Mineração de dados com documentos históricos

Uma abordagem sobre modelagem de tópicos e processamento de linguagem natural

•

Renato Rocha Souza
Alexandre Moreli
Marcelo Barata Ribeiro

FGV

Sumário da apresentação

1. Projeto e objetivo
2. A coleção
3. Principais passos
4. Resultados

O projeto

- Trabalho conjunto entre CPDOC/FGV e EMAP/FGV.
- Mineração de dados com documentos históricos ligados à área de relações exteriores.
- Integração com o projeto History Lab, organizado pela Universidade Columbia.
- Acervos do History-Lab:
  - CPDOC,
  - Foreign Relations of the United States (FRUS),
  - State Department Central Foreign Policy Files,
  - US Declassified Documents Online (DDO),
  - Kissinger Telephone Conversations,
  - Clinton E-Mail,
  - UK Cabinet Papers.
- Principais ferramentas: Python e MySQL.

HISTORY LAB

History as Data Science

CPDOC

Antônio Azeredo da Silveira

•

Café Filho

•

Ernesto Geisel

•

Getúlio Vargas

•

João Goulart

•

•

•

Coleção Antonio Azeredo da Silveira

- Foi selecionada a série AAS-MRE (Ministério de Relações Exteriores) da coleção Antonio Azeredo da Silveira como piloto do projeto de integração do banco de dados ao History-Lab.
- Antonio Azeredo da Silveira foi ministro das Relações Exteriores no governo de Ernesto Geisel, de 1974 a 1979.
- 45 mil documentos.
- Ano de doação da coleção: 1996

Azeredo da Silveira and Henry Kissinger, 1974

Os documentos

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

Etapas

4.1 Digitalização

4.2 OCR

4.3 Limpeza de dados

4.4 Modelagem de tópicos

4.5 Extração de Entidades

4.6 Integração dos resultados ao History-Lab

Digitalização

- Em 2009, foi contratado o serviço para digitalizar cada documento da série MRE da coleção Antonio Azeredo da Silveira.

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

Valente, M. Gurgel

Montevideu, 2/2/74

Meu caro Silveira,

Izabel e eu enviamos, o Mary e a Você, nosso

M. ET Mme GURGEL VALENTE

abraço de parabens, com votos de plena realização pessoal — já que a profissional é garantida — na carreira onde Você estará

I-2

após 15 de março. Aperfeiçoado este tratamento intitulo pois, aquela data passada o Excelentíssimo será respectivamente de direito. Que acha de reunirmos Embaixadores da área do agrião?

Nesse interim, "tanti auguri e belle cose".

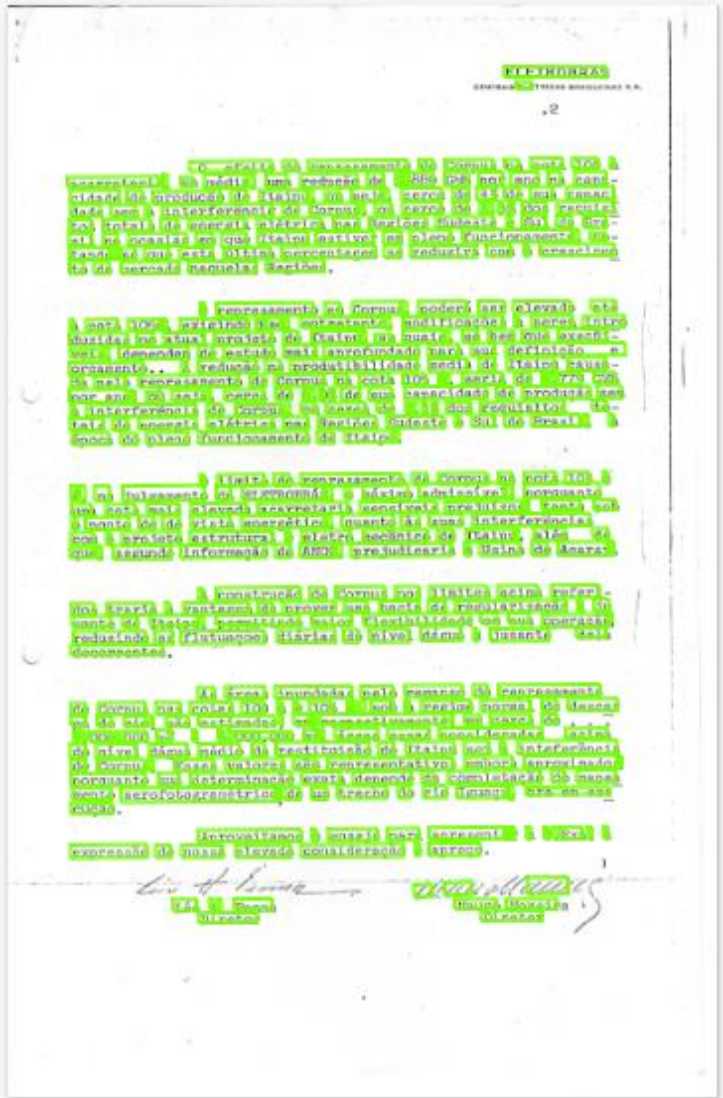
Lew, de sempre

Faly

OCR

- Ferramenta utilizada: Tesseract
- Alternativa: Cloud Vision API
- Alternativa (se utilizasse manuscritos): Transkribus

