

Marcelo Bianchi Barata Ribeiro

**Modelagem de tópicos e interpretabilidade:
Uma proposta de visualização de resultados
implementada em D3.js**

Rio de Janeiro, Brasil

2020

Marcelo Bianchi Barata Ribeiro

**Modelagem de tópicos e interpretabilidade:
Uma proposta de visualização de resultados
implementada em D3.js**

Dissertação apresentada à Escola de Matemática Aplicada da Fundação Getúlio Vargas para a obtenção de Título de Mestre em Modelagem Matemática, na Área de Análise da Informação.

Fundação Getúlio Vargas – FGV

Escola de Matemática Aplicada – EMAp

Programa de Pós-Graduação em Modelagem Matemática

Orientador: Asla Medeiros e Sá

Coorientador: Renato Rocha Souza

Rio de Janeiro, Brasil

2020

Marcelo Bianchi Barata Ribeiro

Modelagem de tópicos e interpretabilidade:

Uma proposta de visualização de resultados implementada em D3.js/ Marcelo
Bianchi Barata Ribeiro. – Rio de Janeiro, Brasil, 2020-

87 p. : il. (algumas color.) ; 30 cm.

Orientador: Asla Medeiros e Sá

Dissertação (Mestrado) – Fundação Getúlio Vargas – FGV

Escola de Matemática Aplicada – EMap

Programa de Pós-Graduação em Modelagem Matemática, 2020.

1. Palavra-chave1. 2. Palavra-chave2. I. Orientador. II. Universidade xxx. III.
Faculdade de xxx. IV. Título

Marcelo Bianchi Barata Ribeiro

Modelagem de tópicos e interpretabilidade: Uma proposta de visualização de resultados implementada em D3.js

Dissertação apresentada à Escola de Matemática Aplicada da Fundação Getúlio Vargas para a obtenção de Título de Mestre em Modelagem Matemática, na Área de Análise da Informação.

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Asla Medeiros e Sá
Orientador

Renato Rocha Souza
Coorientador

Flávio Codeço Coelho
Convidado 1

Luis Gustavo Nonato
Convidado 2

Rio de Janeiro, Brasil
2020

*Este trabalho é dedicado à minha esposa, Mariana, ao meu filho, Eric,
e à memória dos meus pais, Bruno e Rosana.*

Agradecimentos

À professora Asla, pela orientação e pelo empenho dedicado ao meu projeto de pesquisa, que foi essencial desde o início.

Ao professor Renato, coorientador, pelas sugestões e direcionamentos fornecidos, assim como por todos os conselhos desde antes da minha inserção no mestrado.

À Mariana, minha esposa, por todo o seu apoio, incentivo e paciência ao longo desses dois anos, assim como por sua ajuda em desvendar o D3.

A todo o corpo docente da EMAp, pela excelência de ensino.

Aos meus colegas de mestrado, pelas trocas e por caminharem junto comigo nessa empreitada.

*“The city’s central computer told you?
R2D2, you know better than to trust a strange computer!”*
– C3PO (Star Wars)

Resumo

Nos últimos anos, diversos avanços foram promovidos no campo de modelagem de tópicos, seja por meio do desenvolvimento de novos algoritmos, seja em processos de avaliação, assim como pelo surgimento de novas ferramentas de visualização. Esta última frente avança devido à percepção de que modelos de tópicos fornecem nova capacidade exploratória de grandes coleções de documentos, o que, aliado a soluções de visualização, pode trazer nova percepção analítica ao especialista de domínio.

Este trabalho buscou introduzir uma solução interativa e de alta amplitude analítica, tendo como objeto de estudo uma coleção de documentos disponibilizada pela FGV/CPDOC. A metodologia envolveu uso de resultados provenientes de modelagem de tópicos e transformações dos dados para o processo de visualização, o que demandou o uso de distintas ferramentas de programação disponíveis. Após investigação do estado da arte, a hipótese principal é que haveria baixa disponibilidade de ferramentas de visualização de tópicos que incorporassem uma visão global do *corpus* acompanhada por um aumento gradual do nível de detalhamento, passando pela análise de agrupamentos de objetos viabilizada pela modelagem de tópicos, até a exploração de cada objeto.

A principal contribuição está na conceituação de uma nova ferramenta que atende a conceitos de granularidade, usuário-alvo e *data-ink ratio* por meio de uma linguagem de programação que forneça o máximo de flexibilidade. Por fim, conclui-se que haveria muito espaço de melhoria, seja por meio de aumento de interatividade, quanto por maior dedicação a etapas de pré-processamento no caso de coleções de documentos que tenham passado por processo de *OCR*.

Palavras-chaves: Modelagem de Tópicos. Visualização. Dados Textuais

Abstract

In recent years, several advances have been promoted in topic modeling, either through the development of new algorithms, or in evaluation processes, as well as by the emergence of novel visualization tools. This last field advances due to the realization that topic models provide new exploratory capabilities for large collections of documents which, combined with visualization solutions, can bring new insights to domain specialists.

This work sought to introduce a novel interactive and highly analytical solution, having as object of study a document collection provided by FGV/CPDOC. The methodology comprises the use of results from topic modeling and data transformations for the data visualization, which required using distinct programming languages available. After the investigation of the state of the art, the main hypothesis is that there would be low availability of visualization tools aimed at topic models able to incorporate a global view of the *corpus* together with a gradual increase on the level of detail, passing through the analysis of object clusters provided by topic modeling, until the exploration of each unique object.

The main contribution is the implementation of a novel tool that meets the concepts of granularity, target-user and *data-ink ratio* through a programming language that provides maximum flexibility. Finally, it is reckoned that there would be room for improvement, either through increased interactivity, or through greater dedication to pre-processing steps in the case of document collections that have gone through *OCR* processes.

Key-words: Topic Modeling. Visualization. Text Data.

Listas de ilustrações

Figura 1 – Representação de fluxograma de visualização proposto por (WANG et al., 2016).	25
Figura 2 – O modelo probabilístico gráfico do LDA permite demonstrar visualmente as dependências entre as variáveis aleatórias (BLEI, 2012). Segue explicação dos parâmetros introduzidos nesta Figura: N denota o conjunto de palavras em um documento, D representa a coleção de documentos no <i>corpus</i> , K seria o conjunto de tópicos, $\theta_d \sim Dir(\alpha)$ e $\beta_k \sim Dir(\eta)$.	32
Figura 3 – Exemplo ilustrativo que busca demonstrar a intuição do LDA. Os tópicos como distribuições de palavras ($\beta_{1:k}$) estão na parte esquerda da Figura, as proporções de tópicos para os documentos (θ_d) compõem o histograma à direita, as atribuições de tópicos para o d -ésimo documento (z_d) são representadas pelos círculos coloridos próximos ao histograma. Finalmente, as palavras observadas no documento (w_d) se encontram realçadas no meio dessa Figura (BLEI, 2012).	33
Figura 4 – Aplicação do LDA com estabelecimento de 100 tópicos sobre um <i>corpus</i> constituído por 17 mil artigos científicos. Ao lado esquerdo se encontram as proporções de tópicos obtidas por meio do mesmo artigo da Figura 3, enquanto que ao lado direito estão os 15 termos de maior frequência no caso dos 4 tópicos mais relacionados a esse artigo (BLEI, 2012).	33
Figura 5 – Conceito de granularidade abordado por Windhager et al.. Em cada camada, há determinado grau de distância visual da coleção, sendo que na camada inferior, há maior nível de detalhamento de cada objeto (WINDHAGER et al., 2018).	38
Figura 6 – Histograma dos documentos de acordo com contagem de palavras.	47
Figura 7 – Distribuição de documentos de acordo com legibilidade calculada. Obs: Documentos com menos de 10 sentenças foram filtrados devido à imprecisão potencial da medição.	49
Figura 8 – Principais idiomas do <i>corpus</i> .	50
Figura 9 – Tabela de documentos para um tópico selecionado.	56
Figura 10 – Segmento dos dados contidos no arquivo <i>topic_1.json</i> , que contém a totalidade dos dados visuais para o tópico 1.	57
Figura 11 – Dados contidos no arquivo <i>names_list_1.json</i> , relativo a nomes contidos em documentos.	57
Figura 12 – Dados contidos no arquivo <i>tokens_list_1.json</i> , relativo a palavras contidas em documentos.	57

Figura 13 – Visualização dos dados incorporados ao <i>D3.js</i> provenientes do arquivo <i>topic_1.json</i>	58
Figura 14 – Amostra dos dados gerados (10 primeiros documentos) para visualização da coleção (<i>Collection Overviews</i>).	59
Figura 15 – Grafo <i>Jigsaw</i> , representando relações entre documentos e entidades. (WARD; GRINSTEIN; KEIM, 2015).	62
Figura 16 – Visualização 100 Data Stories (< https://lab.christianlaesser.com/ >).	63
Figura 17 – Grafo com conexões entre entidades proveniente de mineração de dados em acervo de e-mails de Hillary Clinton (< http://history-lab.org/clinton >).	65
Figura 18 – Visualização em seu formato inicial, sem interação.	66
Figura 19 – Visualização ao aplicar a função <i>mouse over</i> , ativada quando o cursor do mouse passa por um nó do grafo.	67
Figura 20 – Visualização ao aplicar a função <i>mouse over</i> sobre um nó relativo a documento (camada intermediária).	68
Figura 21 – Página do CPDOC para o qual é feito o redirecionamento, contendo acesso à versão digitalizada do documento.	68
Figura 22 – Página contendo versão digitalizada do documento acessada por meio do endereço eletrônico anterior. Tais páginas são mantidas por prestador de serviços ao CPDOC.	69
Figura 23 – <i>Heatmap</i> representando o <i>score</i> entre documentos e tópicos.	71
Figura 24 – Demonstração da ferramenta <i>Bubble Timeline</i> . Fazendo uso da própria explicação de Schneider: “no eixo horizontal está representado o tempo e no vertical estão os tópicos. Para cada tópico foi definida uma cor diferente. A variação de intensidade de cor em cada círculo do gráfico representa a variação dos valores entre 0 e 1 da probabilidade de ocorrência de cada tópico ao longo do tempo.” (SCHNEIDER, 2014) . .	80
Figura 25 – Demonstração de uso da ferramenta <i>The Topic Browser</i> . Ao selecionar o tópico “125” (janela à esquerda), são expostas mais informações em outras janelas (GARDNER et al., 2010).	81
Figura 26 – Demonstração de uso da ferramenta <i>MetaToMaTo</i> . Na janela à esquerda é possível selecionar um tópico, enquanto que a janela à direita expõe os documentos relacionados (SNYDER et al., 2013).	81
Figura 27 – Navegador desenvolvido por Chaney e Blei que foi testado na Wikipédia. Esta visualização expõe o processo de exploração dos dados por meio dessa ferramenta (CHANAY; BLEI, 2012).	82
Figura 28 – Uma das visualizações possíveis por meio da ferramenta <i>Jigsaw</i> (GORG et al., 2007a).	82

Figura 29 – Exemplo da ferramenta <i>LDAvis</i> (SIEVERT; SHIRLEY, 2014), neste caso aplicado ao próprio <i>corpus</i> do CPDOC.	83
Figura 30 – Demonstração da ferramenta <i>Termite</i> . O lado esquerdo expõe os 30 termos mais frequentes, enquanto que o lado direito expõe os 30 mais salientes (CHUANG; MANNING; HEER, 2012).	84
Figura 31 – Demonstração da ferramenta <i>ThemeRiver</i> , que utiliza a metáfora de um rio de modo a representar variações da importância de tópicos ao longo do tempo (HAVRE; HETZLER; NOWELL, 2000)	85
Figura 32 – Grafo gerado por meio do Gephi, <i>layout Fruchterman Reingold</i>	86
Figura 33 – Grafo gerado por meio do Gephi, <i>layout Yifanhu</i>	87

List of abbreviations and acronyms

AAS	Antonio Azeredo da Silveira
CPDOC	Centro de Pesquisa e Documentação de História Contemporânea
CTM	Correlated Topic Model
DDO	US Declassified Documents Online
EMAp	Escola de Matemática Aplicada
FGV	Fundação Getulio Vargas
FRUS	Foreign Relations of the United States
HDP	Hierarchical Dirichlet Process
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
MCMC	Cadeia de Markov Monte Carlo
MRE	Ministério de Relações Exteriores
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
PMI	Pointwise Mutual Information
PCA	Principal Component Analysis
pLSI	Probabilistic Latent Semantic Indexing
sLDA	Supervised Topic Models
STM	Structural Topic Model
SVD	Singular Value Decomposition

Sumário

Introdução	23
I REFERENCIAIS TEÓRICOS	26
1 MODELAGEM DE TÓPICOS	27
1.1 Exemplos de aplicação	34
2 INTERPRETABILIDADE DE TÓPICOS GERADOS	36
2.1 Métodos de análise automática	36
2.2 Visualização de dados para interpretação de tópicos	37
II PASSOS METODOLÓGICOS	43
3 BASE DE DADOS	46
3.1 Seleção e obtenção dos dados	46
3.2 Preparação dos dados (pré-modelagem)	47
3.2.1 OCR	47
3.2.2 Limpeza de dados	48
4 MODELAGEM DE TÓPICOS	52
5 GERAÇÃO E TRATAMENTO DE DADOS	54
5.1 Geração de metadados	54
5.2 Tratamentos dos dados	55
III VISUALIZAÇÃO	60
6 SELEÇÃO DO MODELO DE VISUALIZAÇÃO	61
Considerações finais	72
REFERÊNCIAS	74
ANEXOS	79
ANEXO A – REFERÊNCIAS DE VISUALIZAÇÃO	80

**ANEXO B – TESTES DE VISUALIZAÇÃO COM A FERRAMENTA
GEPHI 86**

Introdução

O uso de ferramentas de automatização de análise textual, advindas de avanços de modelos de Aprendizado de Máquina (*Machine Learning*) e processamento de linguagem natural (NLP), traz novas oportunidades de trabalho ao ampliar o conjunto de técnicas analíticas disponíveis a campos diversos. No caso das ciências humanas, por exemplo, observa-se contínua expansão da disponibilidade de dados digitais, resultado da digitalização de objetos e documentos que compõem patrimônio cultural e histórico. Tal fenômeno tem motivado o uso de novas tecnologias de modo a aperfeiçoar a geração de conhecimento a partir de grandes volumes de dados e, segundo Windhager et al. (WINDHAGER et al., 2018), foi um fator determinante para o surgimento da área de Humanidades Digitais (*Digital Humanities*), principalmente no campo de visualização da informação. Por outro lado, o acesso a novas técnicas e ferramentas traz desafios para seus usuários, devido à necessidade de preparação dos dados, de adaptação de algoritmos, de especificação de parâmetros e de interpretação de resultados por meio da assistência de métodos estatísticos ou de visualização. Executadas essas tarefas, viabiliza-se a análise de um grande número de documentos, mas com um grau de intervenção humana reduzido se comparado com o trabalho manual tradicionalmente conduzido nessas áreas.

Tal como afirma Alencar et al., atualmente, documentos textuais são amplamente disponíveis em formato digital e proveem fonte valiosa de informação. No entanto, para analistas de diversas áreas, ainda há o desafio de acessar e interpretar os dados (ALENCAR; OLIVEIRA; PAULOVICH, 2012). No âmbito de visualização de dados textuais, existe o problema central sobre como formular uma representação adequada de tais dados, o que é comumente resolvido por meio de seleção de palavras-chave relevantes, coletadas de acordo com as respectivas frequências ou por meio de outras métricas tais como as advindas de modelagem de tópicos¹ (*topic modeling*) (WANNER et al., 2014). Nesta dissertação, será detalhada uma solução de visualização de dados voltada para o aprimoramento do processo de interpretação de resultados gerados por meio de modelagem de tópicos.

O *corpus* utilizado para testar a solução de visualização advém de parte do acervo de Antônio Azeredo da Silveira, série *Ministério de Relações Exteriores* (MRE), coleção que reúne seus documentos como Ministro de Relações Exteriores do Brasil entre 1974 e 1979. A coleção faz parte do acervo da FGV/CPDOC² e a parte de mineração dos

¹ Modelagem de tópicos envolve modelos probabilísticos de Aprendizado de Máquina não supervisionado. O assunto será mais bem explicado ao longo do Capítulo 1.

² O CPDOC concedeu direito de acesso a essa base de dados para o desenvolvimento do trabalho em dezembro de 2018. O CPDOC é a Escola de Ciências Sociais e Centro de Pesquisa e Documentação de História Contemporânea do Brasil, sendo parte da Fundação Getulio Vargas (FGV). As principais atividades do CPDOC são coletar e preservar coleções de arquivos pessoais e história oral considerados

dados é proveniente de trabalho anterior desenvolvido em projeto junto ao FGV/CPDOC, FGV/EMAp³ e *Columbia University*. A motivação desse trabalho prévio foi a utilização de técnicas de mineração de dados (RIBEIRO; MORELI; SOUZA, 2018).

Os resultados desse trabalho anterior serão aproveitados para o desenvolvimento da visualização. Para realizar a extração de informação contida no *corpus* utilizado em tal trabalho, foram utilizadas técnicas como modelagem de tópicos e processamento de linguagem natural. Além disso, foi necessário adotar estratégias de tratamento dos dados de modo a prepará-los para as fases posteriores de geração de metadados. Quanto ao processamento de linguagem natural, foram aplicados procedimentos de extração de entidades de modo a minerar dados sobre pessoas presentes em cada documento (RIBEIRO; MORELI; SOUZA, 2018).

A proposta de visualização tem como base um problema central aos resultados da aplicação de modelagem de tópicos: o fato de demandar por procedimentos de validação que costumam necessitar do trabalho de especialistas de domínio. Existem duas estratégias distintas para lidar com tal questão: análise automática, por um lado, e, por outro, soluções que incorporam visualização da informação. A primeira envolve o uso de métricas de estatística que mensuram, por exemplo, coerência de tópicos, enquanto que a segunda abarca diversos modelos de visualização, em sua maioria interativos, de modo a trazer, da forma mais eficaz possível, uma compreensão dos tópicos gerados, ou, ao menos, melhorar sua interpretabilidade.

Trabalhos de visualização que envolvem dados textuais podem ser parcialmente sintetizados por Matthew Ward (WARD; GRINSTEIN; KEIM, 2015), tais como nuvens de palavras e *TextArcs*⁴. Outra proposta de visualização, que pode ser considerada como inspiração para este trabalho, chamada de *Jigsaw*, é uma ferramenta interativa voltada para visualização e exploração de um *corpus* que permite realçar um documento e descobrir relações com entidades presentes no mesmo. Sua estrutura é apresentada como um grafo de inter-relações, fornecendo apenas uma amostragem da base de dados original⁵. Um resumo mais recente de métodos pode também ser encontrado em Shixia Liu et al. (LIU et al., 2018). Os trabalhos mencionados serão explicados com mais detalhamento ao longo dos Capítulos 2.2 e 6.

Quanto a visualizações voltadas para modelagem de tópicos, diversos exemplos serão fornecidos ao longo desta dissertação, principalmente na parte de referenciais teóricos. De qualquer modo, é importante ressaltar alguns trabalhos que serviram como fontes de inspiração centrais: a ferramenta *LDAvis* desenvolvida por Carson Sievert e Kenneth Shirley (SIEVERT; SHIRLEY, 2014), a ferramenta *Jigsaw*, que serviu de inspiração para

³ importantes para a compreensão da história brasileira.

⁴ Escola de Matemática Aplicada da Fundação Getulio Vargas.

⁴ Eficaz para mostrar sequências repetidas de palavras ao longo de determinado *corpus*.

⁵ No entanto, permite expansões incrementais ao selecionar documentos ou entidades de interesse.

o *layout* da proposta deste trabalho (GORG et al., 2007a), e, por fim, a dissertação de Bruno Schneider que também incorpora a temática de visualização de tópicos, mas sob ótica de seu fluxo ao longo do tempo (SCHNEIDER, 2014).

O presente trabalho está dividido da seguinte forma: na parte I, será feita uma revisão de modelagem de tópicos (Capítulo 1) e de trabalhos relacionados à área de visualização de tópicos (Capítulo 2). Na parte II, será apresentada a preparação dos dados: ferramentas e tecnologias aplicadas, base de dados utilizada, construção de metadados por meio de modelagem de tópicos e extração de entidades, assim como tratamentos e transformações dos dados para o desenvolvimento da visualização. Por fim, a parte III trata da seleção do modelo de visualização e está encarregada de abordar os resultados, tratando também de comparações com outros modelos de visualização. Nas partes II e III, há o princípio de seguir um fluxograma⁶ de visualização de acordo com a imagem a seguir:

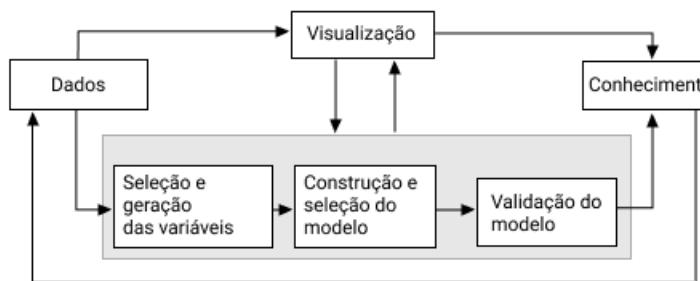


Figura 1 – Representação de fluxograma de visualização proposto por (WANG et al., 2016).

Tal modelo de fluxograma ilustrado na Figura 1 tem como base um modelo generalizado proposto por Xu-Meng Wang et al. (WANG et al., 2016), que, por sua vez, faz referência a modelos mais difundidos na área de visualização, mas que seguem um ordenamento similar, começando por seleção dos dados, tratamento, elaboração de modelo de visualização e finalizado com um modelo de validação ou comparação com modelos de proposta semelhante. Parte desses modelos pode ser consultada em trabalho de Daniel Keim et al. (KEIM et al., 2008), assim como em Jiawei Han et al. (HAN; PEI; KAMBER, 2011) e Ward et al. (WARD; GRINSTEIN; KEIM, 2015).

⁶ Denominei fluxograma como tradução livre do termo *Pipeline*, comum à área de visualização de dados.

Parte I

Referenciais teóricos

1 Modelagem de tópicos

Modelagem de tópicos engloba um conjunto de algoritmos cuja função é descobrir a estrutura temática de um volume comumente extenso de documentos, a partir do qual o modelo pode gerar um conjunto de tópicos interpretáveis, tendo cada documento um grau de associação aos mesmos. A adoção de tais algoritmos permite automatizar um processo que seria comumente mais custoso caso fosse realizado de modo manual, assim como, nas palavras de DiMaggio et al., “tais modelos permitem descobrir tópicos que um pesquisador não conseguiria ao usar métodos manuais” ([DIMAGGIO; NAG; BLEI, 2013](#)).

Segundo David Blei, autor do modelo LDA¹, utilizado neste trabalho, processos de modelagem de tópicos são algoritmos voltados para a busca de temas dentro de uma vasta, e, muitas vezes, não estruturada, coleção de documentos. As variáveis independentes (observáveis) são as palavras contidas nos documentos. A variável dependente (não observável) é a estrutura de tópicos. Práticas avançadas de modelagem permitem estabelecer a inter-relação entre os tópicos gerados, além de possibilitar a observação da variação dos tópicos ao longo do tempo ([BLEI, 2012](#)). A modelagem de tópicos possibilita a classificação dos documentos de acordo com o agrupamento (*clustering*) de temas gerados. Os tópicos em si são desconhecidos a priori, e, após a classificação feita pelo algoritmo, são avaliados e validados por especialistas para formular uma nomeação apropriada para cada tema criado. O processo permite a organização de arquivos de forma automatizada, que não seria possível em curto espaço de tempo se houvesse apenas trabalho manual.

