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

•

Marcelo Bianchi Barata Ribeiro

•

Orientadora: Asla Medeiros e Sá

Coorientador: Renato Rocha Souza

Sumário da apresentação

- Contexto e objetivos
- Referenciais Teóricos e Estado da Arte
- Passos metodológicos
  - Ferramentas
  - Dados
  - Processamento de dados
  - Visualização
- Considerações Finais

Contexto

- 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.

Objetivos

- 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 almejada está na conceituação de uma nova ferramenta que atende a boas práticas indicadas por autores da área de visualização de dados, fazendo uso de uma linguagem de programação que forneça o máximo de flexibilidade.

Referenciais teóricos e estado da arte

Modelagem de Tópicos

- LSI (Latent Semantic Indexing)
- pLSI (probabilistic LSI)
- LDA (Latent Dirichlet Allocation)

LSI

- O LSI foi publicado em 1990 com o intuito de automatizar e aperfeiçoar procedimentos de indexação e recuperação de informação.
- 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 de palavras e conjuntos de palavras.
- Posteriormente, é feita Decomposição em Valores Singulares (SVD).
- Referências:
  - DEERWESTER, S. et al. Indexing by latent semantic analysis. 1990.
  - LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. 1998.
  - STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. 2007.

LSI

$\{X\} =$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

$\{\hat{X}\} =$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. 1998.

