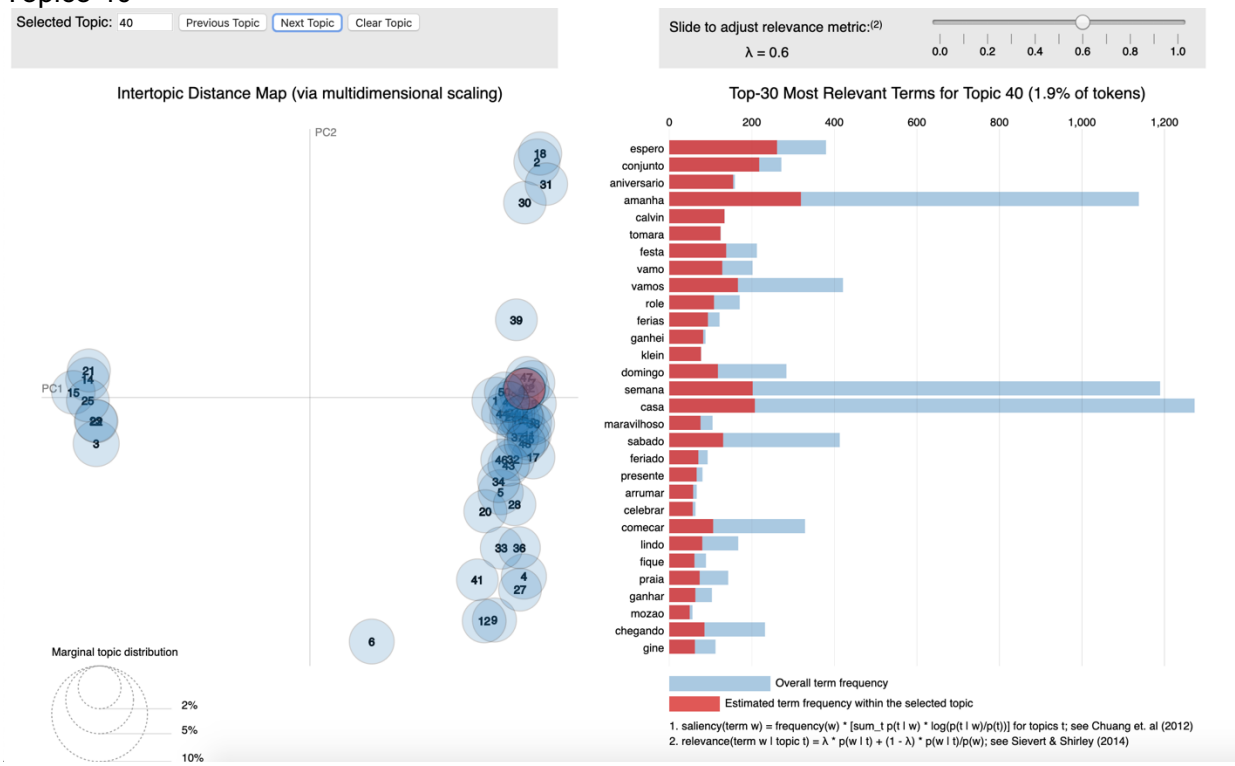
Tópico 38