Labels | Web | Text | Document | Properties | Safe Search | JSON

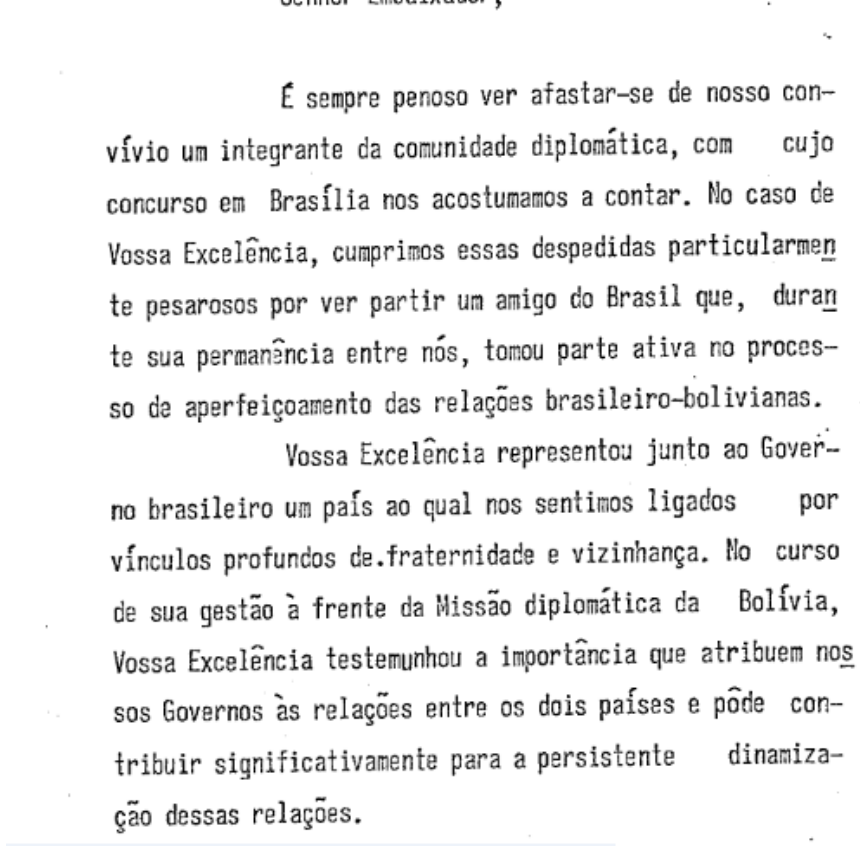


I0016761.JPG

“ ELETROBRAS ELE Oefeito do represamento de Corpus na cota 100 m acarretaria, em inédia, uma redução de 2.889 GHn por ano na capa cidade de producão de Itaipu, ou seja, cerca de 46 de sua capaci dade sem a interferencia de Corpus, ou cerca de 1, 3% dos requisi tos totais de energia elétrica nas Regiões Sudeste e Sul do Bra sil na Ocasião em gue Itaipu estiver em pleno funcionamento, no tando-se que esta última percentagem se reduzira com o crescimen to do mercado naquelas Regiões o represamento em Corpus, poderá ser elevado até a cota 105m, exigindo is so, entretanto, modificações a seren intro duzidas no atual projeto de Itaipu, as quais, se bem que exegii Veis , dependem de estudo mais aprofundado para sua definiçãoe orçamento A redução na produtibilidade media de Itaipu causa

Limpeza de Dados

- Análise exploratória de dados.
- Uso de expressões regulares (regex).
  - Total de linhas de código de regex: 360.
  - docs.python.org/2/library/re.html
  - Regex101
- Mais de 500 mil termos não-dicionarizados, de um total de 2 milhões, foram submetidos ao processo de limpeza.



Senhor Embaixador,

É sempre penoso ver afastar-se de nosso convívio um integrante da comunidade diplomática, com cujo concurso em Brasília nos acostumamos a contar. No caso de Vossa Excelência, cumprimos essas despedidas particularmente pesarosos por ver partir um amigo do Brasil que, durante sua permanência entre nós, tomou parte ativa no processo de aperfeiçoamento das relações brasileiro-bolivianas.

Vossa Excelência representou junto ao Governo brasileiro um país ao qual nos sentimos ligados por vínculos profundos de.fraternidade e vizinhança. No curso de sua gestão à frente da Missão diplomática da Bolívia, Vossa Excelência testemunhou a importância que atribuem nos sos Governos às relações entre os dois países e pôde contribuir significativamente para a persistente dinamização dessas relações.

. Senhor Embaixador,

E sempre penoso ver afastar-se de nosso con-
vivio um integrante da comunidade diplomática, com cujo
concurso em Brasília nos acostumamos & contar. No caso de
Vossa Excelência, cumprimos essas despedidas particularmeg
te pesarosos por ver partir um amigo do Brasil que, duran
te sua permanência entre nos, tomou parte ativa no proces-
so de aperfeiçoamento das relações brasileiro-bolivianas.

Vossa Excelência representou junto ao GoverL
no brasileiro um país ao qual nos sentimos ligados por
vínculos profundos de.fraternidade e vizinhança. No curso
de sua gestão a frente da Missão diplomática da Bolívia,
Vossa Excelência testemunhou a importância que atribuem nos
sos Governos as relações entre os dois países e pôde contribuir significativamente para & persistente dinamiza-

ção dessas relações.