Esses modelos estão baseados no conceito de que documentos seriam mistura de tópicos, onde cada tópico seria uma distribuição probabilística sobre as palavras. Por esse meio, constitui-se um modelo gerativo para documentos, ou seja, ele especifica um processo probabilístico pelo qual documentos podem ser gerados. Esse modelo gerativo é baseado em regras simples de amostragem que descrevem como palavras podem ser geradas em documentos com base em variáveis latentes. Ao ajustar um modelo gerativo, o objetivo é, ao assumir que o mesmo de fato gerou os dados, encontrar o conjunto de variáveis latentes mais verossímeis para a explicação dos dados observados, isto é, as palavras observadas nos documentos ([STEVVERS; GRIFFITHS, 2007](#)). Entre os modelos aplicados para agrupamento de documentos se encontram, por exemplo, *Latent Semantic Indexing* (LSI)², *Probabilistic Latent Semantic Indexing* (pLSI) e *Latent Dirichlet Allocation* (LDA), além de modelos que promovem adaptações ao LDA e que serão brevemente expostos ao final desta seção, tais como *Correlated Topic Models*, *Dynamic Topic Models*, *Bigram*

¹ O modelo LDA foi elaborado por David Blei, Michael Jordan e Andrew Ng. Para mais detalhes do modelo LDA, ver o artigo original ([BLEI; NG; JORDAN, 2003](#)).

² Latent Semantic Indexing também é comumente conhecido como *Latent Semantic Analysis* (LSA).

Topic Models e Hierarchical Dirichlet Process.

O LSI foi desenvolvido em 1990 com o intuito de automatizar e aperfeiçoar procedimentos de indexação e recuperação de informação, fazendo uso da estrutura de associação de termos e documentos, ou seja, da estrutura semântica dos elementos textuais. Uma abordagem concisa do LSI consiste no fato de que a agregação dos contextos em que uma palavra aparece fornece informação relevante que ajuda a determinar a similaridade (proximidade) de palavras e conjuntos de palavras. Deve-se ressaltar que a não ocorrência das palavras em documentos é tão importante quanto a ocorrência positiva e é daí que se estabelece uma distinção significativa frente a métodos que consideram simplesmente a coocorrência das mesmas (LANDAUER; FOLTZ; LAHAM, 1998). Assume-se que existe uma estrutura semântica latente nos dados que é obscurecida pela aleatoriedade com que as palavras são escolhidas e alocadas num texto. O LSI está alicerçado em três considerações: a informação semântica pode ser derivada por meio de uma matriz de coocorrência de palavras e documentos, a redução de dimensionalidade é uma parte essencial dessa derivação, e, por fim, palavras e documentos podem ser representados como pontos de um espaço euclidiano (STEYVERS; GRIFFITHS, 2007).

Inicialmente, o *corpus* é representado em uma matriz onde as linhas compreendem palavras únicas enquanto que as colunas compreendem documentos (ou outro formato de passagem de texto, a depender do *corpus*). Cada uma de suas células contém a frequência de cada palavra (relativa à respectiva linha) quanto a cada passagem (relativa à respectiva coluna). Posteriormente, é feita aplicação de Decomposição em Valores Singulares (SVD³). Pelo modelo SVD, qualquer matriz retangular X, por exemplo, de tamanho $t \times d$, sendo suas dimensões representadas pelo número de termos e documentos, pode ser decomposta no produto de três outras matrizes:

$$X = T_0 S_0 D'_0$$

Analizando as matrizes, T_0 e D_0 possuem colunas ortonormais e formam os vetores singulares, enquanto que S_0 é uma matriz diagonal e forma os valores singulares. Todos os elementos de S_0 são positivos e ordenados de modo decrescente (DEERWESTER et al., 1990).

Por meio da fatoração, é obtida uma matriz de termos e documentos e é construído um espaço “semântico” onde os elementos que se encontram comumente associados são alocados próximos um do outro. O SVD permite o arranjo do espaço de modo a refletir os padrões mais associativos ao mesmo tempo em que ignora padrões de menor influência. Por esta razão, ainda que um termo não ocorra num documento, o mesmo pode se posicionar

³ *Singular Value Decomposition* (SVD) é uma fatoração de matriz complexa em três matrizes distintas (com as matrizes resultantes compostas por vetores singulares e valores singulares) que permite derivar a estrutura latente do modelo, representando termos e documentos como vetores em espaço multi-dimensional. A similaridade entre os vetores é dada pelo produto escalar ou cosseno (DEERWESTER et al., 1990).

próximo ao mesmo na matriz se isso estiver consistente com os padrões de associação dos dados. Isso permite conceber modelos de proximidade, que buscam manter proximidade entre elementos similares, dado um espaço multidimensional (DEERWESTER et al., 1990).

LSI ressalta a importância da escolha da dimensionalidade em que palavras e documentos são representados e assume que a redução de dimensionalidade⁴ dos dados observados, obtida por meio do SVD, frequentemente gerará uma aproximação adequada. Essa redução pode ser feita por meio da eliminação dos coeficientes da matriz diagonal resultante do SVD, começando pelo menor (LANDAUER; FOLTZ; LAHAM, 1998). Em geral, como muitos desses componentes possuem valor baixo, eles podem ser ignorados, o que permite a obtenção de um modelo aproximado constituído por um número bastante reduzido de componentes.

PLSA se diferencia do LSI por possuir sólida fundamentação estatística, por ser baseado no princípio de verossimilhança e por definir um modelo generativo apropriado dos dados. Esse último fator traz vantagens diversas: implica que métodos estatísticos podem ser aplicados para ajuste e seleção do modelo, assim como para controle de complexidade. Por exemplo, pode-se acessar a qualidade de um modelo PLSA por meio da medição da capacidade preditiva (HOFMANN, 2013). PLSA é construído inicialmente pelo que Hofmann denomina de *aspect model*, que seria um modelo de coocorrência dos dados que associam uma classe de variáveis não observada $z_k \in \{z_1, \dots, z_K\}$ a cada observação, nesse caso, a ocorrência de palavras em cada documento. Um modelo generativo de coocorrência de palavras e documentos pode ser definido da seguinte forma:

1. Seleção de um documento d_i com probabilidade $P(d_i)$
2. Escolha de uma classe latente z_k com probabilidade $P(z_k|d_i)$
3. Geração da palavra w_j com probabilidade $P(w_j|z_k)$

Isso permite construir o par de observações (d_i, w_j) , enquanto que a classe z_k pode ser descartada. Esse processo generativo pode ser reescrito em uma probabilidade conjunta ao somar todas as possíveis escolhas de z_k pela qual uma observação poderia ter sido gerada, o que resulta nas equações a seguir:

$$P(d_i, w_j) = P(d_i)P(w_j|d_i), \quad P(w_j|d_i) = \sum P(w_j|z_k)P(z_k|d_i)$$

A principal diferença entre PLSA e LSI está na função objetivo utilizada para a aproximação ótima que, no caso do LSI, seria a norma matricial. Quanto ao PLSA, é feito por meio da função de verossimilhança com o intuito de maximizar o poder preditivo do modelo. Outra diferença é que o espaço latente do PLSA é interpretável, enquanto

⁴ A redução de dimensionalidade é comumente feita para um número ainda considerável.

que o LSI não oferece esse tipo de interpretabilidade ([HOFMANN, 2001](#)). Apesar dessas vantagens, o PLSA apresenta problemas tal como a falta de parâmetros para $P(d)$, o que impossibilita que se assuma probabilidades para um novo documento. Esse problema é resolvido pelo modelo LDA, que generaliza o PLSA ao transformar o d , antes fixo, numa priori de *Dirichlet*.

O LDA requer explicação mais aprofundada devido ao fato de ter sido o modelo aplicado neste trabalho. Como aplicação de modelagem de tópicos, o LDA é um método estatístico que formula a modelagem de modo não-supervisionado, ou seja, os tópicos são definidos automaticamente pelo algoritmo, havendo interferência humana apenas para o estabelecimento de hiper-parâmetros, e na etapa de pré-processamento, tal como para remoção de termos sem importância semântica⁵, criação de filtros de modo a retirar documentos sem pertinência para a análise e em procedimentos. Os tópicos são construídos usando-se a frequência de palavras, suas correlações nos documentos e a relevância (grau de frequência) do tópico no *corpus*. A modelagem é construída, portanto, segundo a distribuição probabilística de tópicos nos documentos, de acordo com um *score* que representa a probabilidade de associação entre tópicos e documentos.

Fazendo uso de uma abordagem mais técnica, a intuição por trás do LDA seria a de que documentos tendem a representar diversos tópicos. Por exemplo, um artigo que aborda análise de dados na área de biologia pode ter diversos temas, que podem ser identificados por combinações de palavras, tais como “computador” e “predição” para um tópico ligado a análise de dados, ou “vida” e “organismo” para um tópico da área de biologia evolucionária, ou ainda “sequenciamento” e “genes” para um tópico relacionado a genética. Ou seja, um mesmo documento trataria nesse caso de três tópicos ([BLEI, 2012](#)).

Modelos como LDA buscam capturar essa intuição ao simular o processo aleatório em que os documentos foram gerados. Assume-se que esses tópicos foram especificados antes da própria captura dos dados. Buscando-se uma explicação mais detalhada: para cada documento na coleção, são geradas as palavras por meio do seguinte processo ([BLEI, 2012](#)):

1. É escolhida aleatoriamente uma distribuição sobre os tópicos.
2. Para cada palavra no documento:
 - a) É escolhido aleatoriamente um tópico a partir da distribuição de tópicos construída no passo 1.
 - b) É escolhida aleatoriamente uma palavra a partir da distribuição correspondente sobre o vocabulário.

⁵ *Stopwords*, por exemplo, que serão explicadas em seção dos passos metodológicos.

Cada documento contém tópicos em diferentes proporções (passo 1), cada palavra em cada documento é obtida a partir de um dos tópicos (passo 2b). O tópico, por sua vez, é escolhido a partir da distribuição de documentos por tópicos (passo 2a).

LDA e outros modelos de tópicos fazem parte da área de modelagem probabilística, segundo os quais os dados são vistos como provenientes de um processo gerativo que engloba variáveis ocultas (*hidden variables*). Tal processo define uma distribuição de probabilidade conjunta quanto às variáveis aleatórias, tanto as observadas quanto às ocultas. Essa distribuição, por sua vez, por meio das variáveis observadas, é utilizada para obter a distribuição condicional das variáveis ocultas, que é denominada de distribuição *a posteriori*⁶.

Sob o modelo LDA, assume-se que os documentos seriam variáveis observadas, enquanto que a estrutura de tópicos, ou seja, a distribuição de documentos e palavras sobre os tópicos, seria uma estrutura oculta (*hidden structure*). Em termos computacionais, a tarefa principal seria encontrar essa estrutura oculta a partir das variáveis observadas (documentos), o que pode ser compreendido como reversão do processo gerativo. Sua resolução passaria pela computação da distribuição *a posteriori*, ou seja, da distribuição condicional das variáveis ocultas dados os documentos que compõem o *corpus*.

O LDA pode ser descrito fazendo uso da seguinte notação (BLEI, 2012): Os tópicos são $\beta_{1:k}$, onde cada β_k é uma distribuição sobre o vocabulário. As proporções de tópicos para o d -ésimo documento são θ_d ⁷, onde $\theta_{d,k}$ é a proporção do tópico k para o documento d . As atribuições de tópicos para o d -ésimo documento são z_d , onde $z_{d,n}$ é a atribuição do tópico para a n -ésima palavra no documento d . Por fim, as palavras observadas relativas ao documento d são w_d , onde $w_{d,n}$ é a n -ésima palavra do documento d , sendo esta um elemento do vocabulário fixo. Com essa notação, o processo gerativo correspondente à distribuição conjunta das variáveis ocultas e observadas pode ser descrito pela fórmula a seguir:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \\ \left(\prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:k}, z_{d,n}) \right)$$

Note-se a dependência dos elementos na fórmula acima. A atribuição de tópico $z_{d,n}$ depende da proporção do tópico para o documento observado θ_d . A palavra observada $w_{d,n}$ depende da atribuição de tópico $z_{d,n}$ e de todos os tópicos $\beta_{1:k}$. Cada tópico K define uma distribuição multinomial sobre o vocabulário e assume-se que foi obtida a partir de

⁶ A probabilidade a posteriori é a probabilidade dos parâmetros θ dada a evidência X : $p(\theta|X)$. Sobre inferência bayesiana, ver (MIGON; GAMERMAN; LOUZADA, 2014)

⁷ $\theta_d \sim Dirichlet(\alpha)$ (HOFFMAN; BACH; BLEI, 2010).

uma distribuição de Dirichlet, $\beta_k \sim Dirichlet(\eta)$ (HOFFMAN; BACH; BLEI, 2010). Tais relações também podem ser demonstradas por meio de um modelo probabilístico gráfico (*probabilistic graphical model*), ilustrado na Figura 2. Além disso, Blei et al. faz uso de exemplos com o objetivo de demonstrar visualmente a intuição do modelo (ver Figuras 3 e 4).

A *posteriori*, explicada mais acima, é dada pela fórmula a seguir:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

Onde o numerador é a distribuição conjunta de todas as variáveis aleatórias, enquanto que o denominador é a probabilidade marginal das observações. A computação da estrutura de tópicos pode ser vista como intratável⁸ e requer métodos de aproximação, compondo dois tipos distintos de algoritmos: os baseados em amostragem e os variacionais. O primeiro tem como principal modelo a amostragem de Gibbs, onde se constrói uma cadeia de Markov Monte Carlo⁹ (MCMC), definida nas variáveis ocultas (tópicos) sobre determinado *corpus* (GRIFFITHS et al., 2005). Métodos variacionais, por outro lado, são métodos de otimização que assumem uma família de distribuições sobre a estrutura oculta e encontra a distribuição que mais se aproxima da *posteriori* (HOFFMAN; BACH; BLEI, 2010).

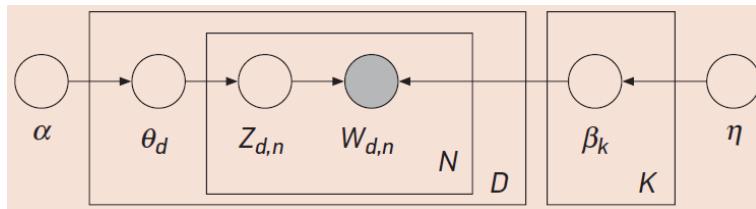


Figura 2 – O modelo probabilístico gráfico do LDA permite demonstrar visualmente as dependências entre as variáveis aleatórias (BLEI, 2012). Segue explicação dos parâmetros introduzidos nesta Figura: N denota o conjunto de palavras em um documento, D representa a coleção de documentos no *corpus*, K seria o conjunto de tópicos, $\theta_d \sim Dir(\alpha)$ e $\beta_k \sim Dir(\eta)$.

É preciso ressaltar também três principais pressupostos de modelos como o LDA. O primeiro, definido como “bag of words”, determina que a ordem das palavras não é relevante, o que é razoável dado os objetivos da modelagem de tópicos. O segundo afirma que a ordem dos documentos não importa, o que tende a ser problemático no caso de

⁸ Considerando que coleções de documentos costumam conter milhões de palavras observadas.

⁹ MCMC engloba um conjunto de técnicas iterativas de aproximação voltadas para a geração de amostras de variáveis aleatórias a partir de distribuições complexas (comumente multidimensionais). Amostragem de Gibbs é um tipo específico de MCMC que simula uma distribuição multidimensional por meio de amostragem em subconjuntos de menor dimensionalidade de variáveis em que cada subconjunto está condicionado no valor de todos os outros. A amostragem é feita de modo sequencial até o momento em que os valores amostrados se aproximam da distribuição almejada (STEYVERS; GRIFFITHS, 2007)

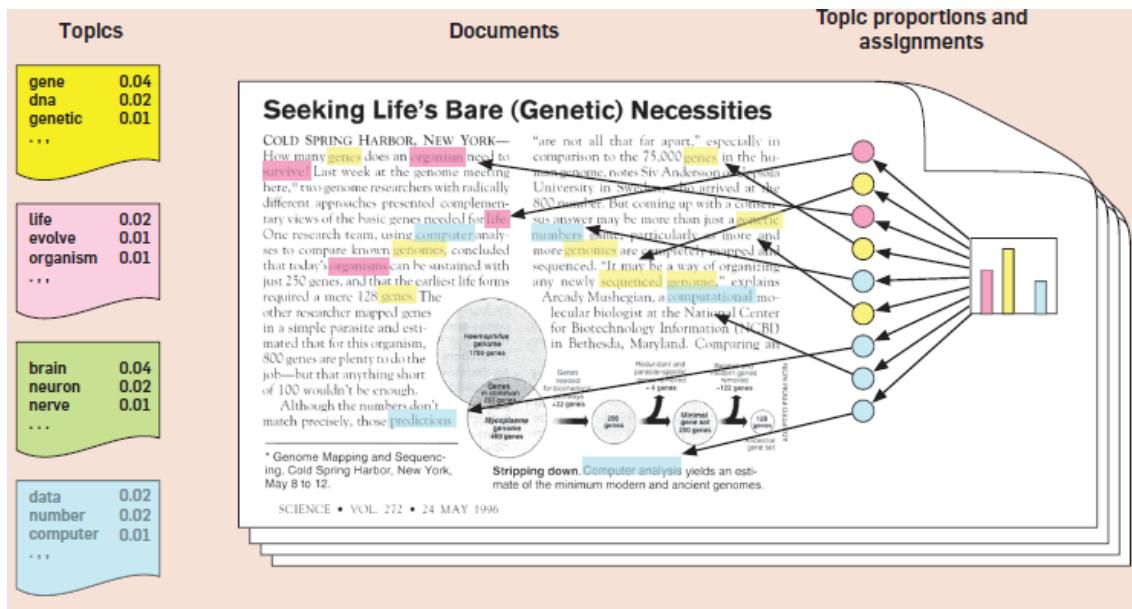


Figura 3 – Exemplo ilustrativo que busca demonstrar a intuição do LDA. Os tópicos como distribuições de palavras ($\beta_{1:k}$) estão na parte esquerda da Figura, as proporções de tópicos para os documentos (θ_d) compõem o histograma à direita, as atribuições de tópicos para o d -ésimo documento (z_d) são representadas pelos círculos coloridos próximos ao histograma. Finalmente, as palavras observadas no documento (w_d) se encontram realçadas no meio dessa Figura (BLEI, 2012).

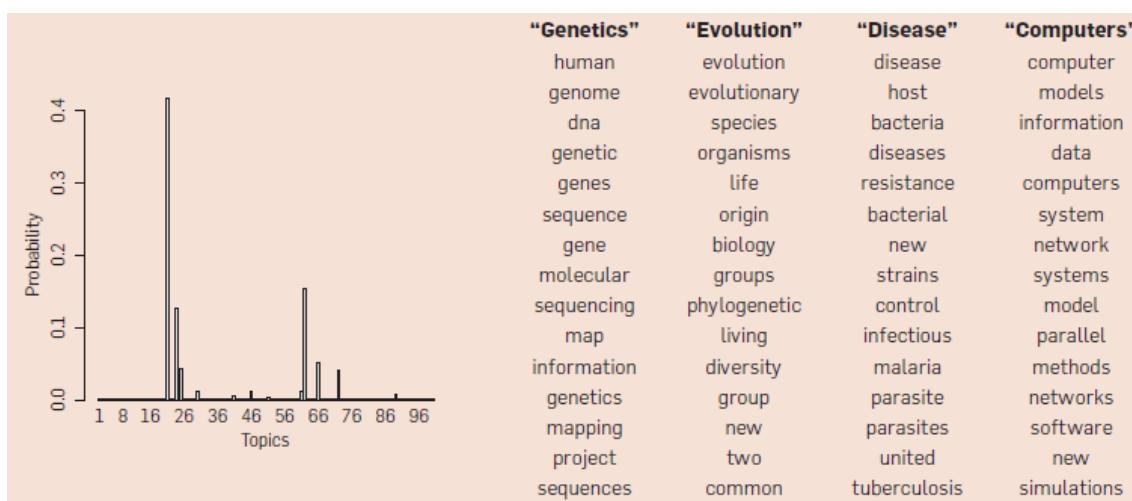


Figura 4 – Aplicação do LDA com estabelecimento de 100 tópicos sobre um *corpus* constituído por 17 mil artigos científicos. Ao lado esquerdo se encontram as proporções de tópicos obtidas por meio do mesmo artigo da Figura 3, enquanto que ao lado direito estão os 15 termos de maior frequência no caso dos 4 tópicos mais relacionados a esse artigo (BLEI, 2012).

coleções que abrangem longos períodos de tempo. Por fim, o LDA assume que o número de tópicos seria determinado e fixo.

Esses pressupostos podem representar também limitações, que se somam a outras de alta relevância, tal como ausência de capacidade de modelar correlações entre tópicos, o que seria algo esperado em casos como o de artigos científicos, tal como o exemplo que foi feito no início deste Capítulo no caso de artigo da área de biologia, em que documentos poderiam fazer referência a tópicos como genética e biologia evolucionária, mas dificilmente poderia incluir um tópico relacionado a astronomia. Essa limitação advém de um dos pressupostos do modelo, no caso, o de independência, que estaria implícito na distribuição de Dirichlet das proporções de tópicos, o que tende a ser inverossímil em muitos casos. Existem atualmente modelos que solucionam tal questão, tal como o *Correlated Topic Model* (CTM), que flexibiliza a distribuição para a proporção de tópicos. (BLEI; LAFFERTY, 2006a).

No caso de outras limitações, também há soluções atualmente desenvolvidas. Por exemplo, quanto à segunda premissa, que ignora a evolução dos documentos no caso de corpora com alta amplitude de tempo, foi desenvolvida uma modelagem de tópicos dinâmica (*Dynamic Topic Model*) (BLEI; LAFFERTY, 2006b). No caso da premissa da “bag of words”, existem modelos que a relaxam e assumem que tópicos geram palavras condicionado à palavra anterior (*Bigram Topic Model*) (WALLACH, 2006). Para a terceira premissa, existem modelos como o *Hierarchical Dirichlet Process*¹⁰ (HDP), que permite elaborar modelagem de tópicos sem a necessidade de determinar previamente o número de tópicos (TEH et al., 2005). Há inclusive atualmente o modelo *Supervised Topic Models* (sLDA), modelo supervisionado proposto para ser usado quando os documentos já possuem categorização prévia (MCAULIFFE; BLEI, 2008).

1.1 Exemplos de aplicação

Existe atualmente uma literatura dedicada a aplicações de modelagem de tópico nas áreas de ciência política, relações internacionais e história. Christopher Lucas (2015) aplica a técnica para formular análise de política comparada, fazendo uso de Structural Topic Model (STM), com a finalidade de automatizar análises de conteúdo textual. A aplicação foi voltada para estudos sobre textos, principalmente fatwas , de modo a encontrar diferenças de padrões de escrita entre jihadistas e não-jihadistas, reunindo um *corpus* com mais de 27 mil textos (LUCAS et al., 2015). Stewart e Zhukov (2009), por sua vez, fazem uso de análise automática de conteúdo e analisam uma base de dados de 8 mil discursos públicos de autoridades militares e políticas da Rússia durante o governo de Vladimir

¹⁰ O modelo HDP é uma generalização do LDA e tem como principal diferencial o fato de permitir que o número total de tópicos seja determinado por meio do próprio processamento em vez de ser estabelecido manualmente (TEH et al., 2005; BEAL; GHAHRAMANI; RASMUSSEN, 2002).

Putin com a finalidade de avaliar a influência da elite militar do país sobre sua política externa (STEWART; ZHUKOV, 2009). Estes são somente alguns exemplos de aplicações de tecnologias que permitem a automatização de análise de grandes volumes de textos. Além desses trabalhos, podemos mencionar também Margaret Roberts et al. (ROBERTS; STEWART; AIROLDI, 2016), que analisa uma coleção de reportagens sobre a China e avalia a evolução temporal dos tópicos construídos.