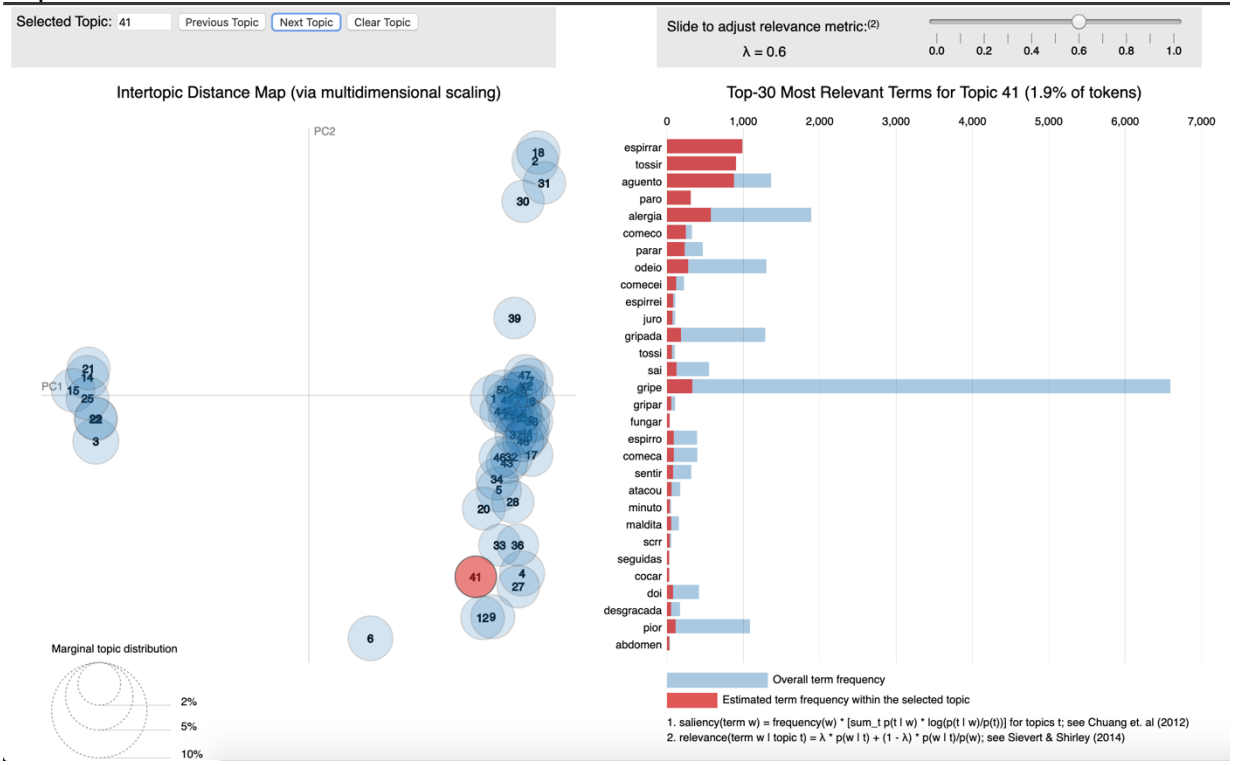
Tópico 39



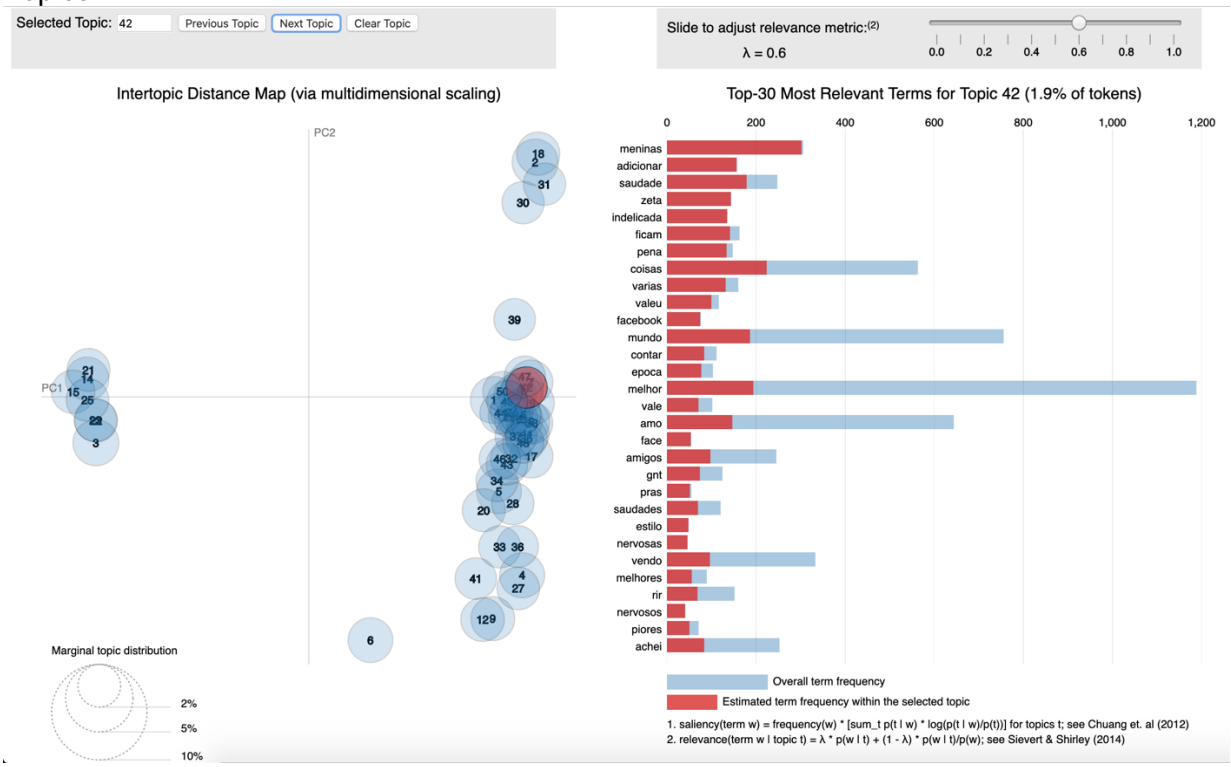