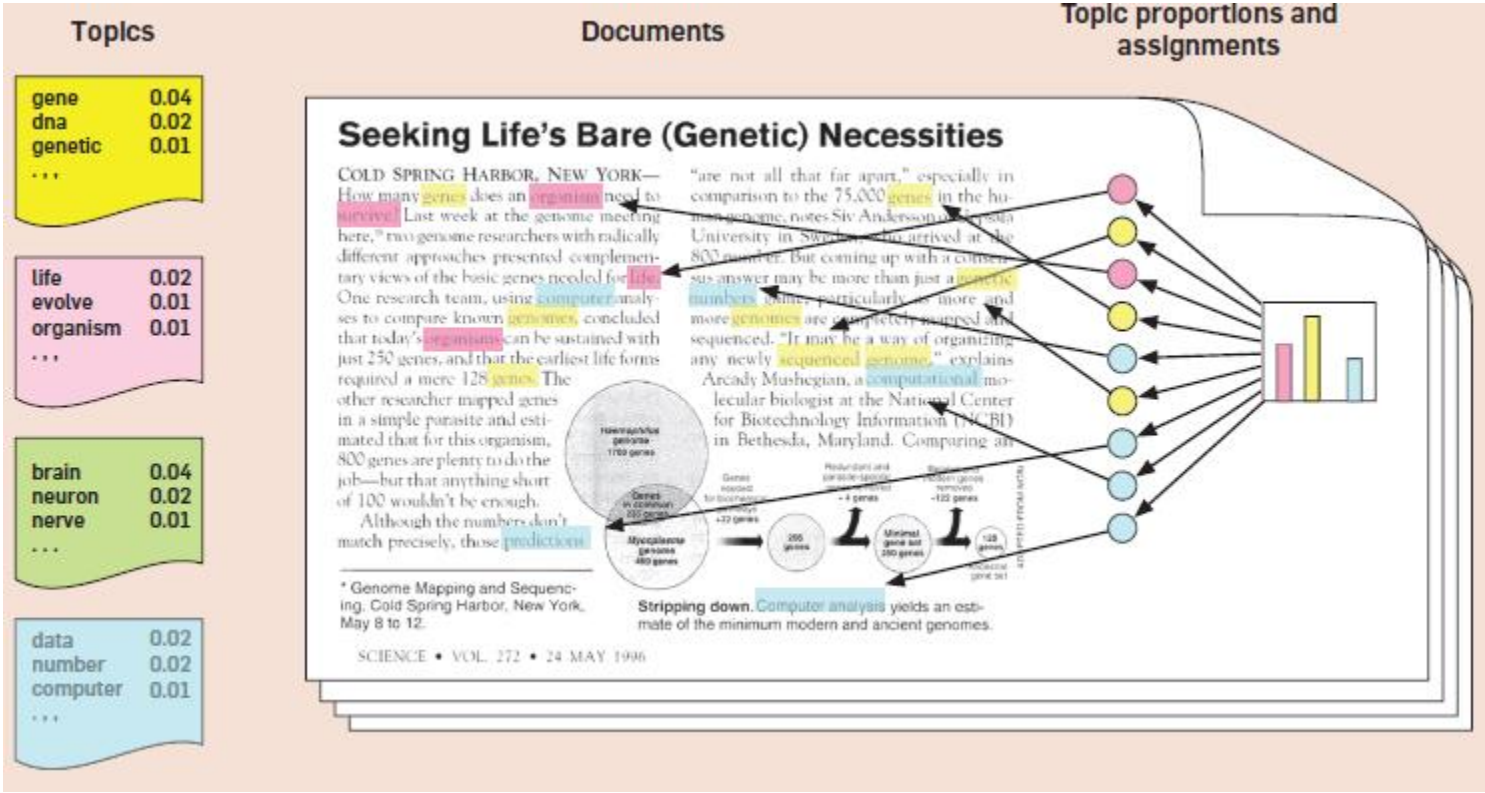
pLSI

- Sólida fundamentação estatística.
- O pLSI apresenta problemas tal como a impossibilidade de assumir probabilidades para um novo documento.
- Referências:
  - HOFMANN, T. Probabilistic latent semantic indexing. 1999.

LDA

- Referências:
  - BLEI, David et al. Latent Dirichlet Allocation. 2003.
  - BLEI, David. Probabilistic Topic Models: Surveying a suite of algorithms that offer a solution to managing large document archives. 2012.
- 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.
- A tarefa principal seria encontrar essa estrutura oculta a partir das variáveis observadas (documentos), o que pode ser compreendido como reversão do processo generativo.

LDA



BLEI, D. *Probabilistic Topic Models: Surveying a suite of algorithms that offer a solution to managing large document archives*. 2012.

Pressupostos do LDA

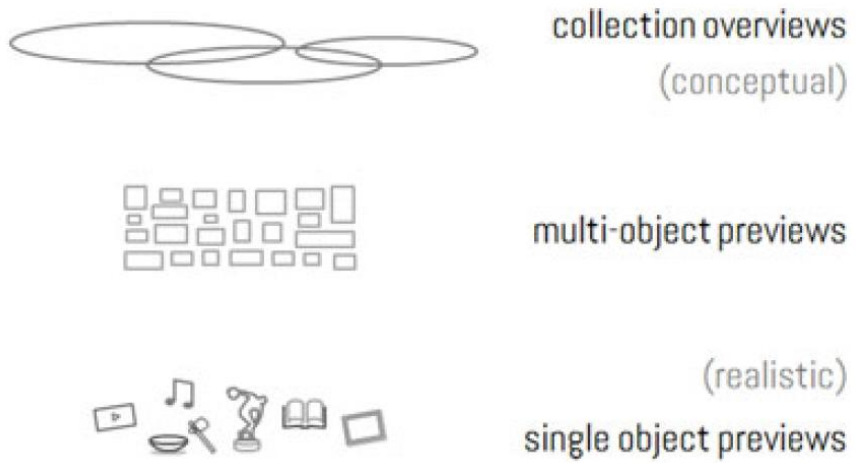
- “Bag of Words”
  - WALLACH, H. M. Topic modeling: beyond bag-of-words. 2006.
- Ordem dos documentos
  - BLEI, D. M.; LAFFERTY, J. D. Dynamic topic models. 2006.
- Número de tópicos
  - TEH, Y. W. et al. Sharing clusters among related groups: Hierarchical dirichlet processes. 2005.

Referências: aplicação

- LUCAS, C. et al. Computer-assisted text analysis for comparative politics.
- ROBERTS, M. E.; STEWART, B. M.; AIROLDI, E. M. A model of text for experimentation in the social sciences.