Trabalhos de classificação de tópicos voltados a documentos históricos devem seguir determinados critérios para facilitar o processo de busca de informação e atender devidamente seus usuários (cientistas sociais, cientistas políticos, historiadores e outros especialistas). Segundo Dustin Hillard et al. (HILLARD; PURPURA; WILKERSON, 2008), os principais critérios são: discriminação, acurácia, confiabilidade, probabilidade e eficiência. Para haver discriminação, os tópicos devem ser mutuamente excludentes. Quanto à acurácia, é necessário que cada tópico indique o conteúdo dos documentos, além de ser necessário estabelecer métricas para mensurar essa acurácia. Já o critério de confiabilidade está relacionado à heterogeneidade dos arquivos presentes em um *corpus*, que ocorre principalmente se a amplitude temporal do *corpus* for longa (acima de duas décadas). Segundo esse critério, documentos similares devem ser classificados de modo similar, mesmo que se encontrem em períodos distintos e mesmo que a terminologia usada se altere ao longo desse período. O autor cita como exemplo o tema “direitos humanos”, conceito cuja terminologia passou por mudanças ao longo de décadas, mas preservando o leque de assuntos tratados. O critério de probabilidade, por sua vez, estabelece que, modelos de classificação de tópicos, além de discriminarem o tópico principal de cada documento, devem identificar probabilisticamente outros tópicos relacionados. Finalmente, não menos importante, o critério da eficiência estabelece que, quanto menos custosa for a adoção de um processo de classificação, maior o seu valor.

Trabalhos manuais de classificação permitem alcançar discriminação, acurácia e confiabilidade. No entanto, para discriminar hierarquicamente o grau de pertencimento de cada documento a um tópico, é preciso recorrer a ferramentas computacionais, tanto para gerenciar bancos de dados quanto para fazer essa identificação, que não é exequível apenas por trabalho humano. Quanto à eficiência, dado o aumento crescente do volume de dados produzidos, a classificação manual de documentos tem se tornado cada vez mais custosa, o que também demanda a adoção de soluções alternativas (HILLARD; PURPURA; WILKERSON, 2008).

2 Interpretabilidade de tópicos gerados

Segundo Wang et al., após o pré-processamento e tratamento dos dados, cabe ao responsável pela sua análise a decisão entre aplicar métodos de análise automática ou recorrer a soluções semi-automáticas que envolvam métodos como o de visualização da informação (WANG et al., 2016). A avaliação de modelos de tópicos é uma tarefa complexa que demanda a assistência de especialista de domínio, incumbido da tarefa de analisar uma amostragem de documentos dado o respectivo *score* com cada tópico. Após a análise, tópicos podem ser nomeados, agregados ou descartados. Além disso, métricas podem ser utilizadas, como o percentual de documentos que condiz com um tópico, dado o estabelecimento do nome.

Um desafio relevante ligado a tópicos gerados por modelos como o LDA seria o de interpretabilidade por humanos. Segundo AlSumait et al (ALSUMAIT et al., 2009), nem todos os tópicos gerados possuem a mesma importância ou correspondem a temas concretos. Na verdade, muitos deles podem representar mera coleção de palavras de fundo ou cabeçalho ou até mesmo temas de baixa relevância para o contexto do pesquisador que aplicou o modelo¹. A baixa interpretabilidade pode trazer algumas dificuldades para alguns procedimentos da modelagem, tal como a seleção do número de tópicos a ser gerado. Seleção do número de tópicos muito abaixo ou acima do nível ótimo tende a gerar um ciclo vicioso de baixa interpretabilidade dos resultados para esse procedimento. Sabe-se que uma seleção excessiva de tópicos tende a produzir tópicos demasiadamente específicos (por vezes com tópicos que compartilham detalhes de um mesmo tema) e de baixa interpretabilidade, enquanto que uma seleção muito parcimoniosa de tópicos tende a gerar tópicos de definições altamente abrangentes, com possibilidade de incorporar múltiplos temas em um, também ocasionando em baixa interpretabilidade do mesmo.

2.1 Métodos de análise automática

Para Chang et al, métodos quantitativos conhecidos não necessariamente correspondem com medidas de interpretabilidade de tópicos feitas por humanos (CHANG et al., 2009). Entre esses métodos, podemos citar AlSumait et al., que propõe uma análise não-supervisionada de tópicos de modo a distinguir tópicos relevantes de tópicos te temas não significativos ou simplesmente caracterizados por ruído nos dados. O trabalho se

¹ Por exemplo, no contexto do trabalho desenvolvido junto ao CPDOC, a primeira versão da modelagem não filtrava idiomas e terminava por agregar tópicos de acordo com idiomas específicos, o que não interessava o projeto.

caracteriza por utilizar medidas de distância² frente a tópicos enquadrados, por meio de três métricas distintas, como não significativos (ALSUMAIT et al., 2009).

Outro trabalho (MIMNO et al., 2011) mensura coerência de tópico de acordo com a coocorrência ao longo do *corpus*. Por fim, Chuang introduz um modelo que permite acessar em grande escala a relevância de tópicos ao quantificar o alinhamento entre um conjunto de tópicos latentes e um conjunto de referências conceituais. No caso, um tópico remeteria a um conceito se houver correspondência de um para um entre ambos (CHUANG et al., 2013). Tal trabalho foi posteriormente aprimorado com a criação de uma ferramenta interativa, o *TopicCheck*, que permite acessar a estabilidade de um modelo de tópicos ao inspecionar se o mesmo consistentemente cobre um conjunto estável de conceitos (CHUANG et al., 2015).

Além desses métodos, podemos mencionar um de maior conhecimento na área de Modelagem de tópicos tal como o *Pointwise Mutual Information* (PMI), que faz ranqueamento de termos contidos em cada tópico, um método semelhante ao de Taddy, que usa uma medida chamada de *lift*, que estabelece uma razão entre a probabilidade de um termo dentro de um tópico e sua probabilidade marginal ao longo de todo o *corpus* (TADDY, 2012). Com um intuito semelhante, Bischof e Airoldi, por sua vez, desenvolvem uma proposta distinta de modelagem de tópicos que permite inferir tanto a frequência de um termo quanto a sua exclusividade, ou seja, em que medida ele ocorre em um grupo reduzido de tópicos (BISCHOF; AIROLDI, 2012).

Tais métodos mencionados podem ser úteis para compreender em que medida tópicos específicos seriam interpretáveis. Entretanto, não trazem uma solução ao usuário no sentido de ajudá-lo a interpretar tópicos individualmente. É nesse sentido que foram desenvolvidas soluções de visualização, que serão abordadas na seção 2.2. Certamente, com visualizações, almeja-se não a substituição de um tipo de solução por outro, mas busca-se complementariedade entre metodologias distintas.

2.2 Visualização de dados para interpretação de tópicos

Trabalhos de visualização que envolvem dados textuais podem ser parcialmente sintetizados por Matthew Ward (WARD; GRINSTEIN; KEIM, 2015), tais como nuvens de palavras e *TextArcs*³. Um resumo mais recente de métodos relacionados a dados textuais pode também ser encontrado em Shixia Liu et al (LIU et al., 2018). Segundo David Blei, existem algumas caminhos de aperfeiçoamento de modelagem de tópicos, tais como nos campos de validação, de visualização e de análise exploratória dos dados. Modelagem de tópicos tem o potencial de oferecer novas formas de exploração dos dados em coleções

² No caso, foram três medidas: coeficiente de correlação, dissimilaridade pelo cosseno e divergência de Kullback-Leibler

³ Eficaz para mostrar sequências repetidas de palavras ao longo de determinado *corpus*.

volumosas de documentos. A junção com visualização de dados tenderia a amplificar essa capacidade. No entanto, algumas questões emergem quanto a formas de apresentação da informação. Por exemplo, é comum que tópicos sejam apresentados com uma listagem de palavras ordenadas pela respectiva frequência nos mesmos, mas podem haver outras medidas relevantes para o ordenamento. Além disso, existem diferentes formas de exposição de documentos que podem assistir à sintetização das informações presentes nos mesmos ([BLEI, 2012](#)).

Windhager et al. descreve categorizações que ajudam a diferenciar as ferramentas de visualização voltadas a patrimônio histórico e cultural (o que inclui arquivos documentais tal como a base aqui utilizada). Entre elas, se encontram duas que serão enfatizadas nesta dissertação: usuários e granularidade. A primeira se refere a experiência prévia, grau de conhecimento e interesses do usuário, o que influenciaria as expectativas e interações com a interface visual. Seriam duas as classes de usuário: especialista e usuário casual (pessoa do público geral). Enquanto que o especialista possui conhecimento de domínio e tem interesse profissional ou científico, o que facilita sua navegação sobre a coleção, o público geral busca informações de interesse pessoal. Quanto à segunda forma de categorização, a granularidade (ilustrado na Figura 5), permite-se diferenciar métodos de acordo com o tipo de perspectiva trazida para a coleção, o que pode ser subdividido em: *Single Object Previews*, que provê representação detalhada dos objetos, *Multi-Object Previews*, que provê representação de uma seleção de objetos, e, finalmente, *Collection Overviews*, que permite uma visão geral da coleção. A visão geral pode ser promovida tanto por meio de transformações visuais discretas tais como glifos, quanto pela utilização de abstrações geométricas de modo a representar algum padrão dos dados ([WINDHAGER et al., 2018](#)).

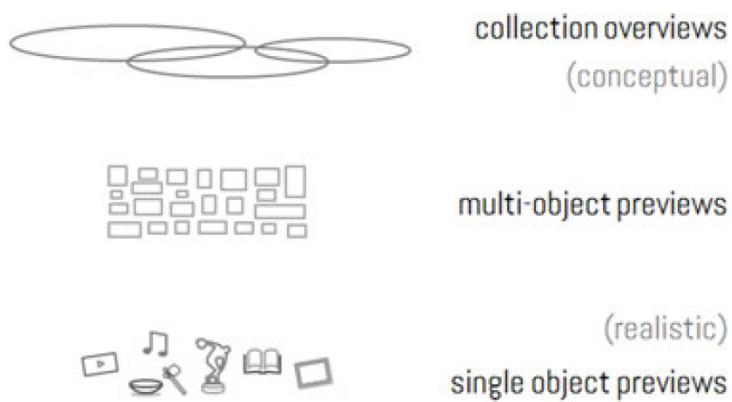


Figura 5 – Conceito de granularidade abordado por Windhager et al.. Em cada camada, há determinado grau de distância visual da coleção, sendo que na camada inferior, há maior nível de detalhamento de cada objeto ([WINDHAGER et al., 2018](#)).

Nos últimos anos, diversas soluções de visualização voltadas para modelagem de tópicos foram desenvolvidas, sendo que grande parcela delas possibilita interação por parte

do usuário com documentos, tópicos e termos, de modo a aumentar sua compreensão sobre relações entre estes elementos. O trabalho de Bruno Schneider, que pode ser considerado essencial fonte de inspiração para o desenvolvimento desta dissertação, traz diversos exemplos de visualização de tópicos sob a ótica de seu fluxo ao longo do tempo até alcançar sua proposta final (Figura 24 do anexo A) (SCHNEIDER, 2014). Entre as ferramentas de visualização analisadas, localizadas no anexo A, estão: *The Topic Browser*⁴ (Figura 25 do anexo A) (GARDNER et al., 2010), *MetaToMATo* (Metadata and Topic Model Analysis Toolkit, Figura 26 do anexo A) (SNYDER et al., 2013) e um navegador desenvolvido por David Blei e testado na Wikipedia (Figura 27 do anexo A) (CHANAY; BLEI, 2012). Essas ferramentas costumam usar listas de termos ou documentos mais prováveis dentro de cada tópico para summarizá-los e as visualizações se apresentam como gráficos de barras, gráficos de dispersão ou nuvens de palavras de acordo com alguma métrica, por exemplo, a probabilidade dos termos para cada tópico ou a probabilidade dos tópicos para cada documento. A interação do usuário com essas ferramentas permite acesso ao *corpus* como um todo, inclusive, como para o caso do *The Topic Browser*, a hiperlinks que fornecem acesso ao corpo de texto original.

Outra proposta de visualização, que pode ser considerada como uma das inspirações para o *layout* da proposta deste trabalho, chamada de *Jigsaw* (Figura 28 do anexo A), é uma ferramenta interativa voltada para visualização e exploração de um *corpus* que permite realçar um documento e descobrir relações com entidades presentes no mesmo. O *Jigsaw* possui seis formas distintas de visualizar a informação de um *corpus*: *List View*, *Graph View*, *Scatter Plot View*, *Text View*, *Time Line View* e *Calendar View*⁵. O mais relevante para este trabalho é o *Graph View* (visualizador de grafo). A estrutura desse visualizador é apresentada como um grafo de inter-relações, fornecendo apenas uma amostragem da base de dados original⁶ (GORG et al., 2007a).

Os autores da ferramenta reconhecem que o *Jigsaw* é mais efetivo na descoberta de conexões entre documentos do que na identificação de temas. A ferramenta oferece alguns filtros. Por exemplo, fornece a opção de remover da análise as entidades que aparecem somente em um documento, o que permite focalizar a análise nas entidades de maior grau de conexão. A interatividade fornece a possibilidade de selecionar uma entidade presente no gráfico gerado e expandir informações relativas a documentos em que a mesma é mencionada. Os documentos por sua vez, ao serem selecionados, realçam as entidades que menciona. Como a expansão dos elementos visuais pode ser muito extensa, há opções de *zoom in* e *zoom out* no gráfico. Os outros visualizadores fornecem uma imersão nos documentos. O *Text View*, por exemplo, gera visualizações dos próprios documentos. A

⁴ Tal ferramenta pode ser acessada no seguinte endereço eletrônico: <<https://agoldst.github.io/dfr-browser/demo/>>

⁵ Demonstrações do uso dos visualizadores está presente no endereço eletrônico a seguir: <<https://www.cc.gatech.edu/gvu/ii/jigsaw/tutorial/>>

⁶ No entanto, permite expansões incrementais ao selecionar documentos ou entidades de interesse.

combinação dos visualizadores tende a trazer um ganho de compreensão da inter-relação dos documentos (GORG et al., 2007b).

Apesar da interatividade do *Jigsaw*, há algumas lacunas presentes no mesmo. Para continuar expandindo entidades e documentos, o gráfico acaba por ser limitado pelo espaço de tela⁷. Essa limitação do espaço também se torna evidente no caso de coleções de documentos muito densas (com grande quantidade de documentos e entidades), o que pode gerar excesso de poluição visual no gráfico, vindo a comprometer a interpretabilidade. Além disso, a identificação de temas⁸, como dito antes, não é possível, dado que a interatividade e o *layout* são voltados para a exploração das inter-relações entre documentos e entidades.

Há outras ferramentas que agregam uma parte dos dados, ao selecionar os termos que serão apresentados aos usuários com o intuito de assisti-los na interpretação dos resultados, dado métricas sobre o termo para o respectivo tópico, tal como relevância. Uma dessas ferramentas, e também considerada mais uma inspiração para esta dissertação, conhecida como LDAvis (Figura 29 do anexo A), foi desenvolvida por Carson Sievert e Kenneth Shirley (SIEVERT; SHIRLEY, 2014), segundo os quais o ranqueamento de termos puramente de acordo com sua probabilidade de ocorrência em determinado tópico seria solução sub-ótima para a interpretação dos mesmos.

A visualização gerada pelo LDAvis possui dois elementos centrais situados à esquerda e à direita do gráfico. O painel esquerdo expõe a distância entre tópicos (quão distintos ou relacionados eles são) e a prevalência dos tópicos. A distância é gerada após aplicação de PCA (Principal Component Analysis)⁹ e possibilita visualizar a distância entre os tópicos em espaço bidimensional. A prevalência é exposta por meio dos tamanhos das áreas dos círculos que representam os tópicos. O painel direito apresenta termos individuais que seriam considerados de maior utilidade para a interpretação de cada tópico ao ser selecionado no campo esquerdo. O *layout* nesse caso é um gráfico de barras, onde cada barra representa um termo. As barras são empilhadas para expor duas variáveis relativas aos termos, que representam a frequência ao longo do *corpus* (barra azul) e a frequência ao longo do tópico (barra vermelha). Os dois painéis são interligados interativamente de modo que ao selecionar um tópico no lado esquerdo, são revelados os termos correspondentes do lado direito, enquanto que a seleção de um desses termos designa, do lado esquerdo, sua distribuição condicional em cada tópico.

A forma usual de interpretação dos termos é baseada no ordenamento dos mesmos de acordo com a ordem de probabilidade de ocorrência no tópico, dado a distribuição multinomial dos tópicos sobre os termos do *corpus*, o que se distingue do modelo proposto

⁷ De fato, na demonstração apresentada pelos autores, foram utilizados quatro monitores LCD.

⁸ É importante ressaltar que esse não é um objetivo central da ferramenta.

⁹ PCA é um método estatístico que faz uma transformação linear (nesse caso, ortogonal) sobre observações e resulta na maior variância possível. Em outras palavras, é uma redução de dimensionalidade que busca o máximo de variância do resultado final da transformação linear do espaço.

pelos autores do LDAvis. Eles ressaltam um problema característico de modelos como o LDA, que é o fato de que os termos comuns ao *corpus* como um todo tendem a aparecer no topo da lista ordenada dos tópicos, o que dificulta a diferenciação dos tópicos (dado que acabam por compartilhar muitos termos em comum), assim como a correta nomeação dos mesmos. Foi nesse sentido que alguns autores, mencionados anteriormente na seção 2.1, propuseram métricas para avaliar a interpretabilidade de tópicos¹⁰. Tais métricas, entretanto, não têm o papel de assistir o especialista para o trabalho de validação dos tópicos em si (SIEVERT; SHIRLEY, 2014).

O LDAvis no caso, é um modelo de visualização que estabelece uma métrica, denominada de relevância, de modo a ordenar termos de acordo com cada tópico. A medida de relevância do termo w frente ao tópico k dado o peso do parâmetro λ (onde $0 \leq \lambda \leq 1$) é dada por:

$$r(w, k|\lambda) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log\left(\frac{\phi_{kw}}{p_w}\right)$$

onde λ denota o peso dado a probabilidade de ocorrência do termo w no tópico k relativo ao seu *lift*, medida já mencionada anteriormente (TADDY, 2012). De modo a esclarecer a intuição desta fórmula, pode-se analisar os casos extremos: ao escolher $\lambda = 1$, a ordenação ocorre unicamente de acordo com a probabilidade do termo relativa ao tópico, enquanto que no caso de $\lambda = 0$, a ordenação é dada unicamente pelo *lift*. Valores entre 0 e 1 implicam em diferentes ordenações dos termos e a recomendação, baseada em diversos testes promovidos pelos autores do LDAvis, seria de $\lambda = 0.6$. Pode-se afirmar que a solução de visualização, assim como a métrica utilizada, fazem do LDAvis uma das ferramentas mais eficazes para a função de interpretação de tópicos (SIEVERT; SHIRLEY, 2014).

Ainda sobre ferramentas de agregação, existem produtos como o *Termite* (Figura 30 do anexo A), que permite acessar um *layout* tabular de modo a comparar termos entre tópicos¹¹. O programa filtra os termos mais prováveis ou salientes¹², enquanto que o usuário tem liberdade para determinar o número de termos que serão expostos (CHUANG; MANNING; HEER, 2012). Os autores da ferramenta advogam que essa abordagem possibilita a filtragem dos tópicos denominados pelos autores de *junk topics*, que seriam aqueles cujos agrupamentos de termos seriam irrelevantes ou incoerentes, dado métricas, sugeridas pelos autores, que são aplicadas aos tópicos, como coerência (ALSUMAIT et al., 2009) ou ainda sua importância relativa (WEN; LIN, 2010). O *Termite* teria um papel mais centrado na identificação da qualidade dos tópicos do que na análise e nomeação de tópicos em si, o que mostra que visualizações podem ser dedicadas a cumprir funções distintas no campo de modelagem de tópicos.

¹⁰ (ALSUMAIT et al., 2009; MIMNO et al., 2011; CHUANG et al., 2013)

¹¹ A ferramenta pode ser acessada no seguinte endereço eletrônico: <<http://vis.stanford.edu/topic-diagnostics/model/silverStandards/>>

¹² Saliência é uma medida de verossimilhança de que um termo w foi gerado por um tópico latente T .

Outras técnicas de visualização visam atacar um ou mais problemas presentes em modelagem de tópicos. Por exemplo, o pressuposto de que a ordem dos documentos não seria importante, no caso de coleções que compreendem décadas, pode ser problemático e requer soluções que gerem aumento de interpretabilidade nesse caso, o que é comumente feito por meio de técnicas de visualização. Uma técnica considerada referência nesse quesito¹³, conhecida como ThemeRiver (Figura 31 do anexo A) (HAVRE; HETZLER; NOWELL, 2000), é um tipo de *stacked graph* que estabelece divisão por temas ou grupos de temas, sendo cada camada dessa divisão representada por uma cor distinta. Tais camadas fluem ao longo de janelas de tempo pré-determinadas, representadas no eixo X, enquanto que a força de cada tema da coleção de documentos é representada verticalmente no eixo Y. Por causa de sua estrutura intuitiva, esta solução permite que usuários identifiquem tendências e padrões em corpora de grande volume. Por outro lado, a determinação da métrica de força do tema é subjetiva e pode causar imprecisões na análise dos temas, além do fato de ser difícil distinguir aos tamanhos de camadas empilhadas quando estas possuem diferenças sutis.

A determinação do número de tópicos é outra dificuldade presente e relevante em algoritmos como o LDA. Se, por um lado, tal questão possui soluções elaboradas por meio de técnicas estatísticas¹⁴ ou até mesmo soluções relativamente recentes de modelagem de tópicos¹⁵, por outro, ainda não foram pensadas técnicas de visualização voltadas precisamente para isso. No entanto, algumas visualizações são eficazes ao permitir identificar a diferença de qualidade entre seleções de números de tópicos. Um dos casos mais notáveis disso é o LDAvis, mencionado mais acima, pelo fato de permitir uma visualização rápida e intuitiva dos tópicos gerados, assim como dos termos dominantes em cada um desses tópicos (SIEVERT; SHIRLEY, 2014).

Todas as ferramentas apresentadas nesta seção se propunham a facilitar a compreensão do resultado gerado por modelagem de tópicos de modo a atacar o problema da interpretabilidade. Pode-se dizer que tais soluções possuem complementariedade, ou seja, que sua utilização em conjunto, somado também à aplicação de soluções estatísticas, tende a gerar aumento de conhecimento sobre o *corpus* e os tópicos gerados. É nesse sentido que esta dissertação se propõe a colaborar, ao trazer mais uma ferramenta de visualização que gere ganho de interpretabilidade.

¹³ Diversos autores baseiam suas soluções de visualização nessa técnica (DOU et al., 2011; CUI et al., 2011), inclusive o trabalho de Bruno Schneider é uma aplicação de visualização da evolução de tópicos (SCHNEIDER, 2014).

¹⁴ A principal resolução envolve medidas de coerência de tópicos (MIMNO et al., 2011).

¹⁵ Por exemplo, o modelo HDP, mencionado na seção 1 deste Capítulo.

Parte II

Passos metodológicos

Para a parte de referenciais teóricos, foi importante trazer uma abordagem de algoritmos de modelagem de tópicos com o intuito de esclarecer ao leitor a dinâmica de geração desses dados, mas o cerne deste trabalho está no desenvolvimento de uma proposta de visualização voltada para interpretação de tópicos. As etapas de pré-processamento e modelagem de tópicos foram realizadas em trabalho anterior apresentado no I Congresso Internacional em Humanidades Digitais no Rio de Janeiro¹⁶ e os resultados pós-modelagem foram, portanto, aqui utilizados (RIBEIRO; MORELI; SOUZA, 2018), sendo tal trabalho relacionado a um projeto desenvolvido junto ao *History Lab*¹⁷. De qualquer modo, é importante trazer uma descrição detalhada de todo o processo desde a coleta dos dados brutos (*raw data*) de forma a garantir plena reproduzibilidade, inclusive para o caso de *corpora* que demandem tarefas voltadas a estágios primordiais, tais como de limpeza de dados textuais ou até mesmo a transformação de imagens de documentos digitalizados em dados textuais.