Tópico 40



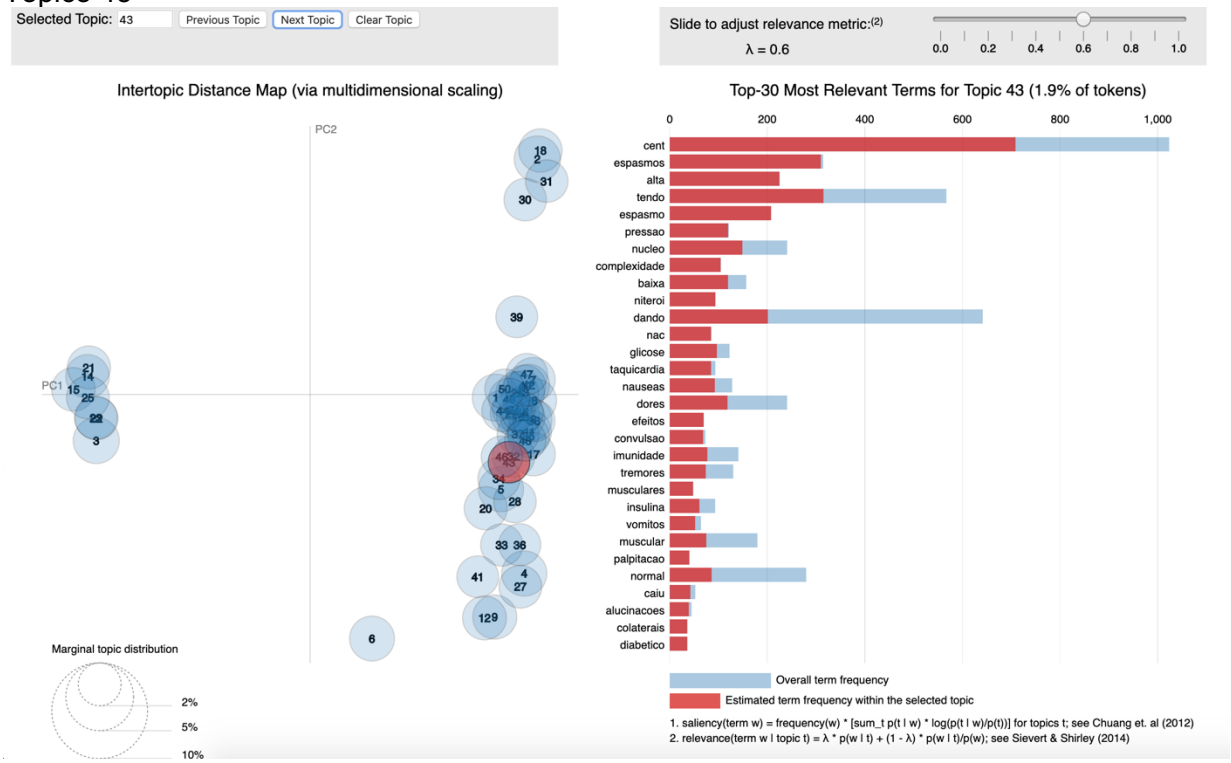
Tópico 41