Referências: visualização

- Camadas de granularidade
  - Collection Overviews
  - Multi-Object Previews
  - Single Object Previews
- Público alvo
  - Usuário casual
  - Especialista de domínio

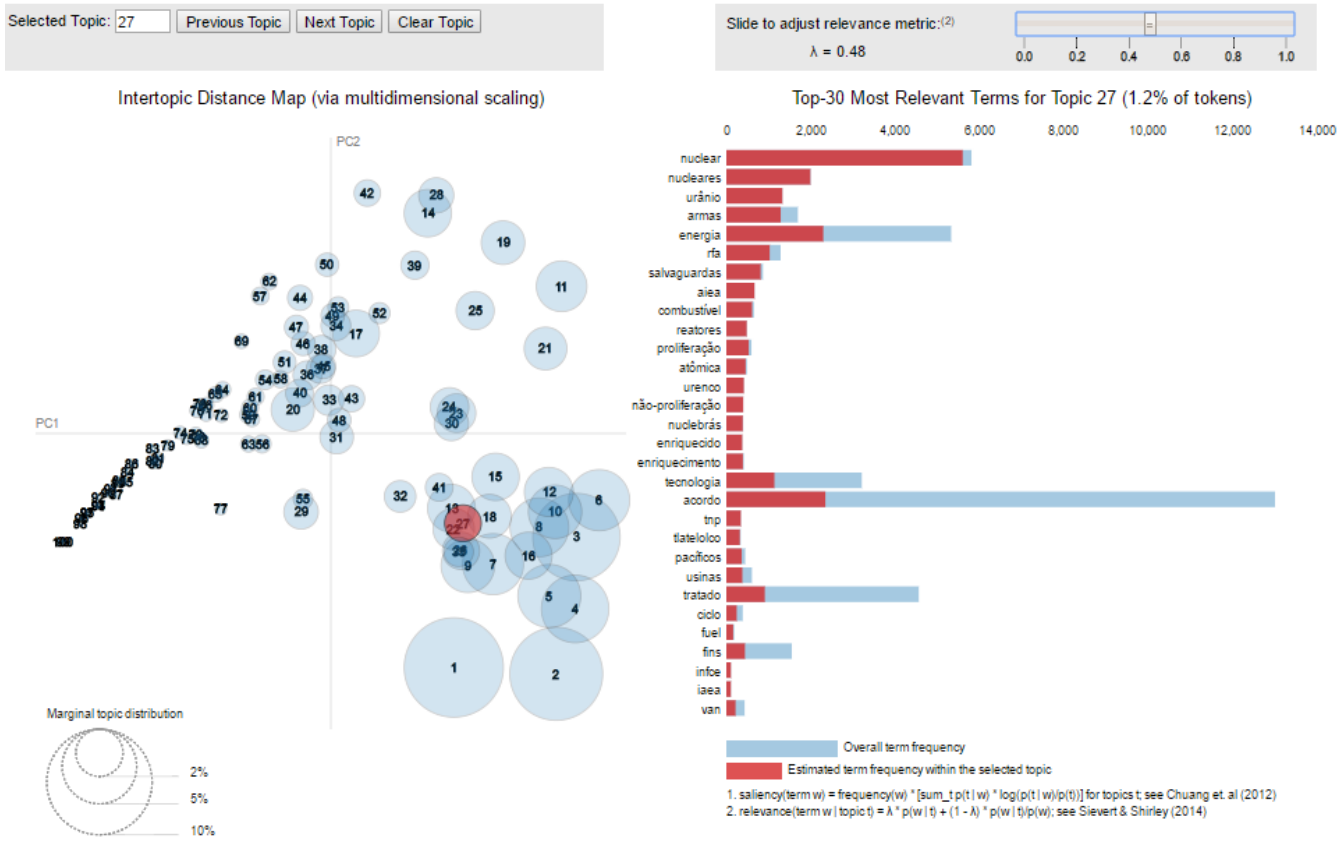


- WINDHAGER, F. et al. Visualization of cultural heritage collection data: State of the art and future challenges. 2018.