Para a realização de cada etapa do trabalho, diversas ferramentas e pacotes foram utilizados, ressaltando-se principalmente a linguagem de programação *Python*, a linguagem de domínio específico voltada para banco de dados relacional *SQL* e a biblioteca de JavaScript *D3.js*. O primeiro teve como intuito fazer tratamento dos dados e foi utilizado sob a versão *Python 3* com a distribuição *Anaconda* e IDE *Jupyter Notebook*. O Anaconda possui diversos pacotes *Python* pré-instalados, mas além destes, também foi necessária a instalação dos seguintes pacotes: *gensim*¹⁸, *pyLDAvis*¹⁹, *PyMySQL*²⁰ e *nltk*²¹. O segundo teve papel central para armazenar os dados. Já o terceiro serve ao objetivo de produzir visualizações dinâmicas e interativas e foi utilizado sob a IDE *Observable*²². A principal referência utilizada como base para os a compreensão do *D3.js* foi o livro *D3.js in Action* (MEEKS, 2018)

Para a modelagem de tópicos, foram utilizados os seguintes pacotes *Python*: *gensim* e *pyLDAvis*. Para a extração de entidades, foi utilizado o programa *Palavras*²³. Todos os *scripts* e *outputs* se encontram em repositório no GitHub²⁴. Quanto à base de dados utilizada, ela se encontra sob propriedade da FGV/CPDOC, que concedeu permissão de uso de parte do seu acervo para este trabalho. No entanto, não há permissão de acesso direto aos arquivos relativos aos documentos, mas as imagens dos documentos podem ser

¹⁶ <<https://eventos.fgv.br/hdrio2018>>

¹⁷ *History Lab* faz parte do projeto *Archives without Borders*, elaborado pela *Columbia Global Policy Initiative*, que obteve parceria junto à FGV/EMAp e FGV/CPDOC. Sobre o *History Lab*, ver <<http://history-lab.org>>

¹⁸ <<https://pypi.org/project/gensim/>>

¹⁹ <<https://pypi.org/project/pyLDAvis/>>

²⁰ <<https://pypi.org/project/PyMySQL/>>

²¹ <<https://pypi.org/project/nltk/>>

²² <<https://observablehq.com>>

²³ Tal programa foi elaborado por Eckard Bick e é considerado referência para o idioma português (BICK, 2000).

²⁴ Link do repositório: <<https://github.com/Marcelobbr/thesis>>

acessadas por meio do seguinte endereço eletrônico ao sistema de busca de acervos do CPDOC <<http://www.fgv.br/cpdoc/acervo/arquivo-pessoal>>²⁵.

As bases de dados finais voltadas para a visualização se encontram em: <<https://github.com/Marcelobbr/thesis/tree/master/outputs>>. As tabelas voltadas para a visualização foram armazenadas em formato JSON, e, para fazer a renderização no *D3.js*, foram capturados em formato *raw* por meio de funcionalidade do GitHub. O *script* feito em *D3.js* está armazenado no *Observable* no seguinte endereço eletrônico: <<https://observablehq.com/@marcelobbr/thesis-visualization>>. O *layout* da visualização teve como inspiração essencial o trabalho desenvolvido por Christian Laesser, *100 data stories*²⁶. Nas seções a seguir, será feita uma descrição mais detalhada de cada uma das etapas do projeto.

²⁵ Detalhes específicos sobre o acervo do Azeredo da Silveira, cuja base de dados foi utilizada no trabalho, assim como suas coleções, pode ser acessada no endereço eletrônico a seguir: <<http://www.fgv.br/cpdoc/guia/detalhesfundamental.aspx?sigla=AAS>>

²⁶ <<https://projects.christianlaesser.com/100-data-stories>>

3 Base de dados

3.1 Seleção e obtenção dos dados

No âmbito do trabalho desenvolvido anteriormente (RIBEIRO; MORELI; SOUZA, 2018), o *corpus* escolhido foi a coleção Antonio Azeredo da Silveira¹, que faz parte do acervo do CPDOC. Antonio Azeredo da Silveira foi Ministro de Relações Exteriores do Brasil durante o governo de Ernesto Geisel (1974 a 1979). Seus arquivos pessoais foram doados por sua família ao CPDOC em 1996 e abrangem 45 mil documentos. Os documentos da FGV que foram disponibilizados por meio do site *History Lab* foram retirados dessa coleção, mais especificamente, da série “Ministério das Relações Exteriores” (MRE), que é, por sua vez, dividida em 11 subséries: Assuntos Gerais, Assuntos Interamericanos, Brasil-EUA, Bacia do Prata, Despachos com o Presidente Geisel, ONU, Política Nuclear, Relações Bilaterais, Relações Multilaterais, Viagem do Ministro e Viagens do Presidente Geisel.

A série MRE se diferencia das restantes do CPDOC por expor uma quantidade maior de documentos ligados à política externa brasileira na década de 1970, tornando-a valiosa no âmbito do *History Lab*. Como a série é extensa, possuindo 10.550 documentos, e, dado o conteúdo do mesmo, conhecido por ter sido previamente catalogado por arquivistas, haveria grande potencial de presenciar-se tópicos de coerência razoável (e relacionados a temas ligados à política externa brasileira. De fato, entre os temas mais recorrentes da série, se encontram: assuntos nucleares²; a construção da Usina Hidroelétrica de Itaipu³; relações entre Brasil e Estados Unidos; relações entre Brasil e países africanos; dentre outros temas. Parte desses temas já havia sido identificado e devidamente transformado em subsérie por especialistas em arquivologia, como por exemplo, a subsérie *Brasil-EUA* e a subsérie *Política Nuclear*, o que corrobora em parte a coerência do método automático frente ao método manual. Por outro lado, a maioria dos temas principais encontrados foi obtida com a aplicação de modelagem de tópicos, o que possibilitou que fossem encontrados os assuntos mais recorrentes no *corpus* de forma automatizada, e, portanto, muito mais rápida do que se tivesse sido aplicada uma abordagem manual sobre essa atividade.

¹ Para mais informações sobre a coleção Antonio Azeredo da Silveira (AAS), ver o endereço eletrônico da FGV/CPDOC, <<http://www.fgv.br/cpdoc/guia/detalhesfundo.aspx?sigla=AAS>>

² Tais assuntos foram discutidos essencialmente entre Brasil e a República Federativa da Alemanha

³ Tema amplamente discutido tanto em âmbito técnico quanto diplomático, entre Brasil, Argentina e Paraguai

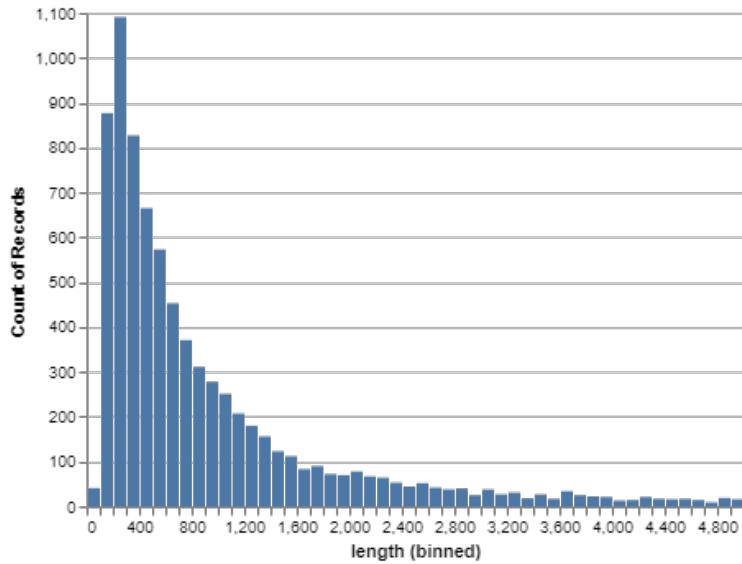


Figura 6 – Histograma dos documentos de acordo com contagem de palavras.

3.2 Preparação dos dados (pré-modelagem)

3.2.1 OCR

A primeira etapa para o início do trabalho envolve o processo de *OCR* (*Optical Character Recognition*)⁴, mas é importante enfatizar que tal procedimento pode por vezes suceder o trabalho prévio de digitalização dos documentos. De todo modo, no caso desse *corpus*, os documentos foram fornecidos digitalizados pelo CPDOC. Além disso, o *OCR* foi feito no âmbito do projeto junto ao *History Lab* (RIBEIRO; MORELI; SOUZA, 2018). Para promover o trabalho de *OCR*, foi utilizada a ferramenta Tesseract⁵, devido a alguns fatores: por ter sua qualidade testada e reconhecida, por ser de livre acesso e pelo fato de ser possível inserir suas funções em *scripts*, possibilitando a automatização de tarefas de *OCR* com grande quantidade de arquivos de imagens. De todo modo, alguns ajustes são necessários para o funcionamento correto do programa. Por exemplo, arquivos “TIF” tendem a trazer resultados melhores em comparação a arquivos “JPG”, devido ao maior contraste (diferenciação entre regiões claras e escuras de uma imagem). Foram promovidos

⁴ OCR envolve o processo de transformação dos dados em formato de imagem para formato de texto. É importante destacar que *OCR* ainda é atualmente um processo imperfeito, pois tal ferramenta pode gerar grande quantidade de sujeira nos dados, principalmente se a digitalização não tiver sido feita em adequação ao posterior *OCR*. É justamente esse fator que cria a necessidade de limpeza dos dados, de modo a aumentar o grau de fidelidade dos dados aos documentos originais e, com isso, aumentar a precisão.

⁵ O Tesseract foi apresentado na UNLV Annual Test of *OCR* em 1995 (RICE; JENKINS; NARTKER, 1995). Em 2005, o Tesseract foi lançado em versão *open-source*. O programa é detalhado no seguinte endereço eletrônico: <<https://github.com/tesseract-ocr>>. Há uma visão geral sobre o programa no seguinte endereço eletrônico: <<https://github.com/tesseract-ocr/docs/blob/master/tesseracticdar2007.pdf>>.

alguns testes com uma alternativa, o *Cloud Vision API*⁶, conjunto de serviços do Google que disponibiliza também *OCR*. Verificou-se que o nível de qualidade seria semelhante ao do *Tesseract* e o fato do serviço ser pago foi fator determinante na escolha da outra opção.

A baixa qualidade da digitalização tende a comprometer o *OCR*, mas esse não é o único fator a reduzir a qualidade da tarefa. A presença de riscos, rasuras e borões de tinta ao longo do texto ou próximos deste são prejudiciais. Além disso, a tipografia, ou seja, a fonte utilizada no texto pode ter influência sobre o resultado (NAGY; NARTKER; RICE, 1999). Textos manuscritos, por sua vez, tendem a ter aproveitamento irrigório em aplicações de *OCR*⁷, ao menos com a tecnologia disponível atualmente. Por fim, a ferramenta de *OCR* utilizada nesse projeto não foi capaz de lidar com gráficos e imagens, o que gerou, junto aos manuscritos, grande quantidade de sujeira nos dados, dado que em vez de ignorar tais imagens, a ferramenta buscou convertê-las em texto.

É importante enfatizar um fator determinante para a escolha da imagem de documento que seria exposta na visualização: dado que a baixa qualidade das imagens de alguns documentos geraria comprometimento do texto resultante do processo de *OCR* em diversos casos, percebeu-se a necessidade de expor a imagem dos documentos originais em vez dos textos resultantes de *OCR*, caso fosse fornecido *hyperlink* ao conteúdo do texto na proposta de visualização. Nesse quesito, torna-se oportuno ressaltar a importância de atentar-se para a organização dos documentos de modo que haja a correta correspondência entre os tipos de arquivo. No caso do CPDOC, havia documentos físicos, as versões digitalizadas e as versões pós-*OCR*. Como havia acesso a uma documentação que permitia estabelecer a correspondência entre os arquivos, assim como havia uma listagem das URLs correspondentes a cada dossiê, foi possível estabelecer a relação correspondente a cada arquivo. A Figura 7 mostra a distribuição de documentos por tamanho e legibilidade⁸.

3.2.2 Limpeza de dados

Nessa etapa, foi feita a organização e a correção de termos não-dicionarizados com a finalidade de aumentar a eficácia da modelagem de tópicos. Inicialmente, foi necessário identificar os principais idiomas presentes no *corpus*, que no caso são: português, inglês, francês, alemão e espanhol, além das variações entre inglês britânico e americano e português europeu e brasileiro. Havia também a questão de que os documentos não foram escritos de acordo com a nova ortografia da língua portuguesa, mas isso especificamente tinha pouca repercussão no processo de limpeza. Essa diversidade de idiomas poderia dificultar

⁶ Endereço eletrônico da ferramenta: <<https://cloud.google.com/vision/>>

⁷ Existem atualmente ferramentas como o *Transkribus*, que permitem trabalhar com manuscritos, mas requerem algum grau de trabalho manual para formar uma base de treino. Além disso, possuem melhor aproveitamento quando há uma quantidade significativa de texto gerado por um mesmo autor (o que permite uma base de treino adequada). Detalhes sobre a ferramenta estão disponíveis no seguinte endereço eletrônico: <<https://transkribus.eu/Transkribus>>

⁸ São mostrados documentos com até 10 mil palavras.

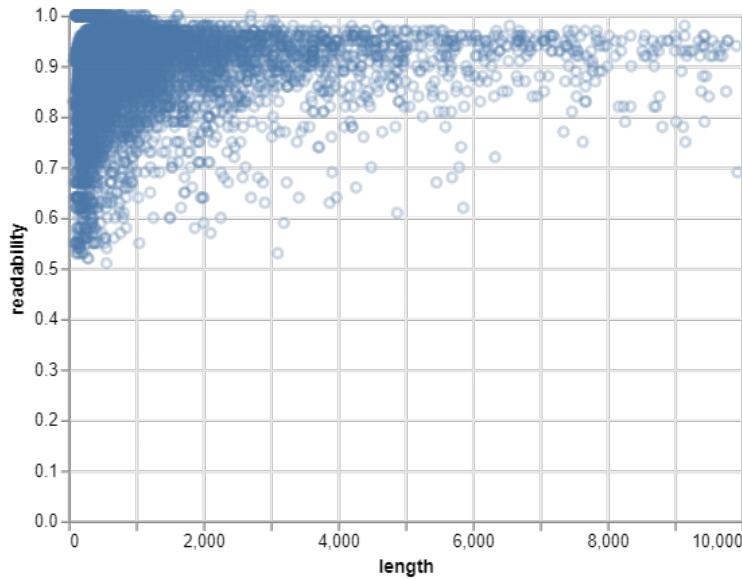


Figura 7 – Distribuição de documentos de acordo com legibilidade calculada. Obs: Documentos com menos de 10 sentenças foram filtrados devido à imprecisão potencial da medição.

consideravelmente as correções caso não fosse criado um código de programação que excluísse termos de outros idiomas e suas variantes da lista de termos não-dicionarizados. Para que a biblioteca do *Python* utilizada, *enchant*⁹, realizasse a tarefa corretamente, esta demandava a devida identificação dos idiomas de modo que fosse formulada a lista de termos não-dicionarizados, a partir dos quais seriam feitas as correções (RIBEIRO; MORELI; SOUZA, 2018). Na Figura 8 se encontra um gráfico com os idiomas principais.

Tal como afirmam Rahm e Do (RAHM; DO, 2000), para a realização de uma limpeza de dados efetiva, é necessário promover análise de dados de modo a detectar erros de modo mais preciso. Além da básica análise manual por meio da verificação de amostras da base de dados, é importante recorrer a programas de modo a encontrar padrões de erros e inconsistências. No caso do projeto anterior, por exemplo, foi utilizada uma biblioteca do *Python*, *enchant*, o que permitiu detectar termos não-dicionarizados, ou seja, palavras que não estavam presentes nos idiomas principais do *corpus*. Existem programas que fazem correção por meio de predição da versão correta mais provável a partir de cada erro, como é o caso de bibliotecas do *Python* como o *soundex*¹⁰.

Com o intuito de corrigir os termos presentes na lista de termos não-dicionarizados, foi necessário e eficaz adotar soluções gerais de modo que abrangesse o maior número possível de variações de erro. Para isso, considerou-se oportuna a estratégia de aplicar expressões regulares no processo de limpeza de dados. Com isso, pôde-se corrigir de modo mais dinâmico, por exemplo, palavras que foram cortadas por quebra de linha, um

⁹ <<https://pypi.org/project/pyenchant/>>.

¹⁰ <<https://pypi.org/project/soundex/>>.

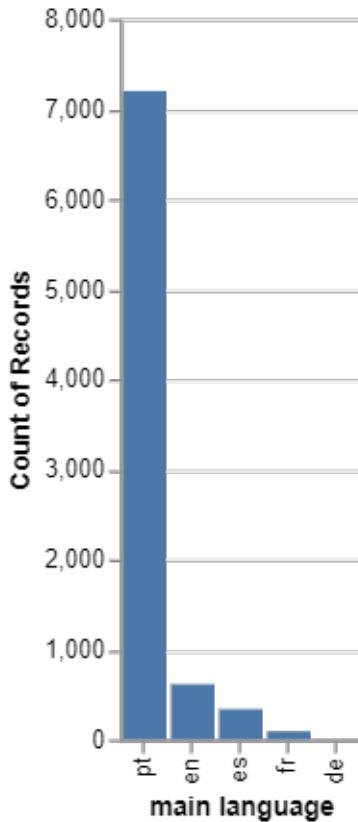


Figura 8 – Principais idiomas do *corpus*.

problema recorrente no caso desse *corpus*. Foram obtidas soluções que permitem a alteração automatizada e geral para tais padrões de erros com a adoção de expressões regulares dentro de *script* escrito em *Python* (RIBEIRO; MORELI; SOUZA, 2018). Ao utilizar esse procedimento, foi possível atingir uma redução considerável do número total de erros (termos não-dicionarizados) em 528 mil (25% de um total de 2,1 milhões) aproximadamente.

É importante descrever outros procedimentos de pré-processamento comuns para trabalhos com modelagem de tópicos: remoção de *stopwords*, procedimentos com *stemming* e *lemmatization*, assim como geração de filtros. O especialista responsável pela configuração do algoritmo de pré-processamento deve sempre avaliar o custo-benefício ao adotar cada procedimento.

Remoção de *stopwords* é uma prática voltada para retirar de um *corpus* os termos que possuem alta frequência e baixa relevância semântica, denominados *stopwords*, como artigos, preposições, pronomes e conjunções. São, portanto, removidos da base de dados antes do processamento por modelagem de tópicos. Há casos em que determinados tipos de *stopwords* possuem relevância num *corpus*. Tal situação ocorre, por exemplo, quando se deseja comparar estilos literários. Em tais situações, pode-se configurar quais *stopwords* serão mantidas na base. É possível também decidir por remover palavras específicas que, mesmo não sendo caracterizadas como *stopwords*, comprometem de algum modo a fase

de modelagem de tópicos, por estarem associadas a algum padrão de documento sem relevância analítica, como termos de cabeçalho ou notas de rodapé. Sob essa etapa, foi realizada a remoção de *stopwords* existentes nos 5 idiomas mais frequentes do *corpus* (português, inglês, espanhol, francês e alemão), além de retirar termos com frequência acima de 80% nos documentos. Termos acima desse percentual tendem a ter características de *stopwords*, ou seja, são de frequência comum ao *corpus* (RIBEIRO; MORELI; SOUZA, 2018).

Stemming e *lemmatization* são métodos utilizados para reduzir a flexão ou derivação de palavras de modo a recuperar sua raiz. Por exemplo, as palavras “democracia”, “democrático”, “democratização” e “democratizar” são derivações com a mesma raiz. O primeiro algoritmo de *stemming* foi desenvolvido em 1968 (LOVINS, 1968) e diversos modelos foram criados desde então, sendo que cada um deles traz diferentes resultados. A biblioteca do *Python*, *nltk* (natural language toolkit), por exemplo, possui três modelos principais: Porter, Lancaster e Snowball. Tal como afirma Manning et al. (MANNING; RAGHAVAN; SCHÜTZE, 2010), enquanto que *stemming* trabalha fundamentalmente com a retirada de sufixos, o método de *lemmatization* funciona com base em análise de vocabulário e morfologia das palavras. Em idiomas europeus de um modo geral, não costuma haver diferença considerável entre os métodos. Os benefícios de sua aplicação dependem do idioma e das particularidades do *corpus* a ser trabalhado. Apesar do potencial desses métodos, não foi verificado ganho significativo após alguns testes, sendo mantido somente o método de remoção de *stopwords* (RIBEIRO; MORELI; SOUZA, 2018).

4 Modelagem de tópicos

Ainda no contexto do projeto ligado ao *History Lab*, para a prática de modelagem de tópicos, foram utilizados os pacotes *gensim* e *pyLDAvis*, sendo que o primeiro possibilita o uso de aplicações de modelagem de tópicos, tais como LDA, LSI e HDP, enquanto que o segundo fornece interfaces de visualização de dados¹. Foi comparado o resultado entre o *corpus* sem filtro de idioma e com filtro de idioma. Na versão sem filtro, a modelagem era efetuada independentemente do conteúdo do documento. Já na segunda versão, buscouse manter somente documentos majoritariamente escritos em português, retirando os documentos restantes da base de dados. Foram obtidos resultados melhores ao usar a versão com filtro, dado que, sem essa característica, a modelagem tendia a criar tópicos relativos a idiomas, pois os termos de outros idiomas presentes no *corpus* terminavam por formar tópicos característicos de um dos idiomas principais do *corpus*. Em alguns casos, tópicos distintos podiam tratar dos mesmos temas, mas em linguagens diferentes. A formação desses agrupamentos tendia a gerar demasiadas distorções e, portanto, foi tomada decisão de manter esse filtro (RIBEIRO; MORELI; SOUZA, 2018).

Outro filtro utilizado retirou da base de dados documentos em que o processo de *OCR* não se adequaria devido à baixa legibilidade. Tais documentos possuíam uma ou mais das seguintes características: eram textos manuscritos, apresentavam grande quantidade de rasuras e riscos no corpo de texto ou eram, na realidade, elementos não textuais, tais como figuras e gráficos. Assim como no caso dos idiomas, esses documentos faziam com que o processo de modelagem criasse tópicos agrupados pelo padrão de sujeira no texto, prejudicando a eficácia da modelagem.

Concluída a elaboração dos filtros, a modelagem com LDA foi testada com versões segundo o número de tópicos construídos. As versões testadas foram: modelagem com 20, 30, 40, 45, 60 e 100 tópicos totais. Também foi estabelecido o número de passos (*passes*) em 50², e, de modo a garantir reproduzibilidade dos resultados, foi fixado o *seed*³ em 0 (zero). O restante dos parâmetros foi deixado sob o padrão do *gensim*⁴. Por fim, foram promovidos testes com o processo de modelagem HDP.

¹ A ferramenta de visualização *LDAvis* será abordada no Capítulo 6.

² Número de passos ao longo do *corpus* durante a etapa de treinamento.

³ *Seed* é referente ao estado aleatório. Como o estado inicial do um modelo iterativo nesse caso não é fixo, a escolha de um *seed* garante que o resultado seja sepicado. Sob o pacote *gensim* do *Python*, essa variável é denominada *random_state*.

⁴ Os parâmetros são: distributed=False, chunksize=2000, update-every=1, alpha='symmetric', eta=None, decay=0.5, offset=1.0, eval-every=10, iterations=50, gamma-threshold=0.001, minimum-probability=0.01, ns-conf=None, minimum-phi-value=0.01, per-word-topics=False, callbacks=None, dtype=<class 'numpy.float32'>. Sobre tais parâmetros do *gensim*, ver a documentação: <<https://radimrehurek.com/gensim/models/ldamodel.html>>.