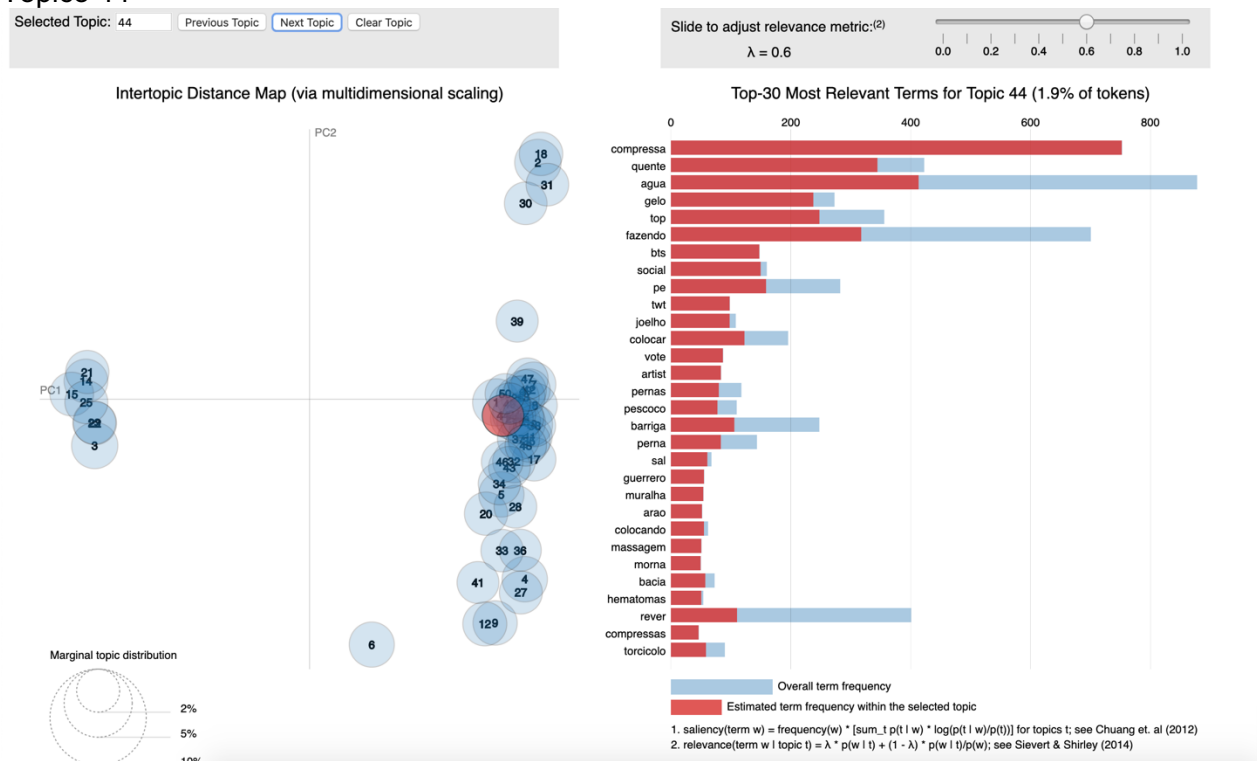
Tópico 42