Referências: visualização

- Data-ink ratio
  - “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” -- Edward Tufte
- TUFTE, E. R. *The visual display of quantitative information*. 2001

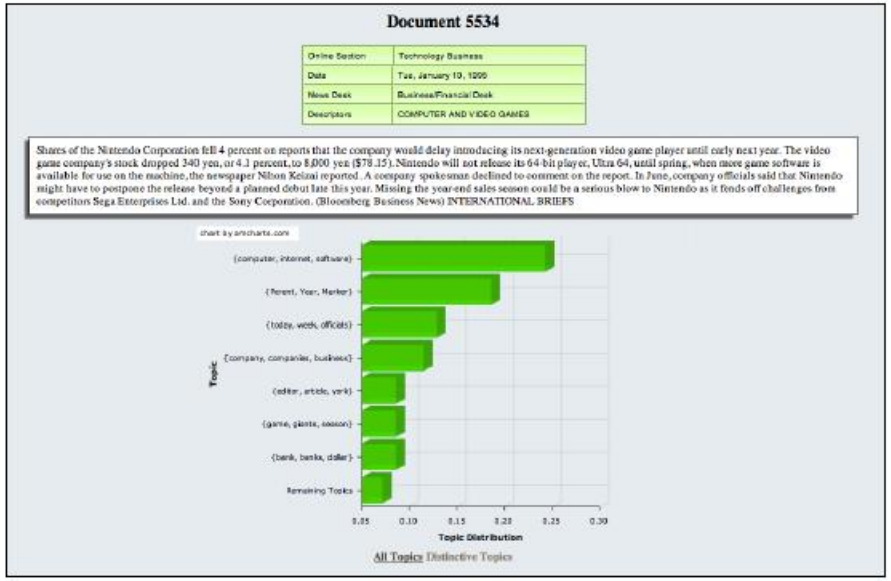
Estado da arte



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

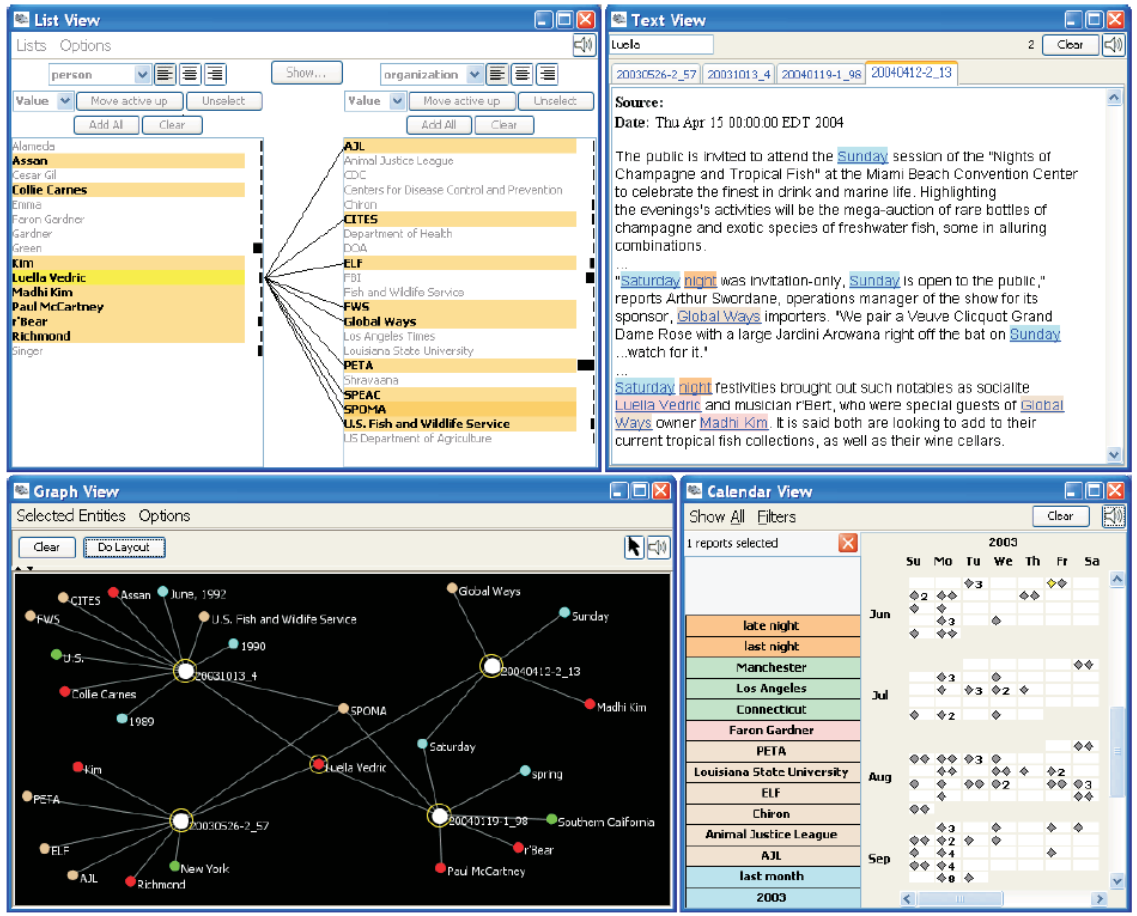
SIEVERT, C.; SHIRLEY, K. Ldavis: A method for visualizing and interpreting topics. 2014.