Tal como explicado no Capítulo 1, os tópicos são definidos probabilisticamente em espaço multidimensional em que as frequências de termos dos tópicos formam um espaço vetorial com múltiplas variáveis. De modo a possibilitar a visualização das diferenças entre os tópicos criados pelo algoritmo, foi feito um gráfico interativo por meio do *PyLDAvis* onde constam os tópicos, representados com aplicação de PCA (*Principal Component Analysis*), um procedimento que permitiu a visualização da distância dos tópicos em espaço bidimensional. Após realização de análise de cada uma das versões por parte de especialista de domínio, foi concluído que o mais adequado para o esse *corpus* seria usar a versão de LDA com 100 tópicos totais. Com o processo de modelagem definido, foram criadas tabelas que relacionavam documentos e tópicos por ordem de *score*, lembrando que *score*, nesse caso, representa a probabilidade de associação entre documento e tópico (RIBEIRO; MORELI; SOUZA, 2018).

5 Geração e tratamento de dados

5.1 Geração de metadados

A etapa de armazenamento dos resultados foi efetuada com a inserção dos metadados primeiramente em tabelas SQL para formar tabelas relacionais e imutáveis, posteriormente em *pandas*¹ para manipulação e transformação dos dados, e, finalmente, em JSON para a alimentação do modelo de visualização. Foram escritos códigos em *Python* para fazer a integração de dados, usando bibliotecas que possibilitaram a alimentação dos dados. Os meta-dados adicionados na tabela de documentos foram: id (identificador do documento, seguindo o nome original dos mesmos), corpo de texto, data (de emissão), coleção, idioma principal, URL (endereço eletrônico da versão digitalizada do documento) e legibilidade. Também foram criadas tabelas para tópicos, além das que relacionavam tópicos e documentos, assim como tópicos e palavras.

Idioma principal e legibilidade são informações que foram geradas por meio de linguagem de programação. O idioma foi identificado com uso de uma biblioteca do *Python*, *langdetect*, que identifica probabilisticamente o idioma de cada frase de um documento. Quando o número de frases identificadas como pertencentes a determinado idioma representa mais de 30% do corpo de texto, o idioma é marcado no documento. No entanto, isso é feito apenas com os idiomas principais do *corpus*. Como a detecção é probabilística, podem ser identificados incorretamente outros idiomas, um problema que ocorre principalmente pelo fato de a detecção ser comprometida pela presença de erros de *OCR* (RIBEIRO; MORELI; SOUZA, 2018).

Já a legibilidade refere-se à proporção de termos corretos frente ao total de termos presentes no documento. Como explicado na seção sobre *OCR*, textos com muita sujeira advinda do processo de *OCR* podem continuar legíveis para um humano, mas dificilmente serão operáveis por ferramentas e códigos de programação, principalmente se não for feito previamente um procedimento de limpeza dos dados. Para estabelecer o critério de legibilidade, foi usado o mesmo pacote de detecção de idioma, que identificava probabilisticamente se cada frase de um texto pertencia a um dos idiomas principais do acervo. Ao comparar a razão de frases detectadas como pertencentes a um idioma e o número total de frases do documento, foi estabelecido um valor percentual referente à legibilidade de cada documento (RIBEIRO; MORELI; SOUZA, 2018).

¹ <<https://pandas.pydata.org/>>

5.2 Tratamentos dos dados

Os tratamentos voltados para visualização de objetos compreendem duas das camadas de granularidade abordadas por Windhager et al. e foram sintetizadas no Capítulo 2 (WINDHAGER et al., 2018), sendo, nesse caso, *Single Object Previews* e *Multi-Object Previews*, enquanto que a camada *Collection Overviews* será abordada em seguida. Tal como será aprofundado na parte III, a ferramenta de visualização principal possibilita acessar um agrupamento de objetos (*Multi-Object Previews*) de acordo com o tópico selecionado pelo usuário, enquanto que a funcionalidade de clique sobre cada documento permite a visualização de sua versão digitalizada (*Single Object Previews*).

Para a geração das tabelas em JSON, algumas transformações foram feitas nos dados por meio do *pandas*. É importante destacar que devido ao fato da visualização trabalhada ter sido voltada para um grafo, com objetos conectados entre si, o formato dos dados deve estar em conformidade. Um formato adequado e bem integrado ao *D3.js* é o *nested data* (dados aninhados), em que, recursivamente, objetos são “filhos” de outros objetos e um objeto “ancestral” pode conter uma lista de “descendentes”, em outras palavras, uma lista de objetos (comumente denominado de “array of objects”). Essa é a estrutura da visualização a ser construída, que será explicada na parte III. Documentos relativos a um tópico estão ligados a um grupamento de palavras (termos contidos no documento) e pessoas (entidades mencionadas no documento). Portanto, será necessário migrar do formato anterior contido nas tabelas *SQL* para o formato *noSQL* que será armazenado nos arquivos em JSON.

Em primeiro lugar, foi necessário reunir os dados, que estavam separados em diversas bases relacionais em *SQL*. Os dados foram coletados para tabela do *pandas* em 20 documentos para cada tópico. O formato nesse caso é *many to many*, ou seja, múltiplos documentos podem estar associados a múltiplos tópicos, dado que a aplicação do LDA não implica em exclusividade de tópicos para documentos, mas numa relação de probabilidade. Para cada documento, foi inserida uma lista de palavras dentre as 20 de maior *score* por tópico (coluna *tokens*) que ocorressem no mesmo, assim como foi inserida uma lista de pessoas mencionadas no mesmo (coluna *names*). A tabela resultante, que será a base para elaboração dos dados seguintes, pode ser visualizada na Figura 9. Tais dados foram salvos em formato *pickle*² sob o arquivo *tables_dict.pkl*.

As seguintes bases de dados foram criadas em formato JSON, nesse caso, contendo um arquivo para cada tipo de dado e para cada tópico (a numeração localizada no sufixo do nome do arquivo indica o tópico): *topic*, *names_list* e *tokens_list*. Quanto ao arquivo *topic.json*, foi gerado um único exemplar, que apresenta os mesmos dados da tabela do *pandas* ilustrado na Figura 9, mas dessa vez transformados em formato JSON (ver Figura

² <<https://docs.python.org/3/library/pickle.html>>

	doc_id	topic_id	topic_score	date	pdf	body	length	tokens	name	year
15	pn_1974.08.15_doc_III-31	5007	0.303900	1978-01-15	http://www.fgv.br/cpdoc/acervo/arquivo-pessoal...	secreto-exclusivo\nnsubsídios para as consult...	4549	[nuclear, acordo, energia, nucleares, brasil, ...]	[James Earl (Jimmy) Carter, Helmut Schmidt]	1978
16	pn_1976.12.28_doc_29	5007	0.300885	1978-01-15	http://www.fgv.br/cpdoc/acervo/arquivo-pessoal...	telegrama recebido\nnnaaaa 97034 . é? \no"...	845	[nuclear, acordo, energia, nucleares, brasil, ...]	[James Earl (Jimmy) Carter]	1978
17	pn_1974.08.15_doc_II-1	5007	0.290109	1976-01-15	http://www.fgv.br/cpdoc/acervo/arquivo-pessoal...	\n\nmmmm\n\n'a. . . , . ..\n\n'o acordo nucl...	1672	[acordo, energia, nucleares, brasil, ...]	[James Earl (Jimmy) Carter]	1976
18	pn_1976.12.28_doc_16	5007	0.286767	1978-01-06	http://www.fgv.br/cpdoc/acervo/arquivo-pessoal...	antonio la' aiii-1315230 da sxmráxminjáwowa i...	330	[acordo, energia, brasil, urânia, rfa, salvagu...]	[Antonio Azeredo da Silveira]	1978
19	d_1974.04.23_doc_XXI-12	5007	0.285340	1979-06-15	http://www.fgv.br/cpdoc/acervo/arquivo-pessoal...	\n\n\nlembrete n? 011\n\npolítica nuclear ...	1590	[nuclear, acordo, energia, nucleares, brasil, ...]	[James Earl (Jimmy) Carter]	1979

Figura 9 – Tabela de documentos para um tópico selecionado.

10). Segue a descrição de cada um dos tipos de arquivo:

- topic.json: Lista cada um dos 20 primeiros documentos (“doc_id”), o respectivo *score* frente ao tópico (“topic_score”), tamanho em número de palavras (“length”), endereço eletrônico para a versão digitalizada (“url”), lista de palavras contidas no documento (“tokens”) e lista de pessoas mencionadas no documento (“names”) (ver Figura 10). Sobre tais listas:
 - names: Lista de objetos que mostra quais dessas entidades são mencionadas por documento (formato *many to many*).
 - tokens: Lista de objetos que mostra quais desses tokens ocorreram por documento (formato *many to many*).
- names_list.json: Tabela contendo nomes de pessoas mencionadas em ao menos um dos 20 documentos de maior *score* para um tópico selecionado e sua respectiva contagem em todo o *corpus* (ver Figura 11).
- tokens_list.json: Tabela contendo 20 *tokens* de maior *score* para um tópico e seu respectivo *score* frente ao mesmo. Os dados passaram por ordenação (decrescente) de acordo com o valor do *score* (ver Figura 12).

Os valores que representam a magnitude dos nós (raio dos círculos), são representados pelos *scores* frente a cada tópico, no caso de nós que representam documentos e *tokens*. Quanto aos nós relativos a pessoas, o valor é dado pela contagem de ocorrência nos documentos. Os *scores* ocorrem num intervalo de 0 a 1, sendo indicativos da correspondência entre tópicos e elementos de interesse (documentos ou *tokens*). De modo a trabalhar com a mesma ordem de valores, foi necessário normalizar os dados da contagem de pessoas

```
[{"doc_id": "pn_1975.04.25_doc_4", "topic_score": 0.545832, "url": "http://www.fgv.br/cpdoc/acervo/arquivo-pessoal/AAS/textual/documentos-referentes-ao-programa-nuclear-brasileiro-durante-a-gestao-de-azeredo-da-silveira-como-ministro-das-relacoes-exterieores-tratando-de-div", "length": 7465, "tokens": ["nuclear", "acordo", "energia", "nucleares", "brasil", "uranio", "armas", "tecnologia", "rfa", "tratado", "cooperacao", "salvaguardas", "programa", "aiea", "combustivel", "proliferacao", "reatores", "alemanha", "atomica", "fins"], "name": [], "year": 1977, "doc": "doc_4", "date": 227232000000}, {"doc_id": "pn_1975.04.25_doc_5", "topic_score": 0.497948, "url": "http://www.fgv.br/cpdoc/acervo/arquivo-pessoal/AAS/textual/documentos-referentes-ao-programa-nuclear-brasileiro-durante-a-gestao-de-azeredo-da-silveira-como-ministro-das-relacoes-exterieores-tratando-de-div", "length": 3766, "tokens": ["nuclear", "acordo", "energia", "nucleares", "brasil", "uranio", "armas", "tecnologia", "rfa", "tratado", "cooperacao", "salvaguardas", "programa", "aiea", "combustivel", "proliferacao", "reatores", "alemanha", "atomica", "fins"], "name": [], "year": 1977, "doc": "doc_5", "date": 227232000000}, {"doc_id": "be_1977.01.27_doc_II-11",
```

Figura 10 – Segmento dos dados contidos no arquivo *topic_1.json*, que contém a totalidade dos dados visuais para o tópico 1.

```
[{"name": "James Earl (Jimmy) Carter", "count": 15.507767337860294}, {"name": "Antonio Azeredo da Silveira", "count": 19.6887445579689}, {"name": "John Hugh Crimmins", "count": 12.94075363634189}, {"name": "Helmut Schmidt", "count": 8.69707517144841}, {"name": "Cyrus Vance", "count": 12.274135660257478}]
```

Figura 11 – Dados contidos no arquivo *names_list_1.json*, relativo a nomes contidos em documentos.

```
[{"token": "nuclear", "score": 0.085}, {"token": "acordo", "score": 0.036}, {"token": "energia", "score": 0.035}, {"token": "nucleares", "score": 0.03}, {"token": "brasil", "score": 0.023}, {"token": "uranio", "score": 0.02}, {"token": "armas", "score": 0.02}, {"token": "tecnologia", "score": 0.017}, {"token": "rfa", "score": 0.016}, {"token": "tratado", "score": 0.014}, {"token": "cooperacao", "score": 0.014}, {"token": "salvaguardas", "score": 0.012}, {"token": "programa", "score": 0.011}, {"token": "aiea", "score": 0.01}, {"token": "combustivel", "score": 0.009}, {"token": "proliferacao", "score": 0.008}, {"token": "reatores", "score": 0.007}, {"token": "alemanha", "score": 0.007}, {"token": "atomica", "score": 0.007}]
```

Figura 12 – Dados contidos no arquivo *tokens_list_1.json*, relativo a palavras contidas em documentos.

com uso de normalização min-max³, de modo a reescalar os valores para o intervalo 0-1. Essa manipulação tornou os dados padronizados para a etapa de visualização e permitiu uma simplificação da escolha das escalas para o tamanho dos círculos.

Assim como todo o restante dos arquivos gerados na dissertação, os dados foram armazenados em repositório do *GitHub*, que são, por sua vez, importados pelo *Observable*, onde está o código em *D3.js*. Com isso, fica mais clara a estrutura dos dados criados no arquivo *topic*, representados na Figura 13. Dentro desses dados, existem listas de objetos que foram importantes para o estabelecimentos das conexões do grafo, pois é por meio

³ A fórmula na normalização min-max é dada por: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$, onde x é um valor em uma lista, $\min(x)$ é o valor mínimo dessa lista e $\max(x)$ é seu valor máximo.

destes que são geradas as arestas entre os nós⁴. No Capítulo 6, sobre a elaboração da proposta de visualização, será explicado detalhadamente como esses dados foram integrados ao modelo visual.

```
topics = ▶ Array(2) [
  0: ▶ Object {
    topic: "topic1"
    documents: ▶ Array(20) [
      0: ▶ Object {doc_id: "pn_1975.04.25_doc_4", topic_score: 0.545832, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pessoal"
      1: ▶ Object {doc_id: "pn_1975.04.25_doc_5", topic_score: 0.497948, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pessoal"
      2: ▶ Object {doc_id: "be_1977.01.27_doc_II-11", topic_score: 0.485872, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pes
      3: ▶ Object {doc_id: "pn_1975.04.25_doc_6", topic_score: 0.484978, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pessoal"
      4: ▶ Object {doc_id: "be_1977.06.01_doc_II-1", topic_score: 0.480625, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pess
      5: ▶ Object {doc_id: "be_1977.04.29_doc_II-27", topic_score: 0.416698, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pes
      6: ▶ Object {doc_id: "rb_1974.04.17_doc_II-8", topic_score: 0.409007, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pess
      7: ▶ Object {
        doc_id: "d_1974.03.26_doc_XXIII-31"
        topic_score: 0.38098
        url: "http://www.fgv.br/cpdoc/acervo/arquivo-pessoal/AAS...com-o-presidente-da-republica-ernesto-geisel-o-do"
        length: 10688
        tokens: ▶ Array(20) ["nuclear", "acordo", "energia", "nucleares", "brasil", "uranio", "armas", "tecnologia", "rfa", "
        name: ▶ Array(2) ["James Earl (Jimmy) Carter", "John Hugh Crimmins"]
        year: 1978
        doc: "doc_XXIII-31"
        date: 261446400000
        x: -117.5570504584946
        y: 161.80339887498948
      }
      8: ▶ Object {doc_id: "pn_1975.04.25_doc_8", topic_score: 0.367693, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pessoal
    ]
  ]
]
```

Figura 13 – Visualização dos dados incorporados ao *D3.js* provenientes do arquivo *topic_1.json*.

Além da visualização de documentos de maior *score* pertencentes a cada tópicos, foi elaborada uma visualização da coleção de modo a trazer uma perspectiva geral dos tópicos (ver imagem 14). Tal manipulação busca a completar as três camadas de granularidade abordadas por Windhager et al. e sintetizadas no Capítulo 2 (WINDHAGER et al., 2018). Enquanto que os dados gerados nos tratamentos e transformações anteriores foram voltados à visualização principal, que atende às camadas de *Single Object Previews* e *Multi-Object Previews*, esta última transformação busca cobrir a camada de *Collection Overviews*. Tais tratamentos envolveram a geração de uma tabela de documentos, nas linhas, e tópicos, nas colunas. Os tratamentos e a visualização foram promovidos por meio da linguagem de programação *Python*. A imagem, salva em formato *png*, foi posteriormente importada para o *Observable*.

⁴ Para os dados contidos em *topic*, as listas se chamam *names* e *tokens*. Quanto aos dados contidos em *names_list* *tokens_list*, as listas são denominadas de *docs*. É importante observar que no primeiro dado é feito nas arestas o “caminho de ida” dos documentos para outros nós, enquanto que nos outros dados é feito o “caminho de volta” dos outros nós para os documentos.

Documentos	Topicos	0	1	2	3	4	5	6	7	8	9	...	90	91	92	93	94	95	96	97	98	99
ag_1973.11.20_doc_l-101	0.215000	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ag_1973.11.20_doc_l-102	0.140926	0.0	0.0	0.0	0.0	0.0	0.213801	0.000000	0.0	0.000000	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ag_1973.11.20_doc_l-103	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ag_1973.11.20_doc_l-106	0.305181	0.0	0.0	0.0	0.0	0.0	0.113188	0.000000	0.0	0.000000	...	0.100333	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ag_1973.11.20_doc_l-108	0.134156	0.0	0.0	0.0	0.0	0.0	0.000000	0.040162	0.0	0.022936	...	0.037407	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ag_1973.11.20_doc_l-109	0.169317	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ag_1973.11.20_doc_l-111	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ag_1973.11.20_doc_l-112	0.418504	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ag_1973.11.20_doc_l-113	0.121600	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.020580	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ag_1973.11.20_doc_l-113	0.334299	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	...	0.043913	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Figura 14 – Amostra dos dados gerados (10 primeiros documentos) para visualização da coleção (*Collection Overviews*).

Parte III

Visualização

6 Seleção do modelo de visualização

A proposta de visualização¹ representa a última etapa do fluxograma sugerido e tem como intuito trazer uma nova ferramenta que permita melhorar a interpretabilidade de resultados de modelagem de tópicos de modo a complementar as soluções existentes atualmente, sejam elas visualizações ou métricas estatísticas que automatizam a parametrização dos modelos. Os referenciais teóricos apresentaram algumas dessas soluções². É importante que a solução não apenas apresente e sintetize informação, mas que forneça também boa experiência de usuário, de modo que a compreensão sobre seu funcionamento tenha baixa curva de aprendizado. Para isso, é necessário incluir instruções concisas, mas precisas, para o devido manuseio da ferramenta.

A solução desenvolvida possui uma visualização principal, em formato de grafo, que permite observar seleções da coleção, nesse caso, documentos principais ligados a cada tópico selecionado pelo usuário, o que atende duas das camadas de granularidade abordadas por Windhager et al. e foram sintetizadas no Capítulo 2 (WINDHAGER et al., 2018), sendo, nesse caso, *Single Object Previews* e *Multi-Object Previews*. Há também uma visualização auxiliar que comprehende a terceira camada (*Collection Overviews*), que comprehende a percepção da coleção como um todo, ordenando os tópicos para o usuário. Devido ao fato de que o usuário central da ferramenta seria o especialista de domínio em vez de pessoa do público geral, não foi necessário fornecer ênfase na camada de *Collection Overviews*. Primeiramente, será abordada a visualização principal, parte que envolveu a pesquisa do estado da arte em visualização de tópicos, antes de abordar a visualização auxiliar.

Os referenciais teóricos trouxeram boa perspectiva sobre o estado da arte, mas podemos considerar uma delas como principal fonte de inspiração para formular uma nova proposta de visualização. Tal ferramenta, denominada de *Jigsaw* pelos autores, é voltada para visualização interativa e exploração de um *corpus* que permite realçar um documento e descobrir relações com entidades presentes no mesmo. Sua estrutura é apresentada como um grafo de inter-relações, fornecendo apenas uma amostragem da base de dados original (WARD; GRINSTEIN; KEIM, 2015).

Além dessa referência, pode-se afirmar que o trabalho de Christian Laesser (ver

¹ Segue endereço eletrônico que expõe as funcionalidades da ferramenta: <<https://observablehq.com/@marcelobbr/thesis-visualization>>.

² No caso de visualizações, ver por exemplo: (SIEVERT; SHIRLEY, 2014), (GARDNER et al., 2010), (SNYDER et al., 2013), (CHANAY; BLEI, 2012). Quanto a métricas, ver por exemplo: (ALSUMAIT et al., 2009), (MIMNO et al., 2011), (CHUANG et al., 2015), (TADDY, 2012) e (BISCHOF; AIROLDI, 2012).

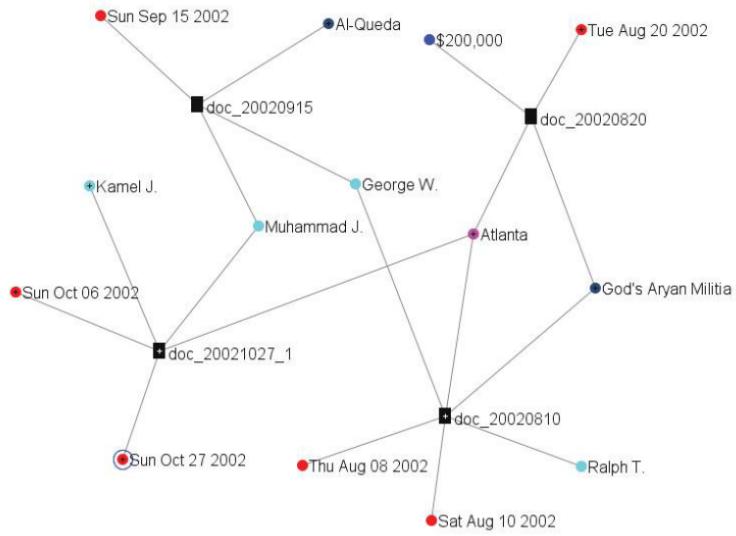


Figura 15 – Grafo *Jigsaw*, representando relações entre documentos e entidades. (WARD; GRINSTEIN; KEIM, 2015).

Figura 16), *100 data stories*³ (LAESSER, 2017) tenha sido a inspiração principal quanto à forma de elaborar a interação e o formato do gráfico. Sua visualização se propõe a promover exploração visual de um pequeno *corpus* composto por *podcasts* e é essencialmente voltado para experiência de usuário e impacto visual. Apesar dessa visualização não ter sido pensada especificamente para modelagem de tópicos, a mesma possui funcionalidades presentes em outras visualizações abordadas na seção sobre referenciais teóricos em visualização (seção 2.2). Por exemplo, assim como no caso da ferramenta *Jigsaw*, esta se propõe a analisar um grafo de inter-relações entre documentos e entidades, nesse caso, pessoas, além de permitir acesso ao conteúdo original de um documento quando se interage com o mesmo. Essa interação também permite a identificação da relação entre documentos e termos-chave, presentes no campo interno do grafo. Tais funções somam à qualidade vista como principal: o fato de ser intuitivo e simples para o usuário.