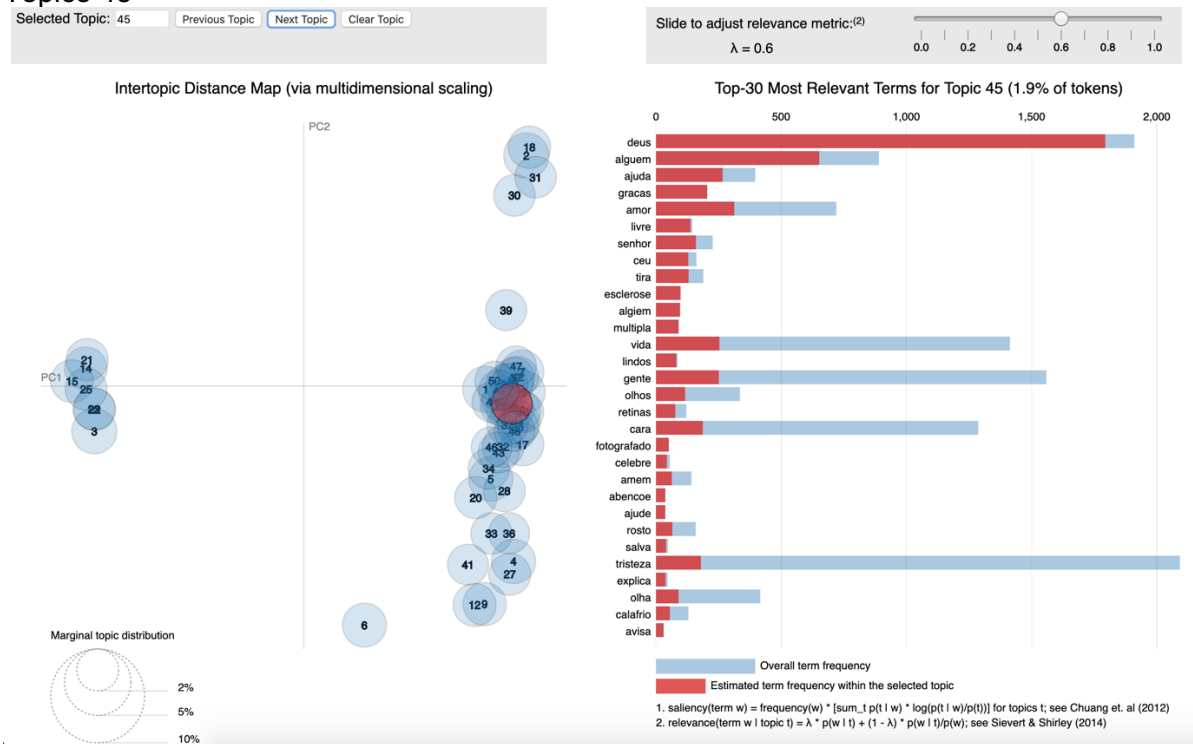
Tópico 43



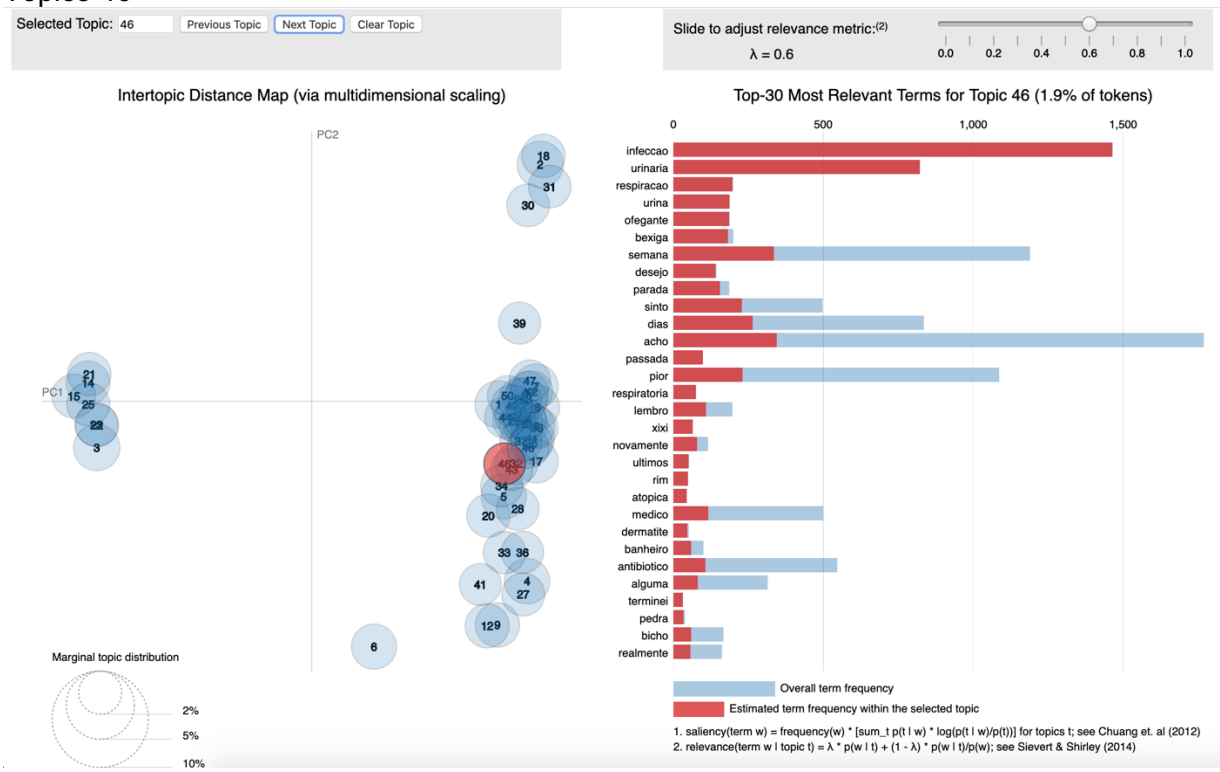