Estado da arte



SNYDER, J. et al. Topic models and metadata for visualizing text corpora. 2013.

Estado da arte



GORG, C. et al. Visual Analytics with Jigsaw. 2007.

Estado da arte

[100 data stories]

Passos metodológicos

Passos metodológicos

- Ferramentas
- Dados
- Processamento
  - OCR
  - Limpeza de dados
  - Extração de entidades
  - Modelagem de tópicos
  - Alimentação da base de dados
- Visualização

Obs: A parte que precede a visualização foi feita no escopo de um trabalho anterior, em projeto entre FGV/EMAp, FGV/CPDOC e Columbia University:

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.

Ferramentas

- Python (Jupyter Notebook)
  - Manipulações de base de dados: sqlite3, Pandas
  - Modelagem de tópicos: Gensim
  - Visualizações auxiliares: Seaborn, Matplotlib
- D3.js (Observable)
  - Geração da visualização principal
- Vega-Lite (Observable)
  - Visualizações auxiliares

Ferramentas

- Questões:
  - Flexibilidade,
  - Curva de aprendizado,
  - Reprodutibilidade,
  - Escalabilidade.
- Vantagens do D3.js:
  - Alta flexibilidade e expressividade como diferencial.
  - Permite gerar grafos, interatividade, dispõe de funções como mouse-over e mouse-click, acesso a hiperlink.
  - Grande disposição de exemplos.
- Alternativas ao D3.js:
  - Vega-Lite,
  - Python: Plotly, Altair, Bokeh.

Obtenção dos dados

- Base de dados foi disponibilizada pela FGV/CPDOC.
- O CPDOC é a Escola de Ciências Sociais e Centro de Pesquisa e Documentação de História Contemporânea do Brasil da Fundação Getulio Vargas (FGV).
- Possui coleções de diversas figuras públicas importantes do país, tais como: Getúlio Vargas, Café Filho, João Goulart, Ernesto Geisel e Anônio Azeredo da Silveira.
- Concedeu direito de acesso a parte de sua base de dados para o desenvolvimento do trabalho (novembro/2018): Coleção Antonio Azeredo da Silveira, Série MRE.