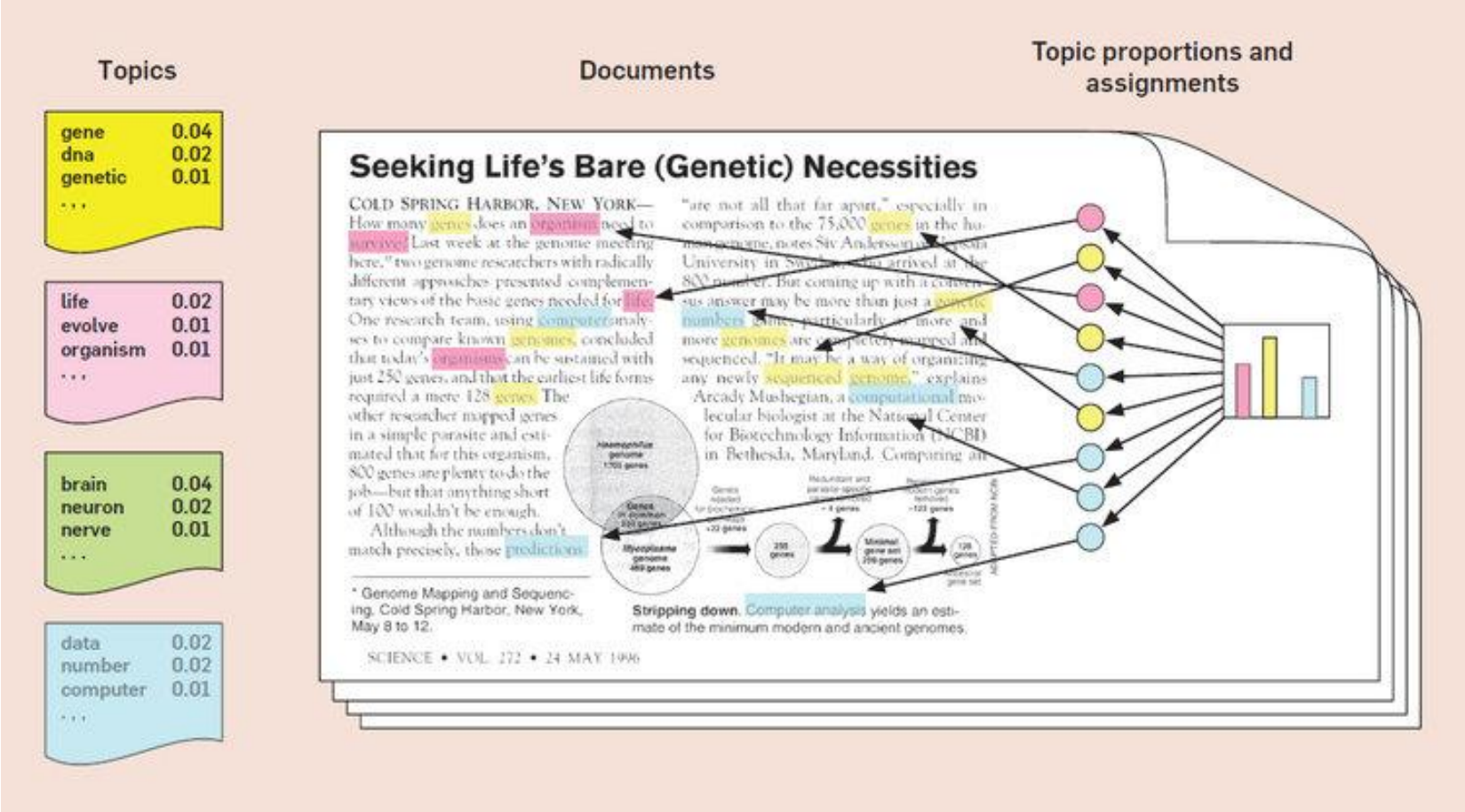
. senhor embaixador,

e sempre penoso ver afastar-se de nosso convívio um integrante da comunidade diplomática, com cujo
concurso em brasília nos acostumamos & contar. no caso de
vossa excelência, cumprimos essas despedidas particularmeg
te pesarosos por ver partir um amigo do brasil que, duran
te sua permanência entre nos, tomou parte ativa no processo de aperfeiçoamento das relações brasileiro-bolivianas.

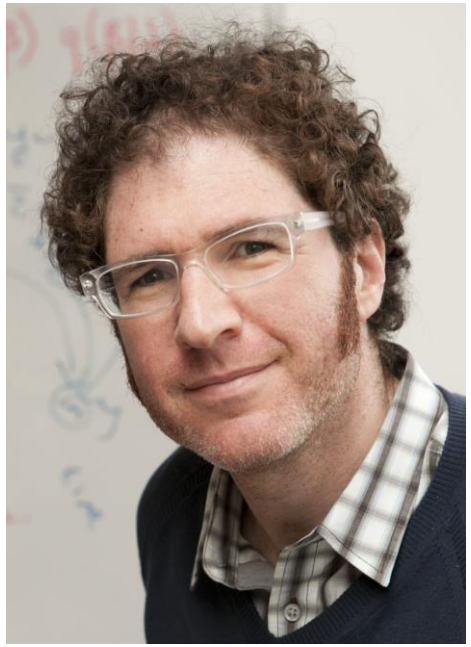
vossa excelência representou junto ao goverl
no brasileiro um país ao qual nos sentimos ligados por
vínculos profundos de. fraternidade e vizinhança. no curso
de sua gestão a frente da missão diplomática da bolívia,
vossa excelência testemunhou a importância que atribuem nos
sos governos as relações entre os dois países e pôde contribuir significativamente para & persistente dinamiza-

ção dessas relações.

Modelagem de Tópicos



Modelagem de Tópicos



Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: John Lafferty

Abstract

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

Modelagem de Tópicos

- Para a modelagem de tópicos, foi utilizado o LDA (Latent Dirichlet Allocation), método criado por David Blei.
- Ferramentas utilizadas: gensim e pyLDAvis (ambos são pacotes do Python).
- Diversos testes com os parâmetros da modelagem, principalmente com o número de tópicos.
- Alternativas: HDP e LSI.