Tópico 44



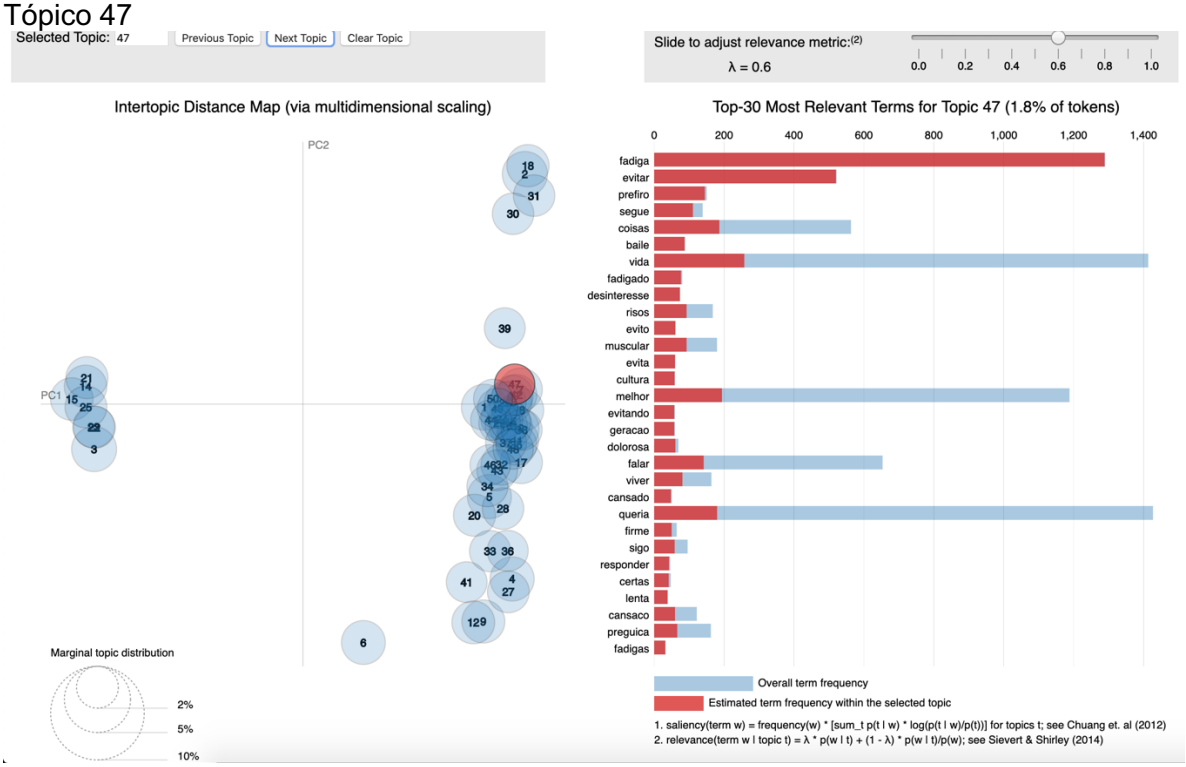
Tópico 45