Coleção Antonio Azeredo da Silveira

- A série AAS-MRE (Ministério de Relações Exteriores) faz parte da coleção Antonio Azeredo da Silveira. Foi utilizada como piloto do projeto ligado ao CPDOC.
- Antonio Azeredo da Silveira foi ministro das Relações Exteriores no governo de Ernesto Geisel, de 1974 a 1979.
- 45 mil documentos em toda a coleção.
- Ano de doação da coleção: 1996

Azeredo da Silveira and Henry Kissinger, 1974

Base de dados

- Dimensões
  - +10 mil documentos
  - +66 mil páginas
  - +14 milhões de tokens/palavras (dicionarizados ou não)
  - 5 idiomas, principalmente português
- Formatos
  - Imagens (.tif e .jpg)
    - Digitalizadas a partir de documentos físicos
    - 24.8 GB
  - Textos (.txt)
    - 114 MB

Base de dados

AAS_1974.08.15
mre1pn

ECT TELEX ECT

BRASEMB WASHINGTON
EM 9/6/75

URGENTISSIMO
DEC/DCS/DE-I/DIE/RIG/
COOPERACAO NUCLEAR
BRASIL-R.F.A.

PARA CONHECIMENTO IMEDIATO DO SENHOR MINISTRO DE ESTADO

2026 - SEGUNDA FEIRA - 14:00 - O ''JOURNJL OF COMMERCE''

DE HOJE PUBLICA O SEGUINTE DESPACHO DE BONN, SOB A ASSINATURA DE JESS LUKOMSKI E COM O TITULO ''WEST GERMANS ANGERED OVER MOVES IN THE US AGAINST A NUCLEAR PACT'':

''THE WEST GERMAN GOVERNMENT, THE PARLIAMENTARY OPPOSITION, AND BUSINESS COMMUNITY HERE ARE BOTH PERTURBED AND ANGERED BY AN OPEN CAMPAIGN IN HE UNITED STATES AGAINST A GERMJN-BRAZILIAN TREATY ON COOPERATION IN NUCLEAR TECHNOLOGY FIELD.

THEY INTERPRET THE EFFORTS OF U.S. SENATORS TO CK THE PENDING SIGNATURE OF THE TREATY AS AN ATTEMPT OF ELIMINATING WEST GERMANY AS AN U UNCOMFORTABLE COMPETITOR ON THE WORLD MARKET FOR NUCLEAR REACTORS, FUEL PROCESSING PLANTS, AND URANIUM ENRICHMENT FACILITIES.

THE BONN GOVERNMENT, VITALY INTERESTED IN KEEPING THE CONTROVERSY FROM SOURING GERMAN-AMERICAN RELATIONS, IS QUICK TO POINT OUT THAT THE AMERICAN GOVERNMENT HAS BEEN CONSULTED AND FULLY INFORMED ABOUT THE NEGOTIATIONS WHICH WERE CONCLUDED ON FEB. 12.

ONLY A WEEK LATER, MARTIN HILLENGRAND, THE U.S. AMBASSADOR HERE, WAS INFORMED ABOUT THE TREATY'S TEXT WHICH WAS SUBSEQUENTLY DISCUSSED WITH A GROUP OF U.S. EXPERTS IN BONN EARLY IN APRIL.

I-7

Processamento de dados

- OCR
  - Tesseract
- Limpeza de dados
  - Expressões regulares
- Extração de entidades
  - Palavras
- Modelagem de tópicos
- Alimentação de dados e metadados em SQL

Modelagem de tópicos

- Modelos testados: LDA e HDP.
- Versões com 20, 30, 40, 45, 60 e 100 quantidades de tópicos.
- Versão final utilizada:
  - LDA
  - 100 tópicos totais
  - 10 tópicos validados junto a um especialista de domínio

SQL

- Textos processados dos documentos :
  - docs
- Resultados de modelagem de tópicos:
  - topics
  - doc_topics: many to many
- Resultados de extração de entidades:
  - persons
  - person_doc: many to many

Tratamento dos dados para visualização

- Pipeline dos dados:
  - SQLite
  - Pandas
  - JSON: doc_ids, person_id, tokens
  - GitHub (raw)
  - Objetos no Observable

Tratamento dos dados

- Dados em JSON aninhados (*nested data*) para formar os grafos:
  - doc_id
    - person_ids
    - tokens
  - person_id
    - doc_ids
  - tokens
    - doc_ids