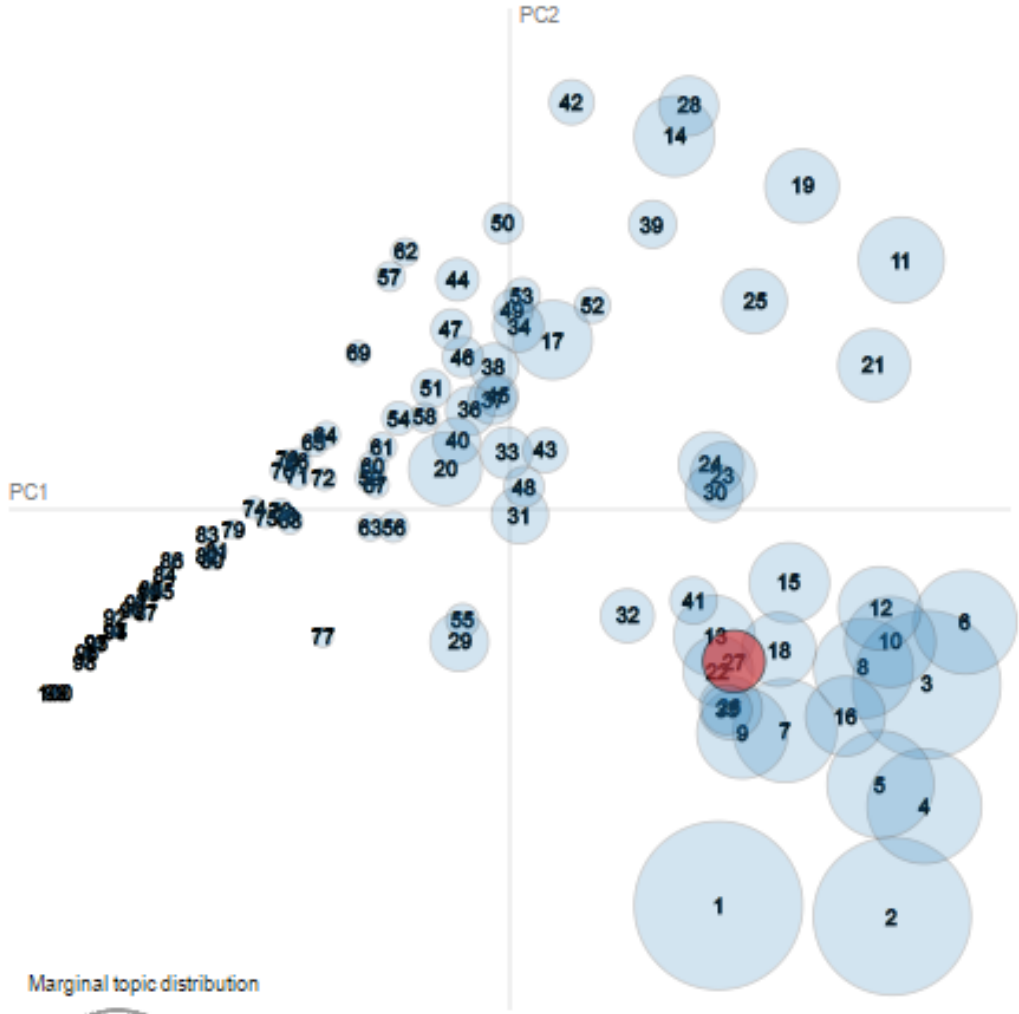
Selected Topic: 27 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)

λ = 0.48

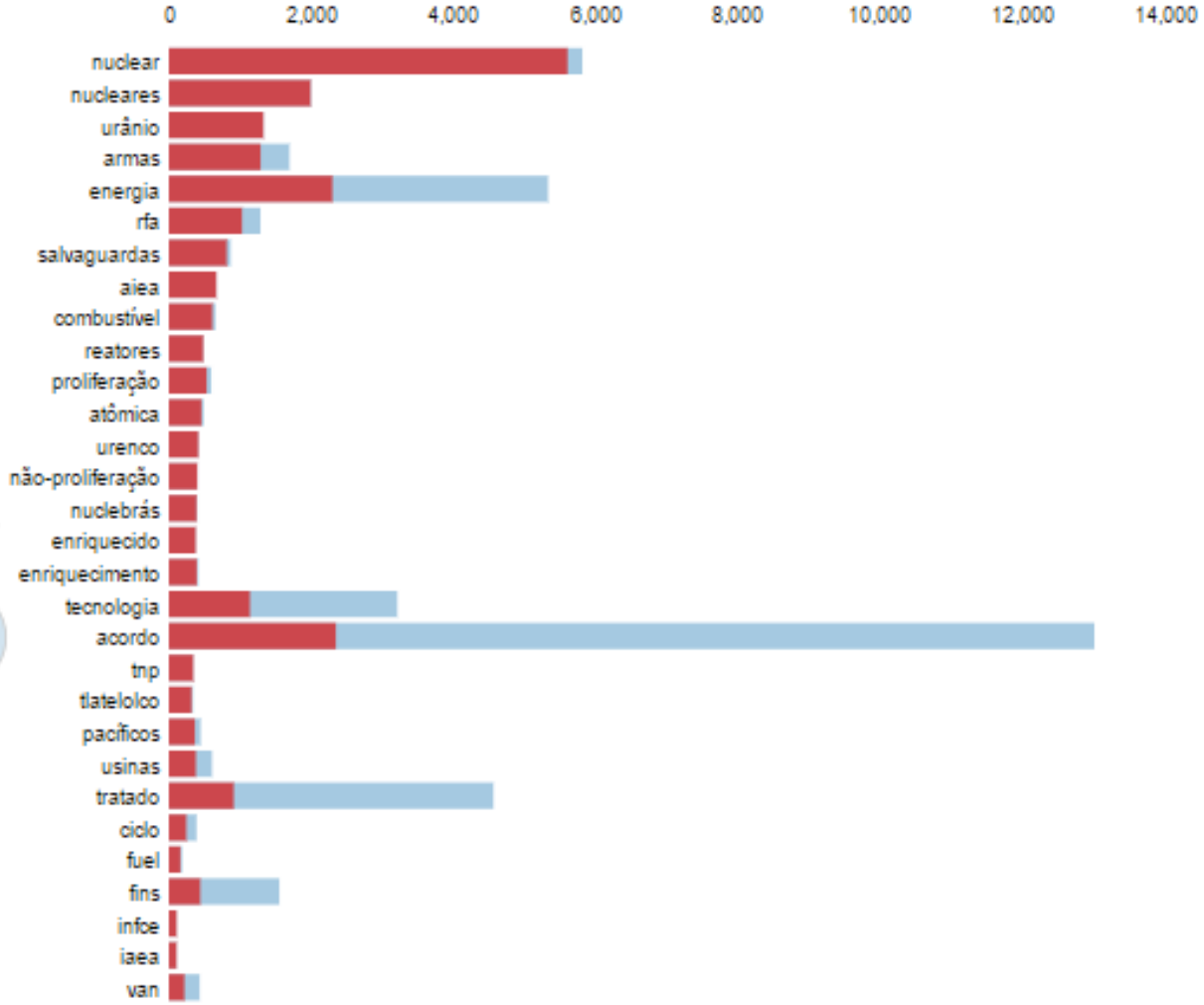
0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution