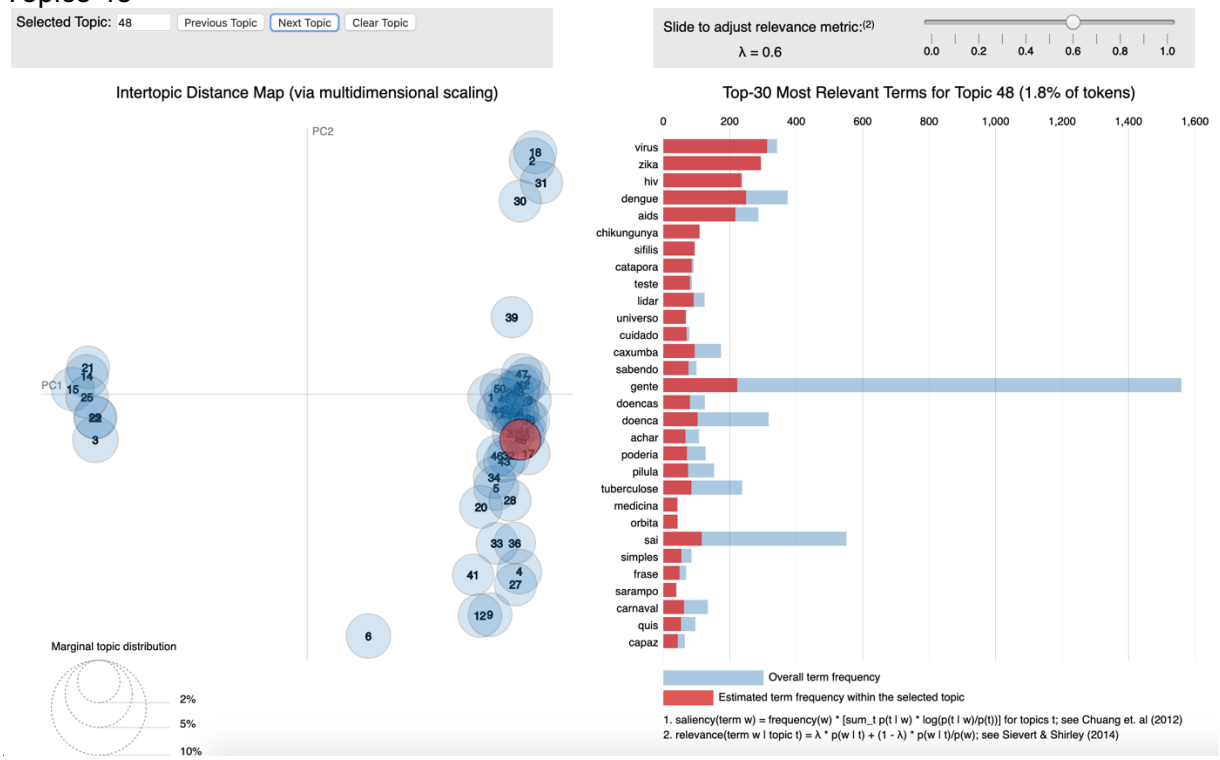
Tópico 46