O trabalho de Christian Laessner serve essencialmente à elaboração do *layout* da visualização que será proposta nesta dissertação, enquanto que as ferramentas que foram enfatizadas na seção 2.2 foram relevantes para refletir sobre quais elementos podem ser considerados relevantes em uma ferramenta visual no sentido de assistir o processo de validação de tópicos. Essas ferramentas foram o *Jigsaw*, mencionado no parágrafo anterior, e o LDAvis. Esta última fornece grande ênfase ao conjunto de palavras que possuem alta probabilidade de ocorrência em cada documento, havendo um parâmetro (λ) que faz ponderação sobre o quanto que esse termo é específico ao tópico, assim como seu grau de ocorrência ao longo de todo o *corpus*. Tal exposição de relação entre tópicos e termos torna a visualização muito esclarecedora sobre o tema dos tópicos assim como podem

³ A visualização foi obtida por meio do seguinte endereço eletrônico: <<https://lab.christianlaesser.com/>>

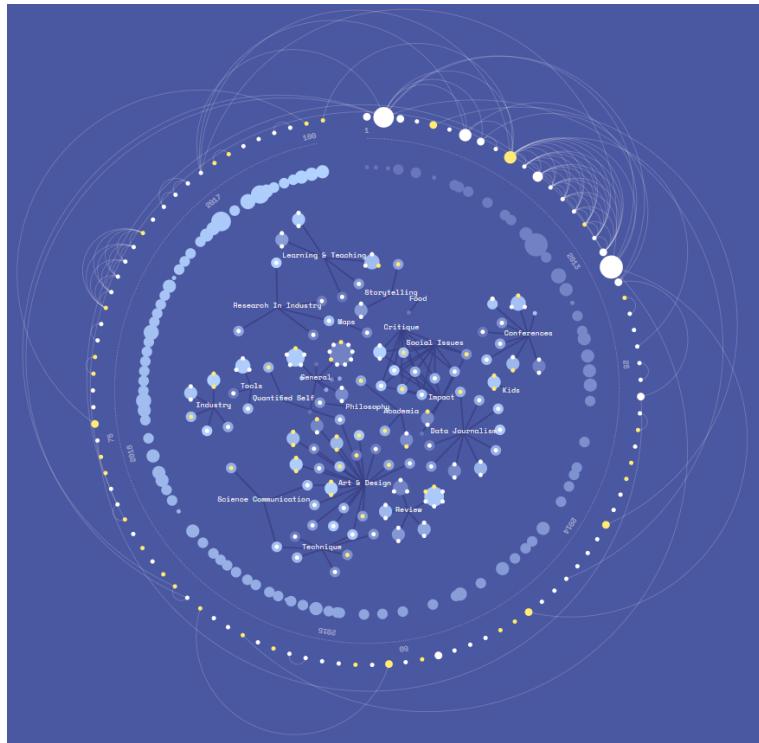


Figura 16 – Visualização 100 Data Stories (<https://lab.christianlaesser.com/>).

sinalizar tópicos de má qualidade quando estes não formam um grupo coeso de termos. A eficácia da ênfase em termos mostra que estes seriam elementos essenciais de uma solução de visualização.

A ferramenta escolhida neste trabalho para a geração da solução de visualização foi o *D3.js*. A opção pelo se justifica pela alta flexibilidade e elevado grau de interatividade frente a outras opções, o que viabilizaria ao usuário final da visualização uma melhor compreensão do *corpus*. O *Python* possui pacotes que permitem construir grafos, tal como o *NetworkX*, que pode ser acrescido de interatividade se usado juntamente ao *Altair*⁴, que é construído por cima do *Vega-Lite*⁵, sendo esta por sua vez uma linguagem de alto nível elaborada por cima do próprio *D3.js*. Apesar do alto grau de complexidade da linguagem, o *D3.js* possui como vantagem principal o fato de garantir a possibilidade de aplicação de todas as funcionalidades esperadas para este trabalho⁶. No próprio repositório do Vega é feita a afirmação de que o *D3.js* possui expressividade como diferencial e seria mais apropriado para elaboração de novas ideias de visualização, enquanto que o *Vega* seria voltado para visualizações mais comuns, ainda que permitam algum grau de personalização⁷.

⁴ <https://altair-viz.github.io/>

⁵ <https://vega.github.io/vega-lite/>

⁶ *D3.js* permite elaborar grafos de alta interatividade, com elementos que podem redirecionar para *hyperlinks*. Ao mesmo tempo, possui grande quantidade de referências de modelos visuais.

⁷ Tal como escrito em página específica do *Vega*: “By design, *D3* will maintain an expressivity advantage and in some cases will be better suited for novel design ideas. On the other hand, we intend Vega

Outra ferramenta de visualização que permite a construção de grafos é o *Gephi*⁸. No entanto, o *Gephi* não implementa interatividade, como ocorre com o *D3.js*. Apesar disso, a ferramenta foi utilizada primordialmente com o intuito de testar possíveis opções de visualização de grafos. Foram feitos testes com a base de dados de entidades extraídas do *corpus*. As entidades nesse caso foram as pessoas citadas nos documentos, sendo que cada nó do grafo representa uma pessoa e cada aresta representa a coocorrência de pessoas presente em um mesmo documento. Não foram utilizados filtros⁹, dado que a utilização dessa ferramenta não foi continuada. Foram geradas visualizações com alguns modelos de algoritmos de *layout*, do gênero *force-directed graph*, disponibilizados pelo *Gephi*: Fruchterman-Reingold¹⁰ (Figura 32 do apêndice B) e Yifan Hu¹¹ (Figura 33 do apêndice B). Além de alterações nos parâmetros dos *layouts*, foram feitas outras configurações de acordo com instruções do *Gephi*¹². Se, por um lado, observa-se que os grafos gerados por meio do *Gephi* trazem certo impacto visual, por outro, não se observa eficácia na geração de informação ao usuário¹³, assim como não é provida interação com os nós tal como ocorre com outras ferramentas apresentadas.

Uma diferença da proposta de visualização desta dissertação frente a algumas inspirações de aplicação genérica para dados textuais se encontra no conteúdo que está a ser analisado, no caso, resultados de modelagem de tópicos, além do uso de dados de entidades ligadas a documentos presentes no tópico, de modo a facilitar o trabalho de especialista de domínio no intuito de interpretar os tópicos gerados. No caso das referências de visualização que aplicaram modelagem de tópicos, estas também foram importantes, mas acredita-se que seria possível incluir adições. Comparado ao *LDAvis*, a visualização aqui elaborada permite navegar diretamente a documentos de interesse ligados a cada tópico. Quanto a visualizações como *MetaToMaTo*, a diferença estaria em na adoção de princípios indicados por Edward Tufte (TUFTE, 2001), assim como Florian Windhager (WINDHAGER et al., 2018), que serão explicados no parágrafo seguinte. A adição de entidades tem também como inspiração visualizações desenvolvidas junto à ferramenta *History Lab*, que aplica modelagem de tópicos e extração de entidades. Apesar do objetivo

to be convenient for a wide range of common yet customizable visualizations.” (Retirado de: <<https://vega.github.io/vega/about/vega-and-d3/>>).

⁸ <<https://gephi.org/>>

⁹ Os gêneros mais evidentes de filtros que poderiam ter sido utilizados são, por exemplo, por quantidade de conexões (com quantas outras entidades cada nó tem coocorrência) ou quantidade de menções (em quantos documentos a entidade é mencionada). O estabelecimento de tais filtros tende a gerar grafos de menor complexidade, e, portanto, mais limpos.

¹⁰ Esse tipo de *layout* é voltado para grafos não-direcionados e de acordo com os seguintes princípios: distribuir os vértices uniformemente, minimizar cruzamento de arestas e reforçar simetria (FRUCHTERMAN; REINGOLD, 1991). Foram reajustados os seguintes parâmetros: Area=10000, Gravity=10.

¹¹ Esse *layout* segue os mesmos princípios do anterior, mas possui número maior de parâmetros (HU, 2005). Foram reajustados os seguintes parâmetros: Optimal Distance=10000, Relative Strength=180.

¹² Instruções estão localizadas em: <<https://gephi.org/users/quick-start/>>.

¹³ O *layout* Fruchterman-Reingold ao menos evidencia que algumas entidades são muito mais conectadas do que outras, ou seja, são mais centrais no aspecto de coocorrência em documentos.

principal deste em oferecer um mecanismo de busca a usuários na área de história das relações exteriores, foram desenvolvidas também visualizações em alguns dos acervos ali armazenados, tal como a coleção *Clinton E-mail* (ver Figura 17).

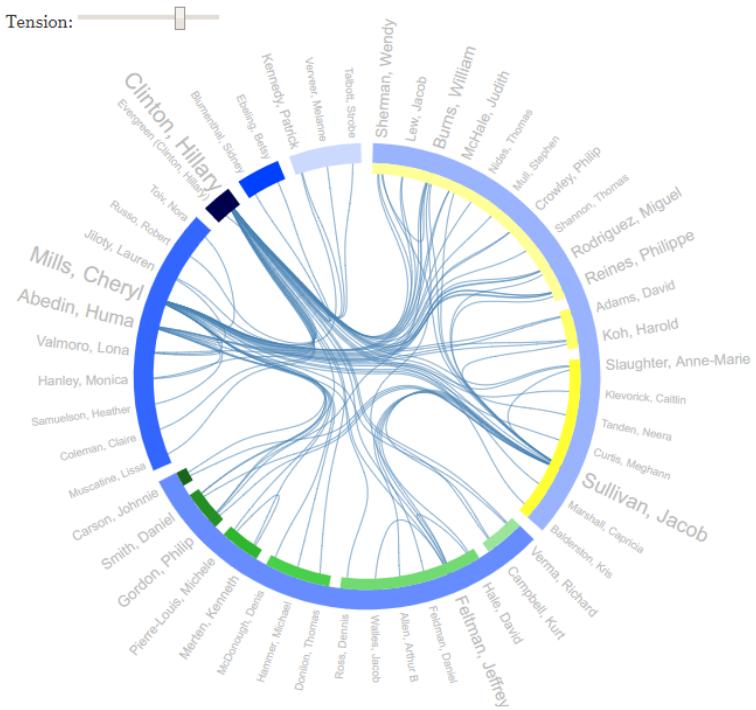


Figura 17 – Grafo com conexões entre entidades proveniente de mineração de dados em acervo de e-mails de Hillary Clinton (<http://history-lab.org/clinton>).

A ferramenta proposta é também dinâmica e interativa, além de expor elementos comuns a outras soluções, como termos-chave, documentos principais e hiperlinks de acesso a imagens originais dos documentos. Pode-se dizer que a ferramenta traz complementariedade às soluções pré-existentes, ou seja, que sua utilização em conjunto com outras soluções de visualização, somado também à aplicação de soluções estatísticas, tende a gerar aumento de conhecimento sobre o *corpus* e os tópicos gerados. Como já afirmado antes, a visualização *100 data stories*, desenvolvida por Christian Laesser, foi essencial para a elaboração desta ferramenta. Por outro lado, duas considerações foram feitas de modo que o resultado final tivesse um perfil mais minimalista. A razão para isso está no público-alvo almejado: enquanto que o *100 data stories* está voltado para categorias mais abrangentes de usuários, esta proposta visa principalmente especialistas de domínio¹⁴. Em primeiro lugar, levou-se em conta o conceito de *data-ink ratio*, segundo o qual deve-se buscar a maximização da proporção de dados por “tinta” utilizada. Em outras palavras, uma grande parcela da tinta em um gráfico deve corresponder à informação relativa a dados, com a tinta se alterando conforme o dado muda. Essa parcela da visualização

¹⁴ Tal como afirma Windhager et al., interfaces voltadas a especialistas devem estar centradas em investigação e curadoria, enquanto que no caso de usuários casuais, estas devem ser principalmente envolventes de modo a capturar sua atenção (WINDHAGER et al., 2018).

ligada aos dados em si representa sua parte principal, que não pode ser retirada, enquanto que o restante da tinta utilizada pode ser considerado redundante (TUFTE, 2001)¹⁵. Em segundo lugar, interações devem ser usadas com o devido cuidado para não gerarem ruído, desviando a atenção da informação que de fato se deseja passar. Estudos enfatizam a possibilidade de desvio de atenção aos dados de interesse quando técnicas de interação buscam alimentar a curiosidade dos usuários, instigando-os a explorá-los, mas sem um processo os guie (WINDHAGER et al., 2018; HINRICHSH; SCHMIDT; CARPENDALE, 2008).

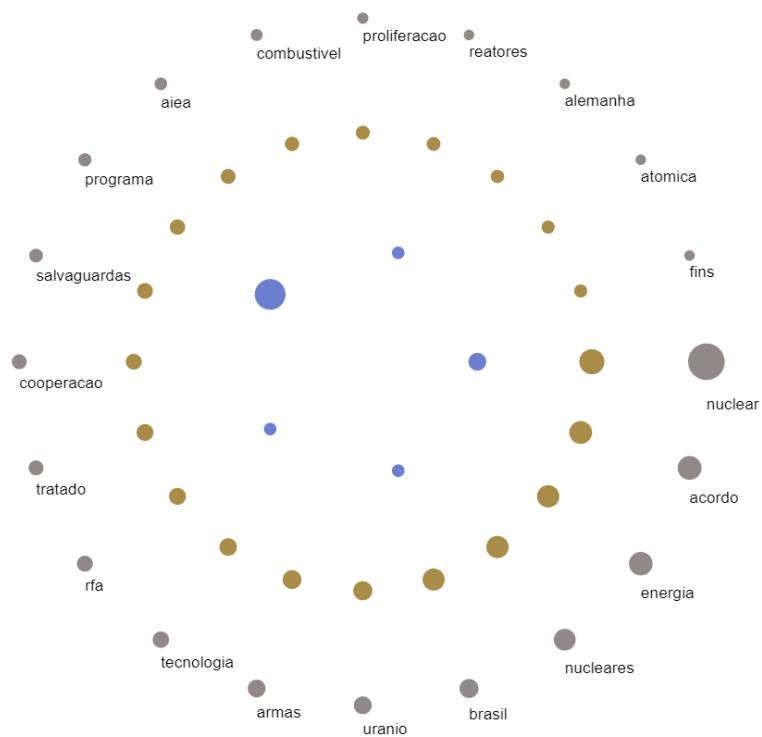


Figura 18 – Visualização em seu formato inicial, sem interação.

O grafo é composto por três camadas: a camada externa, cujos nós representam 20 termos de maior probabilidade de ocorrência para o tópico selecionado, uma camada intermediária representando 20 documentos de maior *score* frente ao tópico e uma camada interna ao centro, que representa as entidades mencionadas nesses 20 documentos. Cada camada foi organizada em forma de círculos com elementos equidistantes, seguindo tanto o *layout* da visualização de Christian Laesser quando ao exemplo contido no projeto *History Lab*. Neste último caso no entanto, em vez da exposição de arestas que conectam camadas distintas que representam diferentes tipos de informação, há uma única camada cujas arestas representam inter-relações de um mesmo tipo de informação.

¹⁵ Nas palavras de Edward Tufte: “A large share of ink on a graphic should present data-information, the ink changing as the data change. Data-ink is the non-erasable core of a graphic, the non-redundant ink arranged in response to variation in the numbers represented” (TUFTE, 2001).

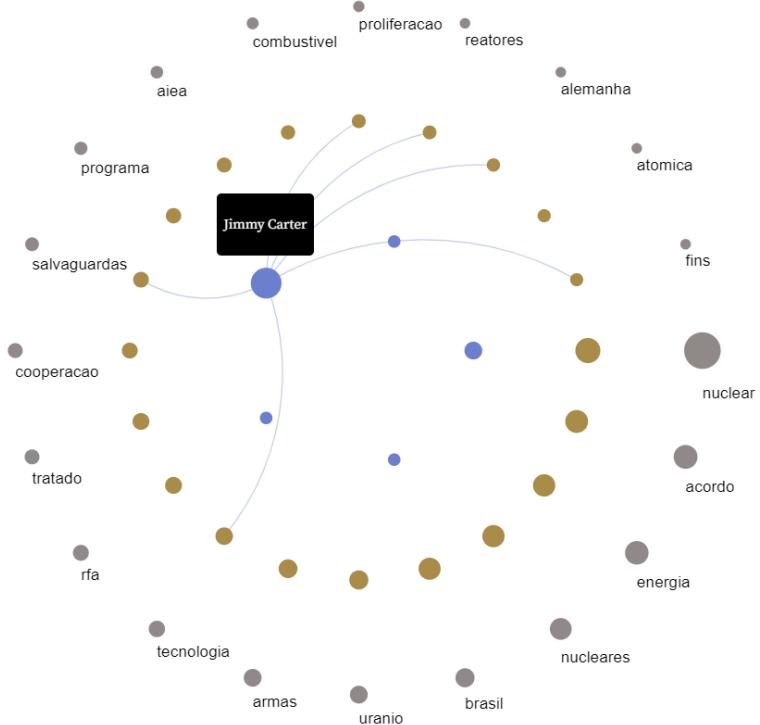


Figura 19 – Visualização ao aplicar a função *mouse over*, ativada quando o cursor do mouse passa por um nó do grafo.

No campo superior esquerdo da ferramenta, o usuário pode escolher um tópico gerado pela aplicação do LDA na coleção do CPDOC. A cada seleção, o usuário pode perceber alterações nas combinações de documentos, entidades e palavras-chave. Quanto à visualização em si, a principal funcionalidade de interação acrescentada foi o *mouseover*, que é ativado toda vez que o usuário passa o cursor do mouse sobre cada nó. A ativação resulta em alguns efeitos. 1) gera uma janela (*pop-up*) que se abre ao lado do nó selecionado e contém informação sobre o mesmo. 2) realça conexões com nós de outras camadas, representadas por arestas. No caso da camada interna, as arestas conectadas mostram quais entidades foram mencionadas no documento realçado. Quanto à camada externa, mostra quais dos termos estão presentes no documento.

A segunda funcionalidade de interação é ativada com o clique do mouse sobre qualquer um dos nós da camada intermediária (documentos). Ao fazê-lo, é acionado um *hiperlink* para o conteúdo completo e original (digitalizado) do documento. Esse novo endereço eletrônico possui o conteúdo do dossiê que contém o documento, cujo acesso é disponibilizado no *hiperlink* “*Ver Documento*”. Ao clicar nesse *hiperlink*, o usuário deve navegar até o documento relativo ao nó (o que foi exposto na janela resultante do *mouseover*), para assim ter acesso ao conteúdo digitalizado e completo relativo ao documento. O primeiro *hiperlink* é pertencente ao CPDOC, o mesmo centro responsável pela conservação dos documentos físicos, enquanto que o segundo é mantido por um prestador de serviços do CPDOC.

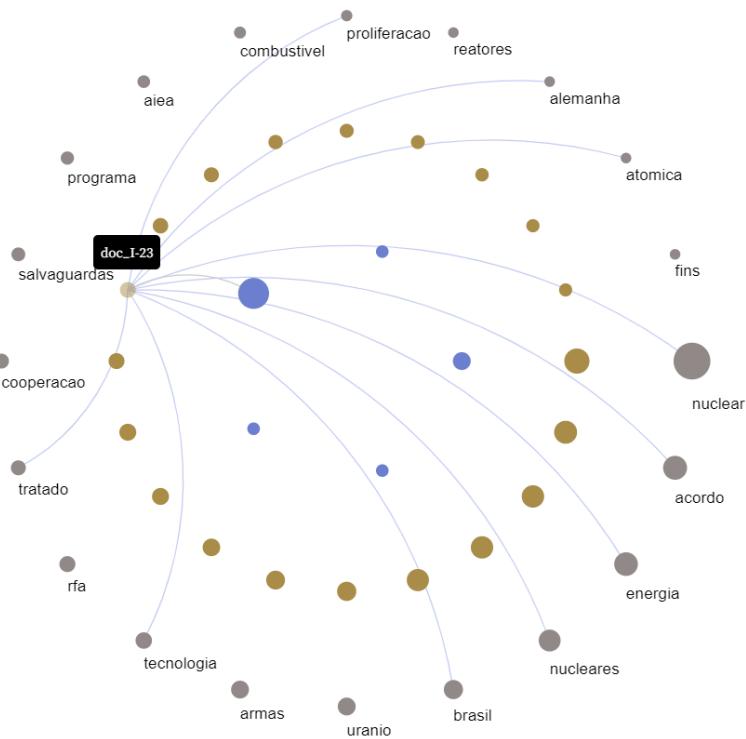


Figura 20 – Visualização ao aplicar a função *mouse over* sobre um nó relativo a documento (camada intermediária).

① Not secure | fgv.br/cpdoc/acervo/arquivo-pessoal/AAS/textual/documentos-referentes-a-atuacao-de-azeredo-da-silveira-como-ministro-das-relacoes-exteriores-na-assinatura-e-implementac...

Documentos referentes à atuação de Azeredo da Silveira, como ministro das Relações Exteriores, na assinatura e implementação do Acordo de Cooperação Nuclear Brasil-Alemanha, em 1975. O dossiê é consti...

[Sobre a consulta](#)

[f](#) [o](#) [g](#) [o](#) [t](#)

Manuscrito

Identificação:
Classificação: [AAS mre pn 1974.08.15](#)
Série: [mre - Ministro das Relações Exteriores](#)
Subsérie: [pn - Política nuclear](#)
Data de produção: [15.08.1974 a 06.02.1979 \(Data certa\)](#)

Quantidade de documentos: 118 (657 folhas) [Ver Documento](#)

Arquivo:
• [Antônio Azeredo da Silveira \(AAS\)](#) [Saiba mais](#)

Resumo:
Documentos referentes à atuação de Azeredo da Silveira, como ministro das Relações Exteriores, na assinatura e implementação do Acordo de Cooperação Nuclear Brasil-Alemanha, em 1975. O dossiê é constituído, basicamente, de correspondência abordando as seguintes questões: negociações no âmbito da política nuclear para a assinatura do Acordo, em junho de 1975; Acordo Trilateral de Salvaguardas entre Alemanha, Brasil e Agência Internacional de Energia Atómica (AIEA), exigido como complemento ao Acordo Brasil-Alemanha; possível suspensão da implementação do Acordo, em 1977, por parte da Alemanha, em decorrência da pressão dos Estados Unidos, que acusava o Brasil de não ter assinado o Tratado de Não Proliferação - TNP. Inclui ainda documentos sobre conversações entre Brasil-EUA, Brasil-Alemanha-EUA, e Alemanha-EUA, a respeito da questão da não proliferação de armas nucleares.

Para enviar uma colaboração ou guardar este conteúdo em suas pesquisas [clique aqui](#) para fazer o login.

[Voltar](#) [Nova consulta](#)

Figura 21 – Página do CPDOC para o qual é feito o redirecionamento, contendo acesso à versão digitalizada do documento.

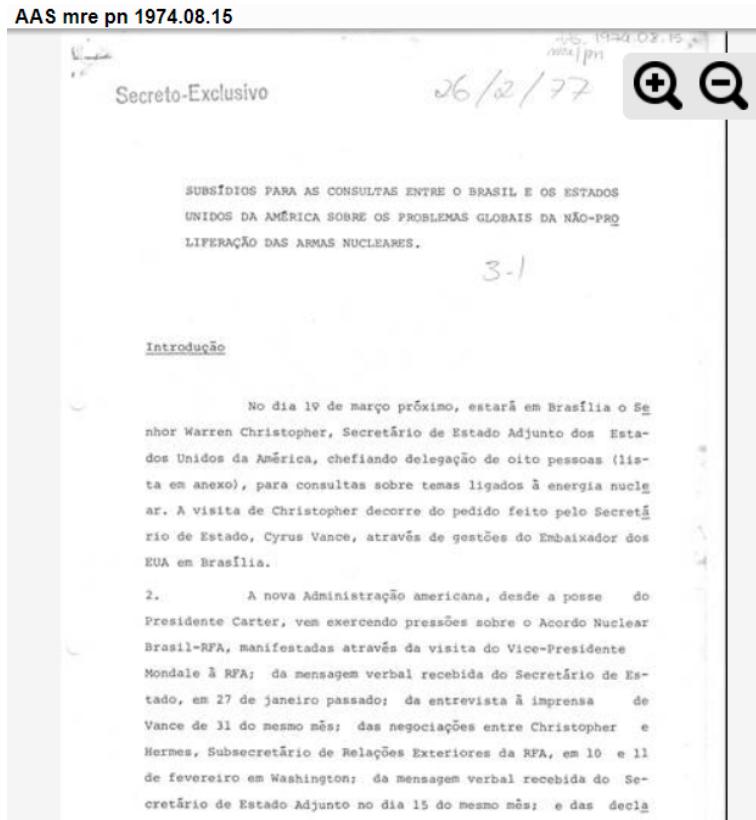


Figura 22 – Página contendo versão digitalizada do documento acessada por meio do endereço eletrônico anterior. Tais páginas são mantidas por prestador de serviços ao CPDOC.