Top-30 Most Relevant Terms for Topic 27 (1.2% of tokens)



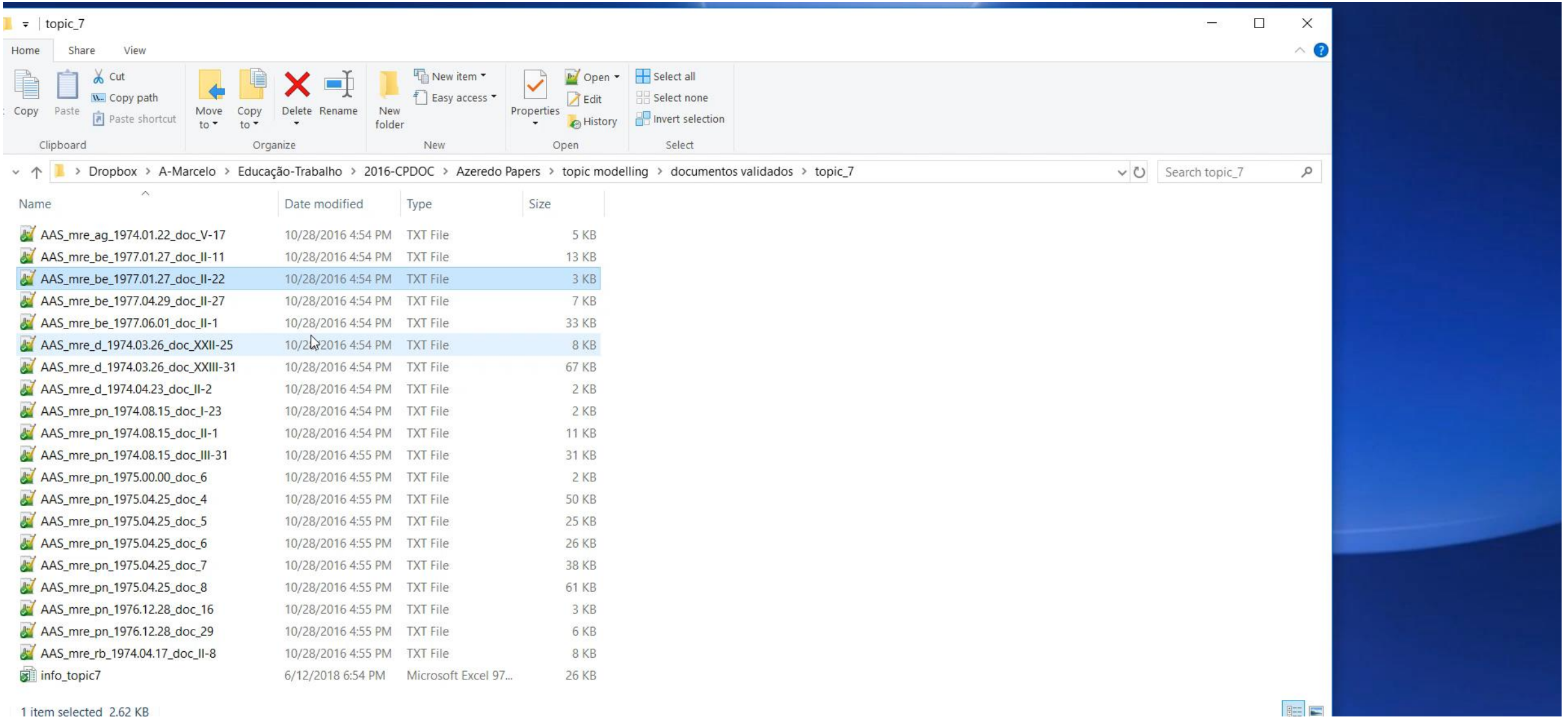
Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Validação