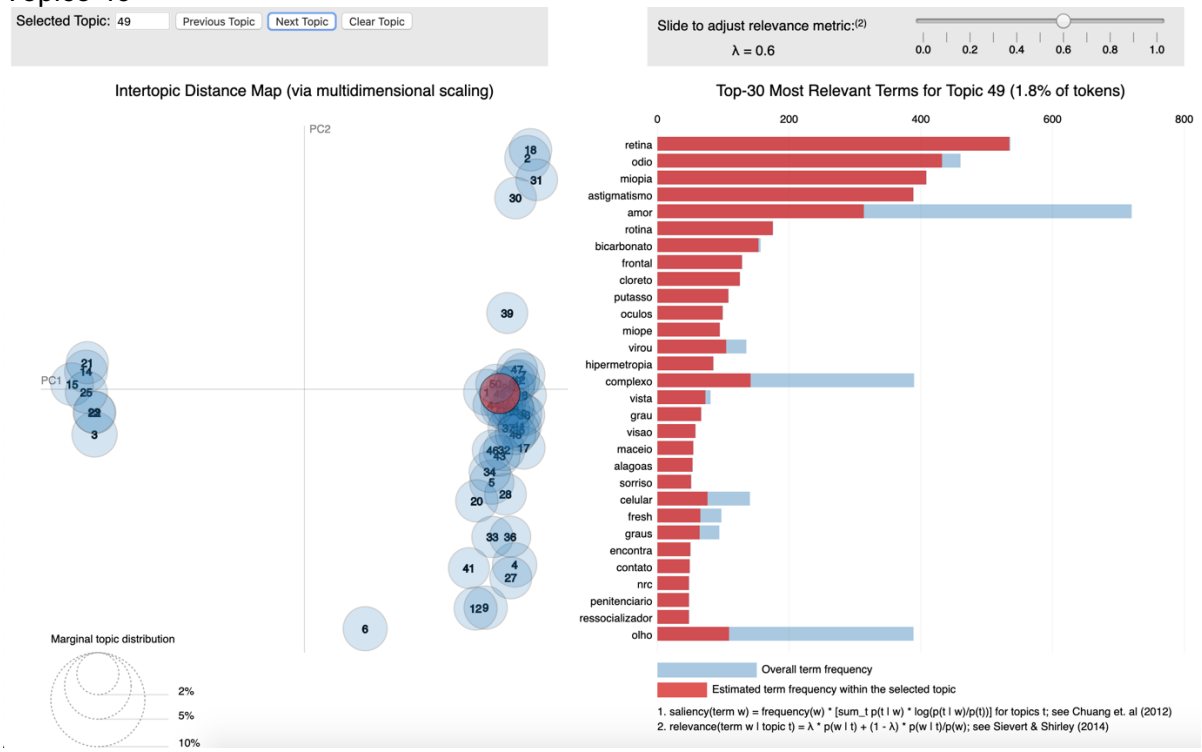
Tópico 47



Tópico 48



Tópico 49



Tópico 50