Como explicado na seção 5.2, os raios dos círculos são representados pelos *scores* no caso de documentos e *tokens*, ou pela contagem de ocorrência nos documentos, no caso de pessoas. Além disso, no caso de *tokens* e pessoas, o raio de cada círculo foi reescalado de modo a reduzir discrepâncias muito elevadas dos tamanhos, evitando assim que casos extremos gerasse ruído na visualização¹⁶. Para a camada externa, que representa termos (*tokens*), o raio é a raiz quadrada do *score* de cada termo relativo ao tópico. Quanto à camada interna, localizada ao centro do gráfico, o raio é a raiz quadrada da contagem de menções de cada entidade frente aos 20 documentos de maior *score* frente a cada tópico. Cada um dos valores obtidos em tais camadas necessitou de um fator multiplicativo para que os nós estivessem devidamente representados no gráfico.

Por fim, para a seleção de cores, foi determinado que deveria ser enfatizado o contraste entre as camadas de nós. Uma ferramenta que atende esse intuito se chama *iWantHue*¹⁷, que possui funcionamento semelhante ao *ColorBrewer*¹⁸. Os valores hexadecimais das três paletas de cores escolhidas são: #a88c47, #6c7fce e #908987, respectivamente,

¹⁶ A ausência dessas transformações poderia resultar em visualizações pouco claras, por exemplo, em nós muito pequenos, no caso de valores consideravelmente abaixo da média, ou em nós demasiadamente grandes, no caso de valores muito acima.

¹⁷ <<https://medialab.github.io/iwanthue/>>.

¹⁸ <<https://colorbrewer2.org/>>.

em escalas de marrom, azul e cinza. Entre as variáveis selecionadas no programa, foi ativada a opção “Improve for the colorblind”, de modo a tornar a experiência de usuário disponibilizada pela visualização independente do fato deste ser daltônico ou não. Já o método de *clusterização* utilizado foi o “soft (k-means)”. A escolha do fundo branco, assim como o fato de as arestas serem mostradas apenas por meio de ativação da função *mouseover* atende ao conceito de *data-ink ratio* (TUFTE, 2001).

De modo a fornecer maior amplitude de compreensão da coleção, foi elaborada uma visualização auxiliar, que fornece um gráfico de *heatmap* do *score* entre documentos, representados no eixo vertical, e tópicos, representados no eixo horizontal. Tal desenvolvimento completa as camadas de granularidade propostas por Windhager et al. (WINDHAGER et al., 2018). A visualização foi construída por meio do *seaborn*¹⁹, pacote do *Python*. Cada ponto é um valor de *score* entre 0 (nenhuma correspondência) e 1 (correspondência máxima) entre documentos e tópicos. A escala de cores²⁰ vai de branco (valor 0) e preto (valor 1), tendo sido escolhida por ser uma cor neutra. A decisão de reverter a escala, ou seja, de não ser preto até branco, se justifica pelo fato de ser mais apropriado para a impressão, por realçar melhor os *scores*, assim como pelo fato de corresponder ao conceito de *data-to-ink ratio*. A escala de cores relativa ao *score* possui uma legenda no canto direito do gráfico, de modo a melhorar a compreensão do mesmo para o usuário. Os tópicos foram reordenados de modo que os *scores* fossem concentrados ao lado esquerdo, por ordem de magnitude. Tal ordenamento foi realizado com a captura da mediana de *score* dos 20 documentos com maior valor por tópico. O gráfico foi alocado acima da visualização principal, de modo que o usuário possa primeiramente ter uma visão geral da totalidade dos tópicos e documentos.

¹⁹ <<https://seaborn.pydata.org/>>.

²⁰ A escala é denominada no *seaborn* como “gist_gray_r”

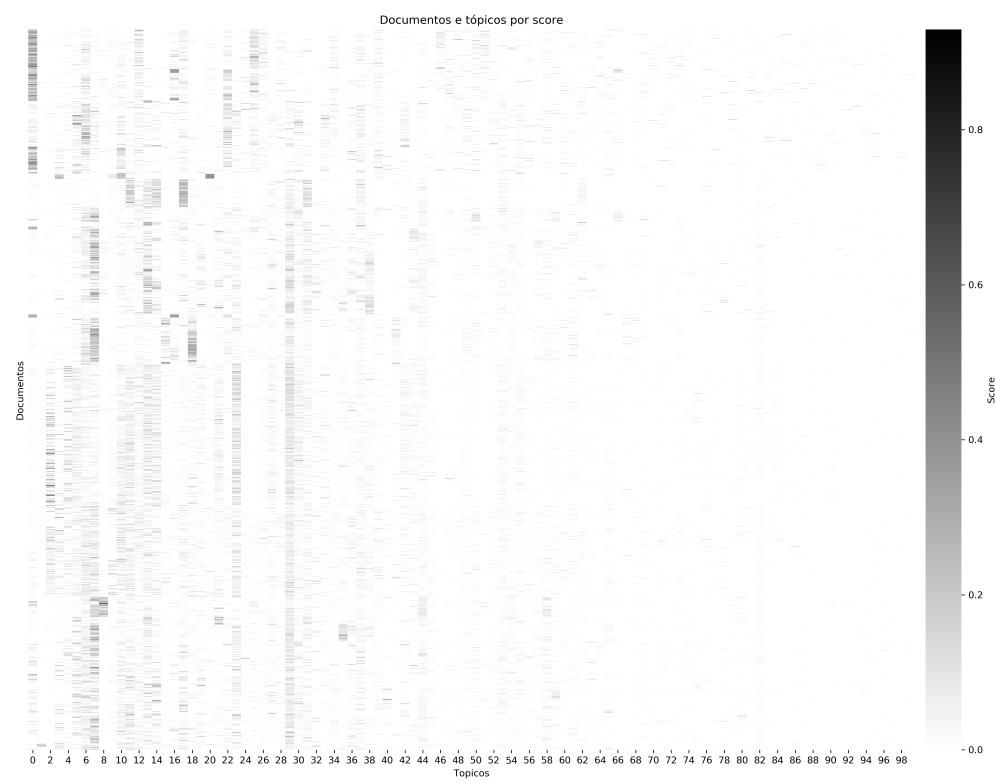


Figura 23 – *Heatmap* representando o *score* entre documentos e tópicos.

Considerações finais

O paradigma atual de gestão de arquivos históricos trouxe novos desafios que demandam a utilização de recursos modernos. Desse modo, a explosão da quantidade de dados produzidos na era da informação deve ser compensada pela adoção de ferramentas de *Big Data* e Aprendizado de Máquina, de forma a preservar a qualidade do trabalho de gestão de arquivos. Essa área se encontra em constante expansão e a quantidade e variabilidade de tecnologias disponíveis muda a todo instante.

Neste trabalho, foi apresentado um ciclo completo de aplicação de ferramentas para realizar extração e organização de informação a partir de um conjunto de mais de dez mil documentos históricos relativos à política externa brasileira na década de 1970. Tais etapas foram necessárias para chegar à solução de visualização proposta nesta dissertação, cuja função principal foi facilitar o trabalho de interpretação de tópicos gerados pelo processo de modelagem de tópicos. A visualização criada permite conectar rapidamente um tópico específico aos principais documentos e personagens ligados ao mesmo por meio de interatividade, sendo minimalista em elementos visuais, porém informativo o suficiente para guiar o usuário. A possibilidade de acessar a versão digitalizada de documentos por meio de nós do grafo, que redirecionam a endereço eletrônico específico, ajuda o processo de análise exploratória e interpretação dos tópicos. Além disso, como o redirecionamento é ligado à própria organização responsável pelo armazenamento dos documentos, há uma valorização da mesma, ao aumentar o acesso de usuários.

As ferramentas utilizadas neste trabalho são apenas uma amostra dentro de um mosaico diverso de opções disponíveis, sendo que cada uma delas tende a se adequar melhor a um contexto específico. Diversas melhorias podem ser feitas para atingir um resultado mais satisfatório, principalmente na parte de pré-processamento e modelagem de tópicos. Por exemplo, poderia ter sido feita uma pré-filtragem das imagens que geraram o *corpus* com uso de ferramentas de *computer vision*, tal como o *Google Cloud Vision API*. Desse modo, seria possível retirar páginas de documentos que referenciam elementos não-textuais como gráficos, desenhos, inclusive manuscritos, que, apesar destes serem elementos textuais, não há solução trivial por meio de *OCR* para fazer o tratamento dos mesmos.

Na parte de pré-processamento, a limpeza de dados, que envolvia correção ou recuperação de palavras, poderiam ter sido testadas ferramentas como *soundex* ou *enchant.predict* de modo a identificar probabilisticamente a versão correta de termos que não puderam ser identificados, tendo o devido cuidado e estudo sobre sua correta utilização, levando em conta especificidades do *corpus* trabalhado. Além disso, para a modelagem de

tópicos, poderiam ter sido realizados testes com diferentes versões dos documentos, por exemplo, com filtragem de elementos não-dicionarizados. Para isso, seria necessário incorporar um dicionário de termos específicos ao *corpus*, principalmente nomes de entidades, que não necessariamente estariam previamente inclusos em dicionários de ferramentas que fazem a identificação de termos não-dicionarizados, tal como o *enchant*, biblioteca do *python*.

No campo da visualização desenvolvida, uma adição relevante estaria no aumento da interatividade. Dado que o usuário almejado seria o especialista de domínio, pode-se considerar oportuno que este tivesse flexibilidade na escolha do número de nós a serem expostos em vez de estar submetido a um número pré-determinado, assim como na escolha do tópico diretamente por meio da visualização auxiliar. Do mesmo modo, poderia ter também esse poder de escolha sobre o número de palavras expostas. Por fim, uma implementação importante seria ter um sistema que permita rapidamente refazer a modelagem de tópicos. Tais acréscimos tenderiam a aumentar a capacidade de análise exploratória, além de personalizar a experiência de usuário.

Muitos avanços ainda devem ser gerados na área de modelagem de tópicos, assim como no campo de visualização desses resultados. É possível que o contínuo desenvolvimento do *vega-lite* torne interessante futuramente que ferramentas semelhantes sejam desenvolvidas sob essa linguagem em detrimento do *D3*, caso a mesma se torne mais flexível em gêneros possíveis de visualização. Entretanto, o *D3* ainda pode ser considerado uma linguagem de alta relevância, pois apesar de seu alto nível de complexidade de manipulação, permite idealizar formas inovadoras de visualização. Além disso, devido a efeitos de escala, ou seja, da grande quantidade de usuários, há disponibilidade de muitas referências para formular inovações.

Todos os resultados desenvolvidos estão devidamente documentados no início da parte 2, relativa aos passos metodológicos. De todo modo, os endereços eletrônicos principais podem ser novamente expostos para consulta:

- Repositório com códigos elaborados em *Python* e bases de dados geradas: <<https://github.com/Marcelobbr/thesis>>
- Visualização gerada em *D3.js*: <<https://observablehq.com/@marcelobbr/thesis-visualization>>

Referências

ALENCAR, A. B.; OLIVEIRA, M. C. F. de; PAULOVICH, F. V. Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 2, n. 6, p. 476–492, 2012. Citado na página [23](#).

ALSUMAIT, L. et al. Topic significance ranking of lda generative models. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2009. p. 67–82. Citado 4 vezes nas páginas [36](#), [37](#), [41](#) e [61](#).

BEAL, M. J.; GHAHRAMANI, Z.; RASMUSSEN, C. E. The infinite hidden markov model. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2002. p. 577–584. Citado na página [34](#).

BICK, E. The parsing system palavras. *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, University of Arhus, 2000. Citado na página [44](#).

BISCHOF, J.; AIROLDI, E. M. Summarizing topical content with word frequency and exclusivity. In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. [S.l.: s.n.], 2012. p. 201–208. Citado 2 vezes nas páginas [37](#) e [61](#).

BLEI, D. *Probabilistic Topic Models: Surveying a suite of algorithms that offer a solution to managing large document archives*. [S.l.]: Apr, 2012. Citado 7 vezes nas páginas [15](#), [27](#), [30](#), [31](#), [32](#), [33](#) e [38](#).

BLEI, D.; LAFFERTY, J. Correlated topic models. *Advances in neural information processing systems*, Citeseer, v. 18, p. 147, 2006. Citado na página [34](#).

BLEI, D. M.; LAFFERTY, J. D. Dynamic topic models. In: ACM. *Proceedings of the 23rd international conference on Machine learning*. [S.l.], 2006. p. 113–120. Citado na página [34](#).

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado na página [27](#).

CHANAY, A. J.-B.; BLEI, D. M. Visualizing topic models. In: *Sixth international AAAI conference on weblogs and social media*. [S.l.: s.n.], 2012. Citado 4 vezes nas páginas [16](#), [39](#), [61](#) e [82](#).

CHANG, J. et al. Reading tea leaves: How humans interpret topic models. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2009. p. 288–296. Citado na página [36](#).

CHUANG, J. et al. Topic model diagnostics: Assessing domain relevance via topical alignment. In: *International Conference on Machine Learning*. [S.l.: s.n.], 2013. p. 612–620. Citado 2 vezes nas páginas [37](#) e [41](#).

CHUANG, J.; MANNING, C. D.; HEER, J. Termite: Visualization techniques for assessing textual topic models. In: ACM. *Proceedings of the international working conference on advanced visual interfaces*. [S.l.], 2012. p. 74–77. Citado 3 vezes nas páginas [17](#), [41](#) e [84](#).

- CHUANG, J. et al. Topiccheck: Interactive alignment for assessing topic model stability. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2015. p. 175–184. Citado 2 vezes nas páginas [37](#) e [61](#).
- CUI, W. et al. Textflow: Towards better understanding of evolving topics in text. *IEEE transactions on visualization and computer graphics*, IEEE, v. 17, n. 12, p. 2412–2421, 2011. Citado na página [42](#).
- DEERWESTER, S. et al. Indexing by latent semantic analysis. *Journal of the American society for information science*, Wiley Online Library, v. 41, n. 6, p. 391–407, 1990. Citado 2 vezes nas páginas [28](#) e [29](#).
- DIMAGGIO, P.; NAG, M.; BLEI, D. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, Elsevier, v. 41, n. 6, p. 570–606, 2013. Citado na página [27](#).
- DOU, W. et al. Paralleltopics: A probabilistic approach to exploring document collections. In: IEEE. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. [S.l.], 2011. p. 231–240. Citado na página [42](#).
- FRUCHTERMAN, T. M.; REINGOLD, E. M. Graph drawing by force-directed placement. *Software: Practice and experience*, Wiley Online Library, v. 21, n. 11, p. 1129–1164, 1991. Citado na página [64](#).
- GARDNER, M. J. et al. The topic browser: An interactive tool for browsing topic models. In: WHISTLER CANADA. *NIPS Workshop on Challenges of Data Visualization*. [S.l.], 2010. v. 2. Citado 4 vezes nas páginas [16](#), [39](#), [61](#) e [81](#).
- GORG, C. et al. Jigsaw meets blue iguanodon—the vast 2007 contest. In: IEEE. *2007 IEEE Symposium on Visual Analytics Science and Technology*. [S.l.], 2007. p. 235–236. Citado 4 vezes nas páginas [16](#), [25](#), [39](#) e [82](#).
- GORG, C. et al. Visual analytics with jigsaw. In: IEEE. *2007 IEEE Symposium on Visual Analytics Science and Technology*. [S.l.], 2007. p. 201–202. Citado na página [40](#).
- GRIFFITHS, T. L. et al. Integrating topics and syntax. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2005. p. 537–544. Citado na página [32](#).
- HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado na página [25](#).
- HAVRE, S.; HETZLER, B.; NOWELL, L. Themeriver: Visualizing theme changes over time. In: IEEE. *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*. [S.l.], 2000. p. 115–123. Citado 3 vezes nas páginas [17](#), [42](#) e [85](#).
- HILLARD, D.; PURPURA, S.; WILKERSON, J. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, Taylor & Francis, v. 4, n. 4, p. 31–46, 2008. Citado na página [35](#).
- HINRICHES, U.; SCHMIDT, H.; CARPENDALE, S. Em dialog: Bringing information visualization into the museum. *IEEE transactions on visualization and computer graphics*, IEEE, v. 14, n. 6, p. 1181–1188, 2008. Citado na página [66](#).

HOFFMAN, M.; BACH, F. R.; BLEI, D. M. Online learning for latent dirichlet allocation. In: *advances in neural information processing systems*. [S.l.: s.n.], 2010. p. 856–864. Citado 2 vezes nas páginas 31 e 32.

HOFMANN, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, Springer, v. 42, n. 1-2, p. 177–196, 2001. Citado na página 30.

HOFMANN, T. Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*, 2013. Citado na página 29.

HU, Y. Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, Redwood City, Ca.: Advanced Book Program, Addison-Wesley Pub. Co., c1990-, v. 10, n. 1, p. 37–71, 2005. Citado na página 64.

KEIM, D. et al. Visual analytics: Definition, process, and challenges. In: *Information visualization*. [S.l.]: Springer, 2008. p. 154–175. Citado na página 25.

LAESSER, C. *100 Data Stories*. 2017. <<https://projects.christianlaesser.com/100-data-stories>>. Accessed: 2019-10-01. Citado na página 62.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. *Discourse processes*, Taylor & Francis, v. 25, n. 2-3, p. 259–284, 1998. Citado 2 vezes nas páginas 28 e 29.

LIU, S. et al. Bridging text visualization and mining: A task-driven survey. *IEEE transactions on visualization and computer graphics*, IEEE, v. 25, n. 7, p. 2482–2504, 2018. Citado 2 vezes nas páginas 24 e 37.

LOVINS, J. B. Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics*, v. 11, n. 1-2, p. 22–31, 1968. Citado na página 51.

LUCAS, C. et al. Computer-assisted text analysis for comparative politics. *Political Analysis*, Cambridge University Press, v. 23, n. 2, p. 254–277, 2015. Citado na página 34.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. *Natural Language Engineering*, Cambridge university press, v. 16, n. 1, p. 100–103, 2010. Citado na página 51.

MCAULIFFE, J. D.; BLEI, D. M. Supervised topic models. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2008. p. 121–128. Citado na página 34.

MEEKS, E. *D3.js in action: data visualization with JavaScript*. [S.l.]: Manning Publications, 2018. Citado na página 44.

MIGON, H. S.; GAMERMAN, D.; LOUZADA, F. *Statistical inference: an integrated approach*. [S.l.]: Chapman and Hall/CRC, 2014. Citado na página 31.

MIMNO, D. et al. Optimizing semantic coherence in topic models. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the conference on empirical methods in natural language processing*. [S.l.], 2011. p. 262–272. Citado 4 vezes nas páginas 37, 41, 42 e 61.

- NAGY, G.; NARTKER, T. A.; RICE, S. V. Optical character recognition: An illustrated guide to the frontier. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Document Recognition and Retrieval VII*. [S.l.], 1999. v. 3967, p. 58–69. Citado na página 48.
- RAHM, E.; DO, H. H. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, v. 23, n. 4, p. 3–13, 2000. Citado na página 49.
- RIBEIRO, M.; MORELI, A.; SOUZA, R. R. Data mining with historical data. In: FGV/CPDOC - UNIRIO. *Anais do I Congresso Internacional em Humanidades Digitais no Rio de Janeiro*. [S.l.], 2018. p. 384–392. Citado 10 vezes nas páginas 24, 44, 46, 47, 49, 50, 51, 52, 53 e 54.
- RICE, S. V.; JENKINS, F. R.; NARTKER, T. A. *The fourth annual test of OCR accuracy*. [S.l.], 1995. Citado na página 47.
- ROBERTS, M. E.; STEWART, B. M.; AIROLDI, E. M. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, Taylor & Francis, v. 111, n. 515, p. 988–1003, 2016. Citado na página 35.
- SCHNEIDER, B. Visualização em multirresolução do fluxo de tópicos em coleções de texto. In: . [S.l.: s.n.], 2014. Citado 5 vezes nas páginas 16, 25, 39, 42 e 80.
- SIEVERT, C.; SHIRLEY, K. Ldavis: A method for visualizing and interpreting topics. In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. [S.l.: s.n.], 2014. p. 63–70. Citado 7 vezes nas páginas 17, 24, 40, 41, 42, 61 e 83.
- SNYDER, J. et al. Topic models and metadata for visualizing text corpora. In: *Proceedings of the 2013 NAACL HLT Demonstration Session*. [S.l.: s.n.], 2013. p. 5–9. Citado 4 vezes nas páginas 16, 39, 61 e 81.
- STEWART, B. M.; ZHUKOV, Y. M. Use of force and civil–military relations in russia: an automated content analysis. *Small Wars & Insurgencies*, Taylor & Francis, v. 20, n. 2, p. 319–343, 2009. Citado na página 35.
- STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. *Handbook of latent semantic analysis*, v. 427, n. 7, p. 424–440, 2007. Citado 3 vezes nas páginas 27, 28 e 32.
- TADDY, M. On estimation and selection for topic models. In: *Artificial Intelligence and Statistics*. [S.l.: s.n.], 2012. p. 1184–1193. Citado 3 vezes nas páginas 37, 41 e 61.
- TEH, Y. W. et al. Sharing clusters among related groups: Hierarchical dirichlet processes. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2005. p. 1385–1392. Citado na página 34.
- TUFTE, E. R. *The visual display of quantitative information*. [S.l.]: Graphics press Cheshire, CT, 2001. v. 2. Citado 3 vezes nas páginas 64, 66 e 70.
- WALLACH, H. M. Topic modeling: beyond bag-of-words. In: ACM. *Proceedings of the 23rd international conference on Machine learning*. [S.l.], 2006. p. 977–984. Citado na página 34.

WANG, X.-M. et al. A survey of visual analytic pipelines. *Journal of Computer Science and Technology*, Springer, v. 31, n. 4, p. 787–804, 2016. Citado 3 vezes nas páginas [15](#), [25](#) e [36](#).

WANNER, F. et al. State-of-the-art report of visual analysis for event detection in text data streams. In: *EuroVis (STARs)*. [S.l.: s.n.], 2014. Citado na página [23](#).

WARD, M. O.; GRINSTEIN, G.; KEIM, D. *Interactive data visualization: foundations, techniques, and applications*. [S.l.]: AK Peters/CRC Press, 2015. Citado 6 vezes nas páginas [16](#), [24](#), [25](#), [37](#), [61](#) e [62](#).

WEN, Z.; LIN, C.-Y. Towards finding valuable topics. In: SIAM. *Proceedings of the 2010 SIAM International Conference on Data Mining*. [S.l.], 2010. p. 720–731. Citado na página [41](#).

WINDHAGER, F. et al. Visualization of cultural heritage collection data: State of the art and future challenges. *IEEE transactions on visualization and computer graphics*, IEEE, v. 25, n. 6, p. 2311–2330, 2018. Citado 10 vezes nas páginas [15](#), [23](#), [38](#), [55](#), [58](#), [61](#), [64](#), [65](#), [66](#) e [70](#).