- Tivemos a colaboração de um especialista, que analisou 10 tópicos e 20 documentos com melhor score para cada um deles.
- Depois foi aplicado um índice de coesão, de acordo com o número de documentos que de fato abordavam o mesmo assunto.
- Alternativa: automatizar parte do processo. Ver, por exemplo, o paper "Exploring the Space of Topic Coherence Measures"


topic_7
Home
Share
View
Copy
Paste
Cut
Copy path
Paste shortcut
Clipboard
Move to
Copy to
Delete
Rename
Organize
New item
Easy access
New folder
New
Properties
Open
Edit
History
Open
Select all
Select none
Invert selection
Select
Dropbox › A-Marcelo › Educação-Trabalho › 2016-CPDOC › Azeredo Papers › topic modelling › documentos validados › topic_7
Search topic_7
Name
Date modified
Type
Size
AAS_mre_ag_1974.01.22_doc_V-17 10/28/2016 4:54 PM TXT File 5 KB
AAS_mre_be_1977.01.27_doc_II-11 10/28/2016 4:54 PM TXT File 13 KB
AAS_mre_be_1977.01.27_doc_II-22 10/28/2016 4:54 PM TXT File 3 KB
AAS_mre_be_1977.04.29_doc_II-27 10/28/2016 4:54 PM TXT File 7 KB
AAS_mre_be_1977.06.01_doc_II-1 10/28/2016 4:54 PM TXT File 33 KB
AAS_mre_d_1974.03.26_doc_XXII-25 10/28/2016 4:54 PM TXT File 8 KB
AAS_mre_d_1974.03.26_doc_XXIII-31 10/28/2016 4:54 PM TXT File 67 KB
AAS_mre_d_1974.04.23_doc_II-2 10/28/2016 4:54 PM TXT File 2 KB
AAS_mre_pn_1974.08.15_doc_I-23 10/28/2016 4:54 PM TXT File 2 KB
AAS_mre_pn_1974.08.15_doc_II-1 10/28/2016 4:54 PM TXT File 11 KB
AAS_mre_pn_1974.08.15_doc_III-31 10/28/2016 4:55 PM TXT File 31 KB
AAS_mre_pn_1975.00.00_doc_6 10/28/2016 4:55 PM TXT File 2 KB
AAS_mre_pn_1975.04.25_doc_4 10/28/2016 4:55 PM TXT File 50 KB
AAS_mre_pn_1975.04.25_doc_5 10/28/2016 4:55 PM TXT File 25 KB
AAS_mre_pn_1975.04.25_doc_6 10/28/2016 4:55 PM TXT File 26 KB
AAS_mre_pn_1975.04.25_doc_7 10/28/2016 4:55 PM TXT File 38 KB
AAS_mre_pn_1975.04.25_doc_8 10/28/2016 4:55 PM TXT File 61 KB
AAS_mre_pn_1976.12.28_doc_16 10/28/2016 4:55 PM TXT File 3 KB
AAS_mre_pn_1976.12.28_doc_29 10/28/2016 4:55 PM TXT File 6 KB
AAS_mre_rb_1974.04.17_doc_II-8 10/28/2016 4:55 PM TXT File 8 KB
info_topic7 6/12/2018 6:54 PM Microsoft Excel 97... 26 KB
1 item selected 2.62 KB

Extração de entidades

- Ferramenta principal: palavras.
  - Ferramenta muito eficiente, mas textos “sujos” reduzem a precisão.
  - Necessidade de adotar diferentes estratégias de extração.

AAS 1973. 11. 20
mre/ag

EMPRESA BRASILEIRA DE CORREIOS E TELÉGRAFOS

RECIBO DO TELEGRAMA ABAIXO DISCRIMINADO

DESTINO	Espaço reservado a autenticação mecânica
Embaixador Azeredo da Silveira - Ipanema - Rio Jan. - GB Será preenchida pelo expedidor	
E C T HORA DA TRANSMISSÃO INICIAIS DO OPERADOR	Espaço reservado a autenticação mecânica
INDICAÇÕES DE SERVIÇOS TAXADOS	URGENTE

TEXTO E ASSINATURA — ENDEREÇO

DESTINATÁRIO: Embaixador Azeredo da Silveira

Avenida Vieira Souto 408 apt 202 - Ipanema
(Rua, Av., etc.) (Bairro)

CIDADE: RIO de JANEIRO ESTADO: GUANABARA
(ou nome da estação móvel, no radiograma) (ou nome da estação terrestre, no radiograma)

Congratulando-nos futuro Governo Geisel pela acertadíssima et aplaudida escolha seu ilustre nome para cargo Ministro Relações Exteriores vg enviamos queridos amigos Dom Antonio Dona May afetuosos abraços de felicitações et melhores votos para sua continuada felicidade pessoal et todos êxitos desempenho grande Missão lhe foi confiada pt Com mais elevada estima continuaremos sempre seus leais gratos admiradores