Objetos: documentos


```
36: ▼Object {
  topic: "topic35"
  documents: ▼Array(20) [
    0: ▶Object {doc_id: "ag_1974.01.22_doc_V-17", topic_score: 0.8990649934, url: "http://www.fgv.br/cpdoc/acervo/arquivo-
    1: ▶Object {doc_id: "pn_1975.04.25_doc_4", topic_score: 0.5458315189, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pes
    2: ▶Object {doc_id: "pn_1975.04.25_doc_5", topic_score: 0.4979462396, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pes
    3: ▶Object {doc_id: "be_1977.01.27_doc_II-11", topic_score: 0.4858728345, url: "http://www.fgv.br/cpdoc/acervo/arquivc
    4: ▶Object {doc_id: "pn_1975.04.25_doc_6", topic_score: 0.4849777956, url: "http://www.fgv.br/cpdoc/acervo/arquivo-pes
    5: ▶Object {doc_id: "be_1977.06.01_doc_II-1", topic_score: 0.4806304466, url: "http://www.fgv.br/cpdoc/acervo/arquivo-
    6: ▶Object {doc_id: "be_1977.04.29_doc_II-27", topic_score: 0.4167021578, url: "http://www.fgv.br/cpdoc/acervo/arquivc
    7: ▶Object {doc_id: "rb_1974.04.17_doc_II-8", topic_score: 0.4089984822, url: "http://www.fgv.br/cpdoc/acervo/arquivo-
    8: ▼Object {
      doc_id: "d_1974.03.26_doc_XXIII-31"
      topic_score: 0.3809786886
      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", "
      names: ▶Array(2) ["John Hugh Crimmins", "Jimmy Carter"]
      doc: "doc_XXIII-31"
      topic_id_renamed: 35
    }
```

Objetos: entidades


```
36: ▼Object {
  topic: "topic35"
  documents: ▶Array(20) [Object, Object, Object, Object, Object, Object, Object, Object, Object, Object, Object, Object, (
  names: ▼Array(5) [
    0: ▶Object {name: "Antonio Azeredo Da Silveira", count: 0.1, docs: Array(2), x: 100, y: 0}
    1: ▶Object {name: "Cyrus Vance", count: 0.05, docs: Array(1), x: 30.901699437494745, y: 95.10565162951535}
    2: ▶Object {name: "Helmut Schmidt", count: 0.05, docs: Array(1), x: -80.90169943749473, y: 58.77852522924732}
    3: ▼Object {
      name: "Jimmy Carter"
      count: 0.25
      docs: ▼Array(5) [
        0: "pn_1974.08.15_doc_I-23"
        1: "pn_1974.08.15_doc_III-31"
        2: "pn_1976.12.28_doc_29"
        3: "pn_1974.08.15_doc_II-1"
        4: "d_1974.03.26_doc_XXIII-31"
      ]
      x: -80.90169943749474
      y: -58.7785252292473
    }
    4: ▶Object {name: "John Hugh Crimmins", count: 0.05, docs: Array(1), x: 30.901699437494724, y: -95.10565162951536}
  ]
  tokens: ▶Array(20) [Object, Object, Object, Object, Object, Object, Object, Object, Object, Object, Object, Object, Obje
}
```