Anexos

ANEXO A – Referências de visualização

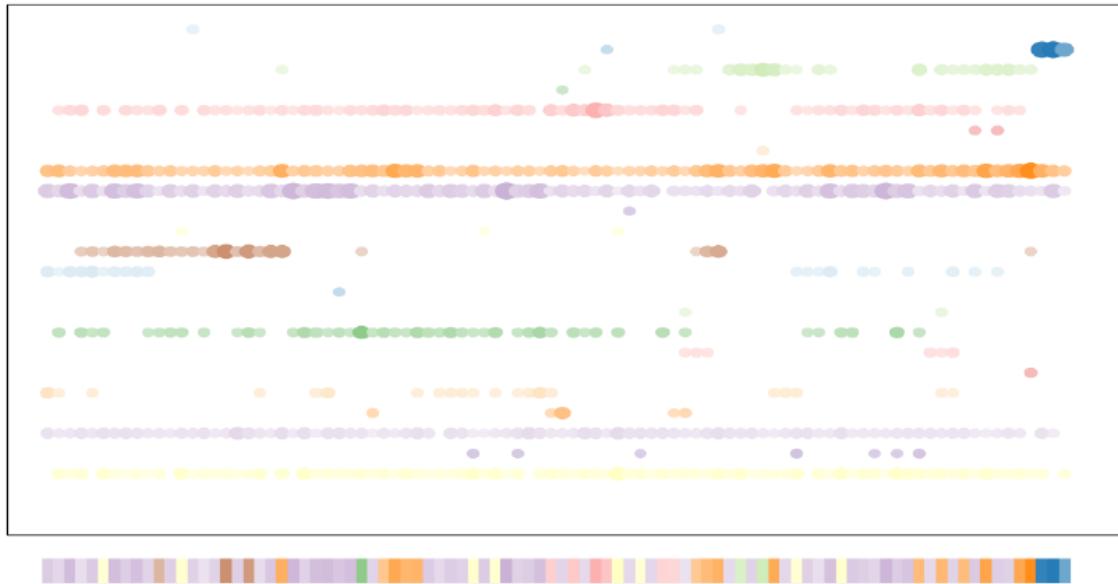


Figura 24 – Demonstração da ferramenta *Bubble Timeline*. Fazendo uso da própria explicação de Schneider: “no eixo horizontal está representado o tempo e no vertical estão os tópicos. Para cada tópico foi definida uma cor diferente. A variação de intensidade de cor em cada círculo do gráfico representa a variação dos valores entre 0 e 1 da probabilidade de ocorrência de cada tópico ao longo do tempo.” ([SCHNEIDER, 2014](#))

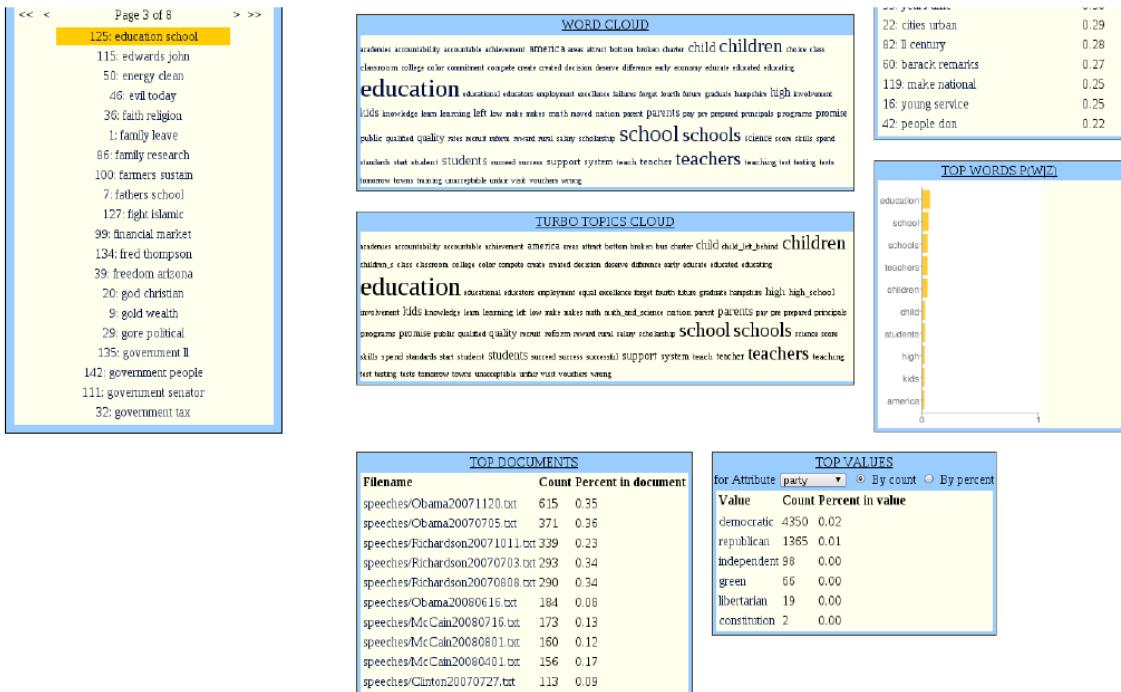


Figura 25 – Demonstração de uso da ferramenta *The Topic Browser*. Ao selecionar o tópico “125” (janela à esquerda), são expostas mais informações em outras janelas (GARDNER et al., 2010).

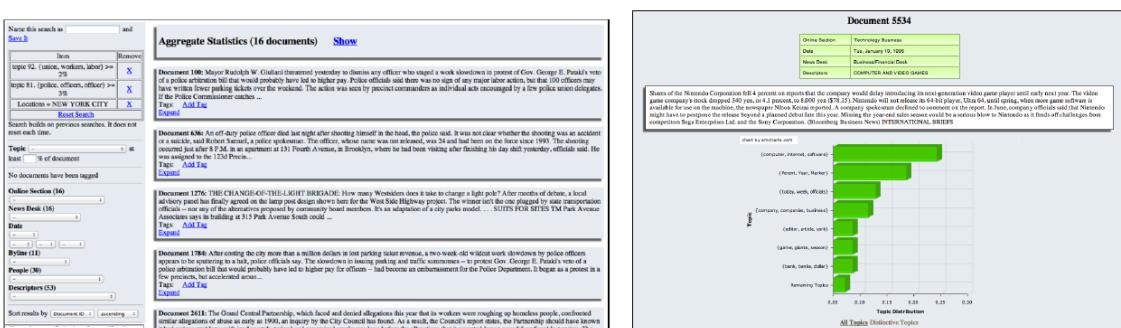


Figura 26 – Demonstração de uso da ferramenta *MetaToMaTo*. Na janela à esquerda é possível selecionar um tópico, enquanto que a janela à direita expõe os documentos relacionados (SNYDER et al., 2013).

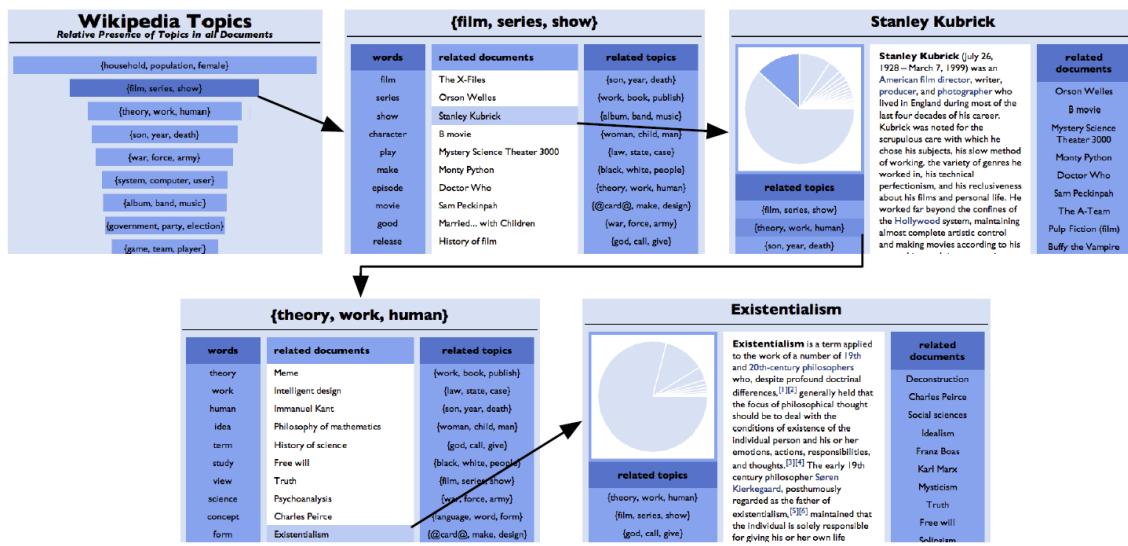


Figura 27 – Navegador desenvolvido por Chaney e Blei que foi testado na Wikipédia. Esta visualização expõe o processo de exploração dos dados por meio dessa ferramenta (CHANAY; BLEI, 2012).

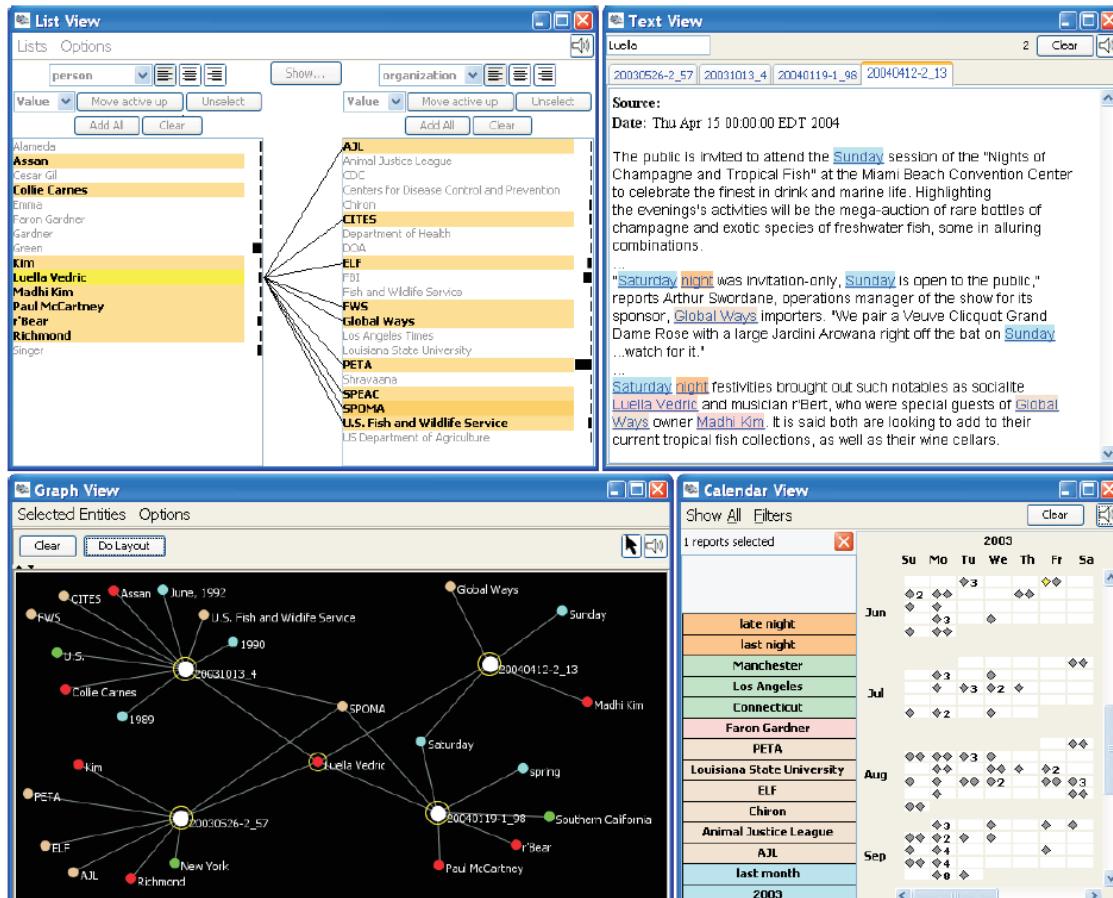


Figura 28 – Uma das visualizações possíveis por meio da ferramenta Jigsaw (GORG et al., 2007a).

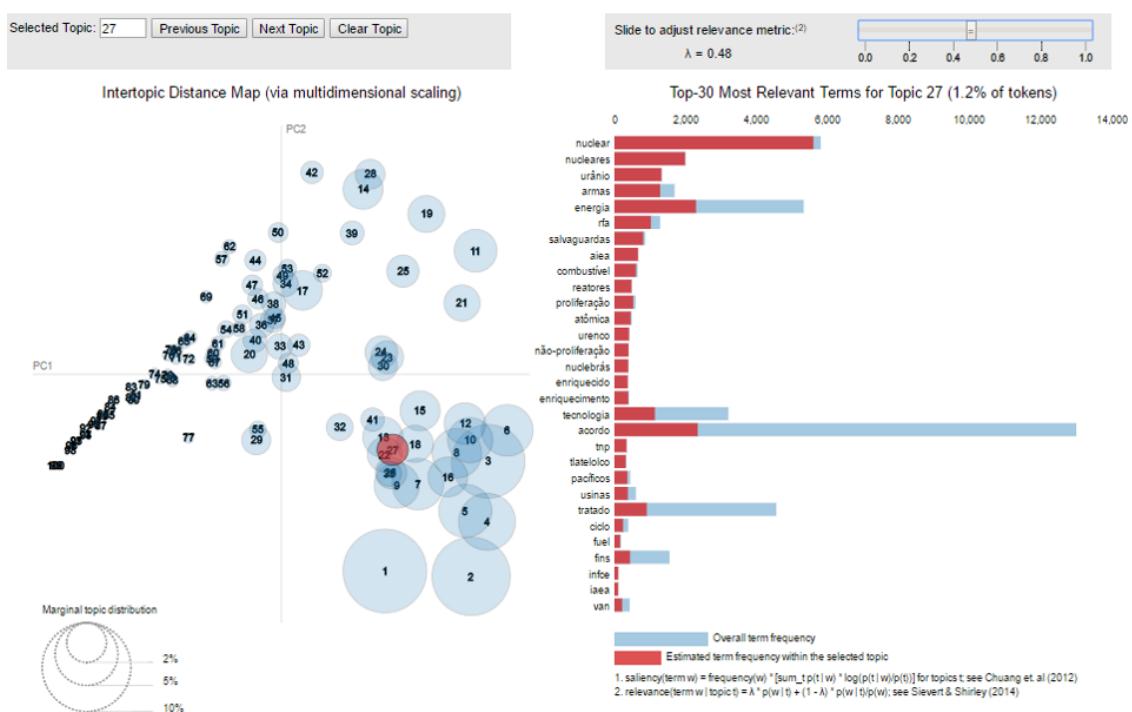


Figura 29 – Exemplo da ferramenta *LDAvis* ([SIEVERT; SHIRLEY, 2014](#)), neste caso aplicado ao próprio *corpus* do CPDOC.

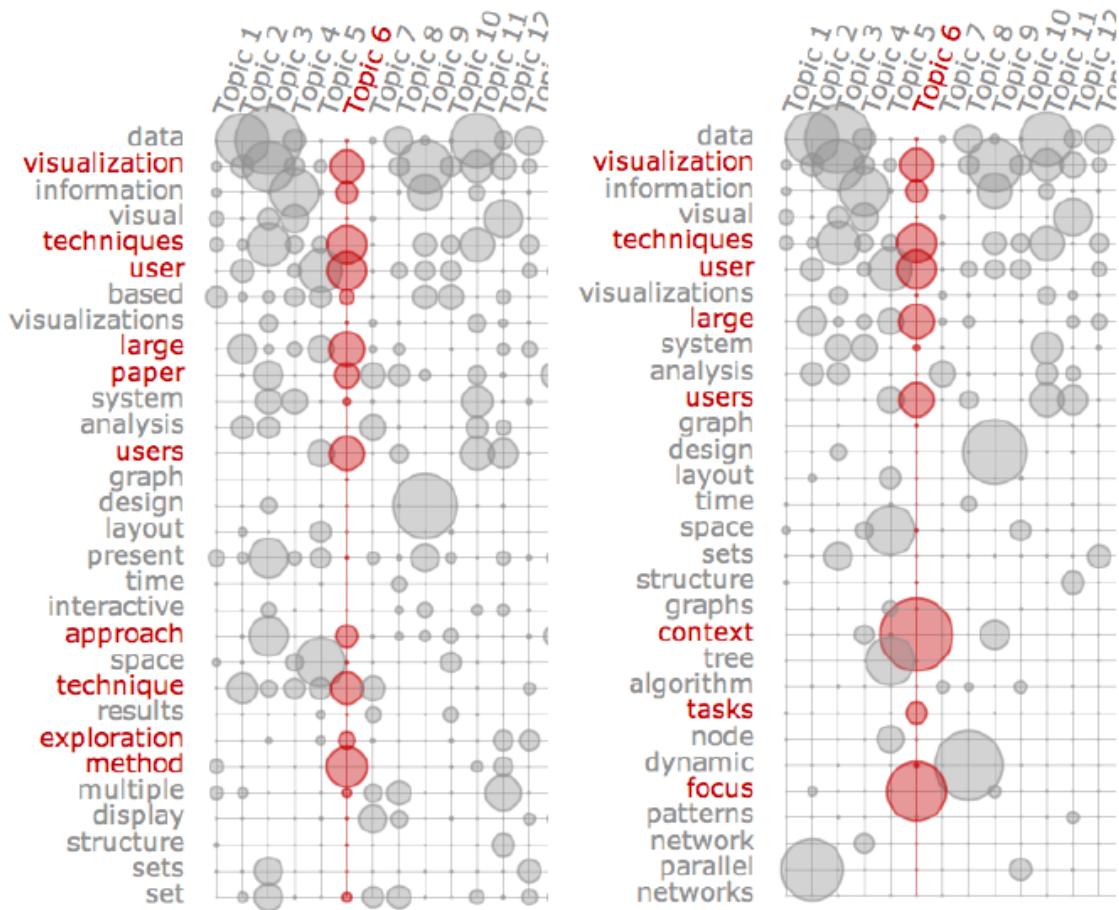


Figura 30 – Demonstração da ferramenta *Termite*. O lado esquerdo expõe os 30 termos mais frequentes, enquanto que o lado direito expõe os 30 mais salientes (CHUANG; MANNING; HEER, 2012).

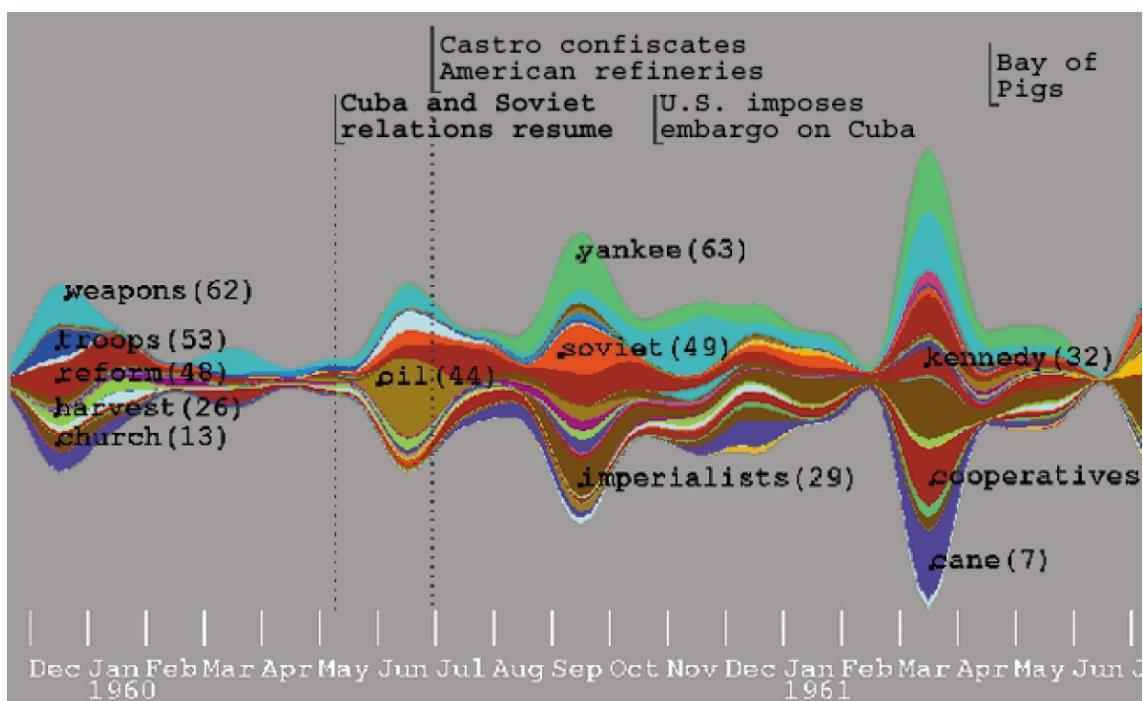


Figura 31 – Demonstração da ferramenta *ThemeRiver*, que utiliza a metáfora de um rio de modo a representar variações da importância de tópicos ao longo do tempo (HAVRE; HETZLER; NOWELL, 2000)

ANEXO B – Testes de visualização com a ferramenta Gephi

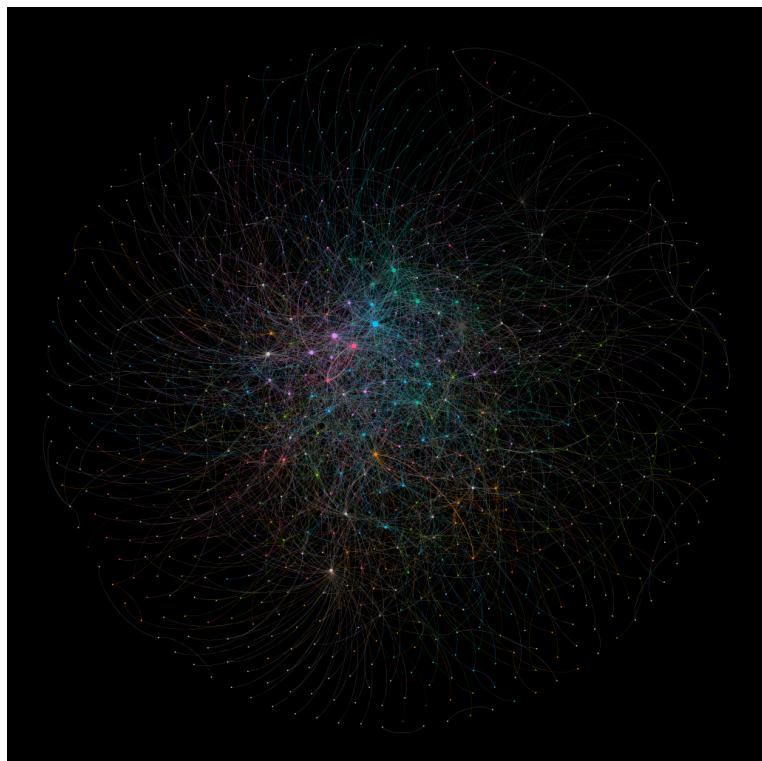


Figura 32 – Grafo gerado por meio do Gephi, *layout* Fruchterman Reingold.

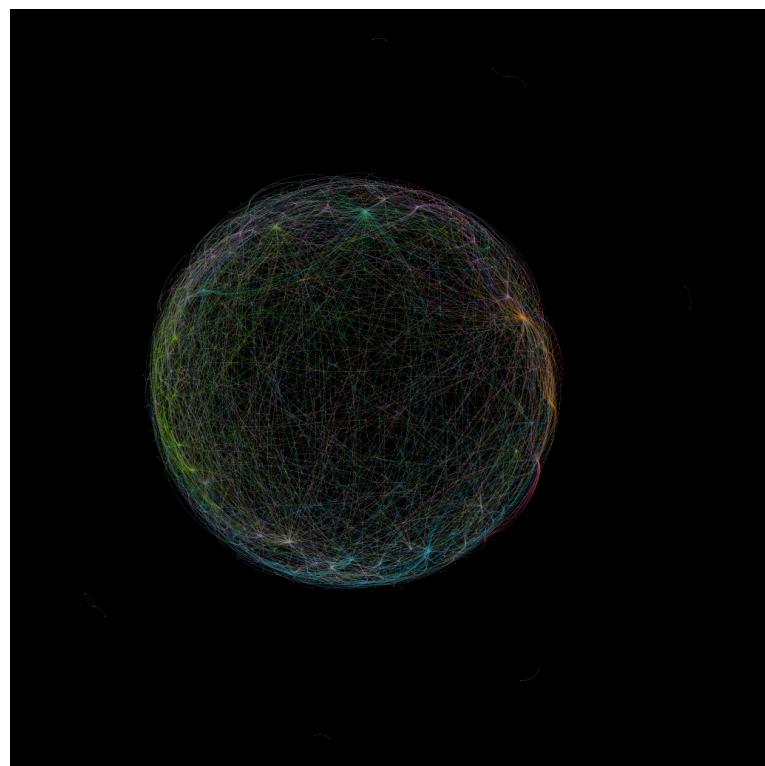


Figura 33 – Grafo gerado por meio do Gephi, *layout Yifanhu*.