ROBERTO et BETTY

Embaixador Roberto Barthel
NOME EXPEDIDOR TELEFONE

Rua Rui Barbosa , 310 /ap. 205 -Fortaleza - Teresópolis/RJ
Rua Bairro Cidade

2530-007-0051

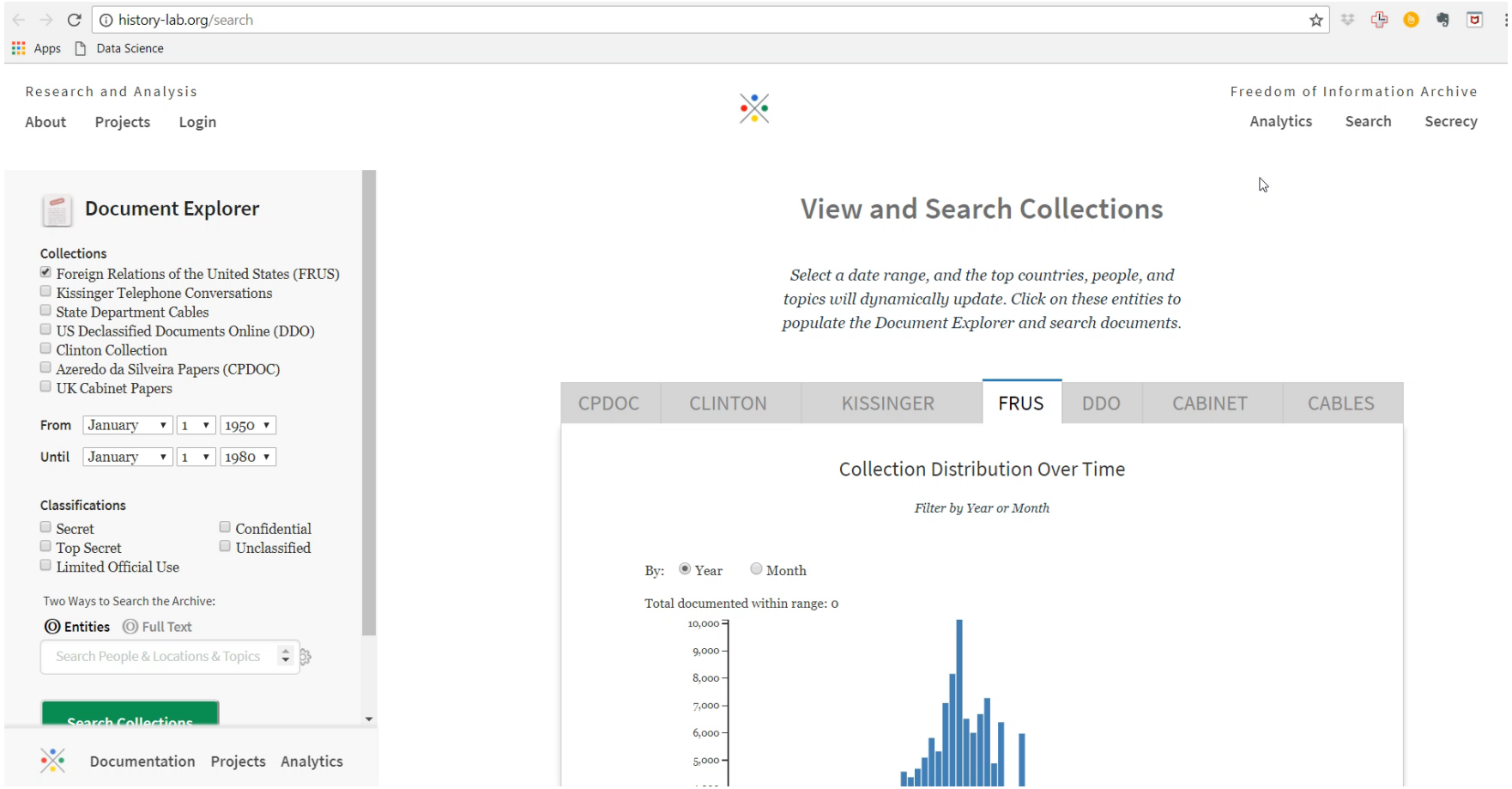
R. T. Barthel

162 x 229 mm.
I - 46 Au

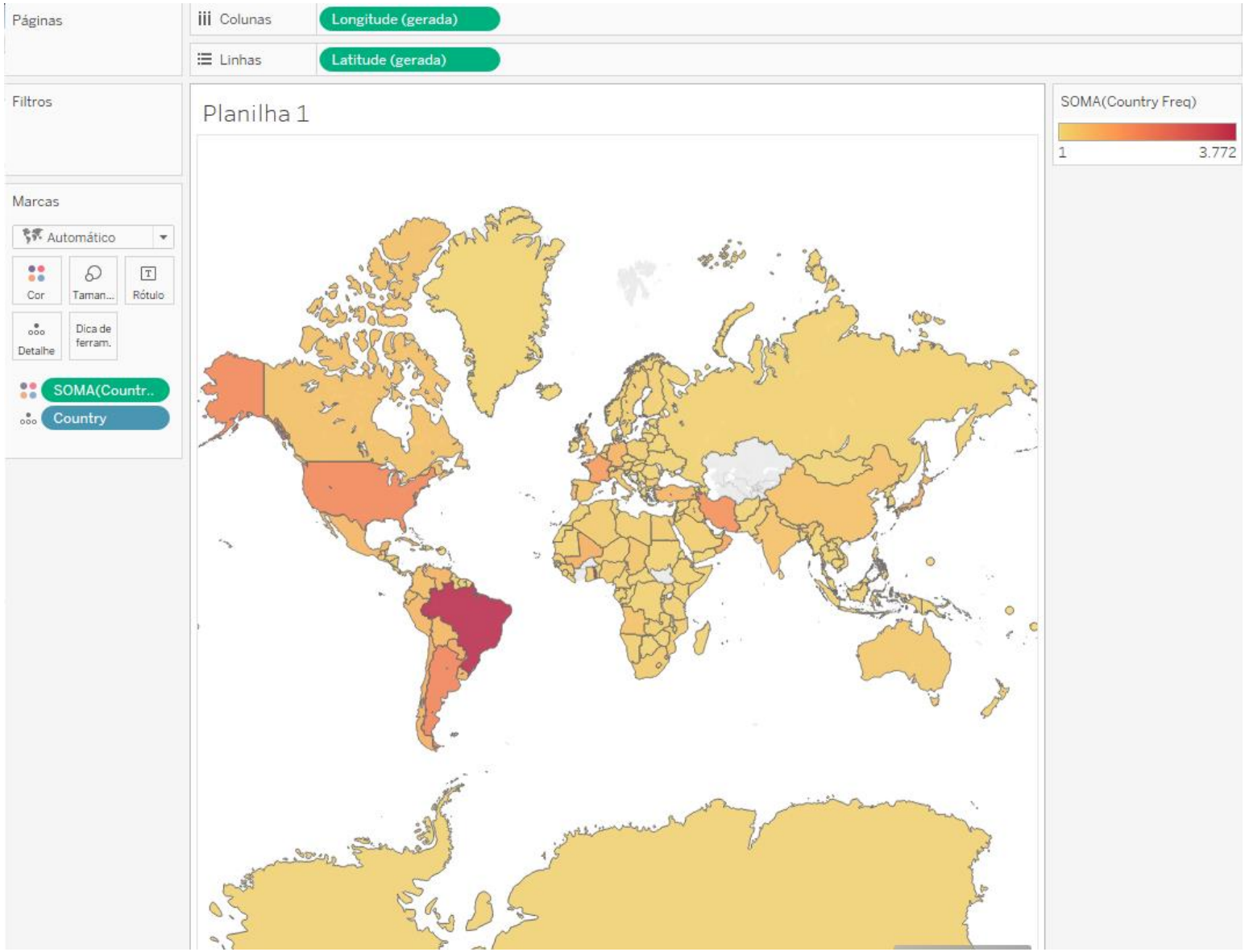
Resultados

- 10 tópicos validados (índice de coesão em parênteses)
  - Itaipu plant (100%)
  - Nuclear Brazil (97.50%)
  - Latin America and Caribbean (95%)
  - International Economic Relations of Brazil (95%)
  - International Cooperation for Development (92.50%)