Objetos: tokens

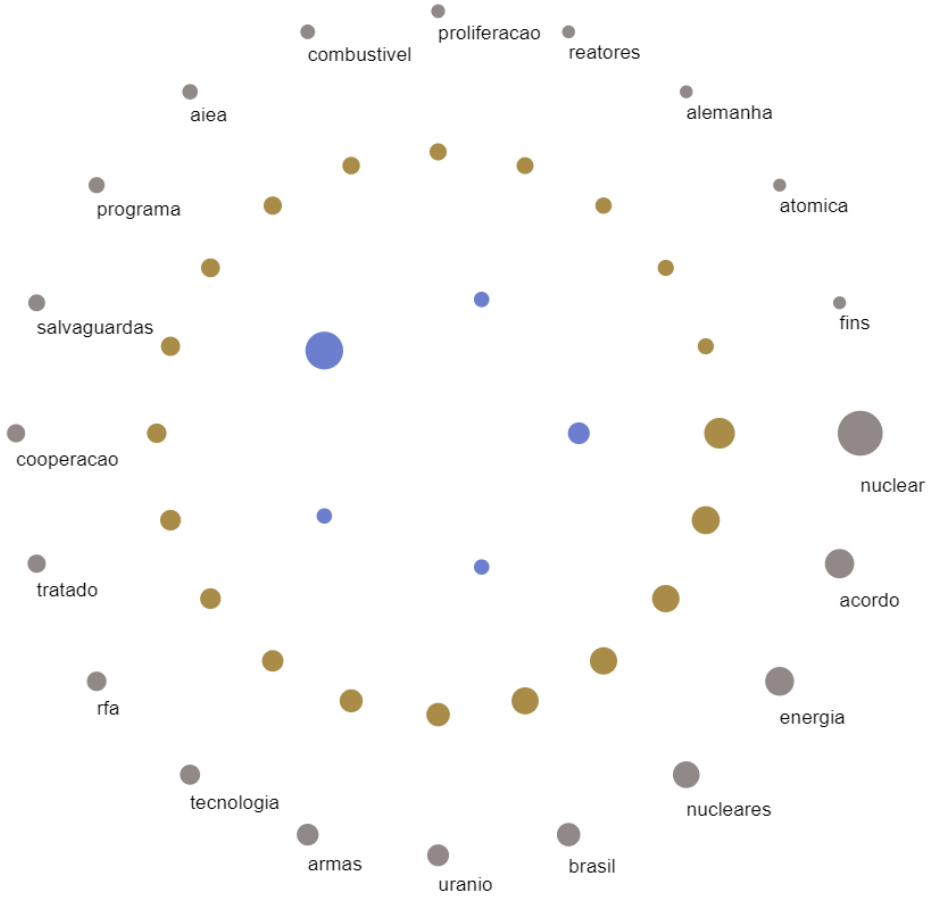

```
36: ▼Object {
  topic: "topic35"
  documents: ▶Array(20) [Object, Object, Object, Object, Object, Object, Object, Object, Object, Object, Object, Object, (
  names: ▶Array(5) [Object, Object, Object, Object, Object]
  tokens: ▼Array(20) [
    0: ▼Object {
      token: "nuclear"
      score: 0.085
      docs: ▶Array(19) ["ag_1974.01.22_doc_V-17", "pn_1975.04.25_doc_4", "pn_1975.04.25_doc_5", "be_1977.01.27_doc_II-11",
      x: 300
      y: 0
    }
```

Tratamento dos dados

Repositório da dissertação no GitHub

Visualização

Visualização



Considerações Finais

Considerações finais

- 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.
- Como o redirecionamento é ligado à própria organização responsável pelo armazenamento dos documentos, há uma valorização da mesma, ao aumentar o acesso de usuários.

Possíveis melhorias

- Processamento
  - Refazer todo o ciclo desde a base de imagens:
    - Pré-filtragem de imagens (antes do OCR): manuscritos, gráficos.
    - Testar alternativas em limpezas de dados: soundex, enchant.predict
    - Testar filtragem de termos: usar somente os termos presentes em dicionário + obtidos por extração de entidades.
    - Aplicação de métricas de coerência para comparar os resultados de tópicos.

Possíveis melhorias

- Visualização
  - Aumentar interatividade
    - Fornecer flexibilidade na escolha do número de nós a serem expostos.
    - Escolha do tópico por meio da visualização auxiliar.
  - Implementar um sistema que permita refazer a modelagem de tópicos.

Links principais

Repositório da dissertação no GitHub

•

Visualização gerada em D3.js (Observable)

Obrigado

Marcelo B. Barata Ribeiro

•

marcelobbribeiro@gmail.com