  - Geisel foreign policy: ideas and action (87.50%)
  - UN system (78%)
  - United States of America (78%)
  - Brazil, Africa and decolonization (75%)
  - Brazilian government and private investment (73%)
- Extração de entidades
  - +49 mil entradas encontradas

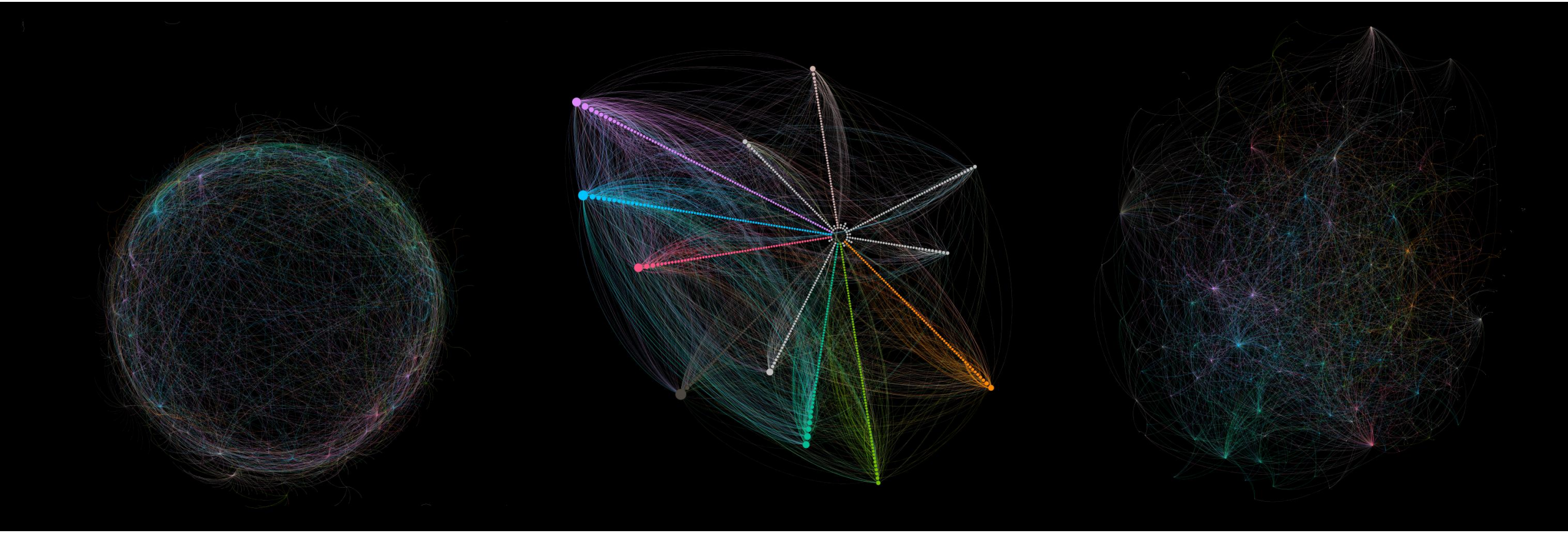
Integração ao History-Lab



Visualização



Visualização



Obrigado

Renato Rocha Souza
Renato.Souza@fgv.br

•

Alexandre Moreli
alexandre.moreli@fgv.br

•

Marcelo Barata Ribeiro
marcelobbribeiro@gmail.com