Text data

Topic Modeling, NLP e outros tratamentos com texto

Hugo Barreto
Marcelo B. Barata Ribeiro

Text data: etapas

1. Digitalização
2. OCR
3. Limpeza de dados / pré-processamento
   - Stem/Lemma
   - Stopwords
   - TF-IDF
4. Predição/Resultados
   - Machine Learning
   - Modelagem de tópicos
   - Extração de Entidades

Computer vision e OCR

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 intitulado 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".

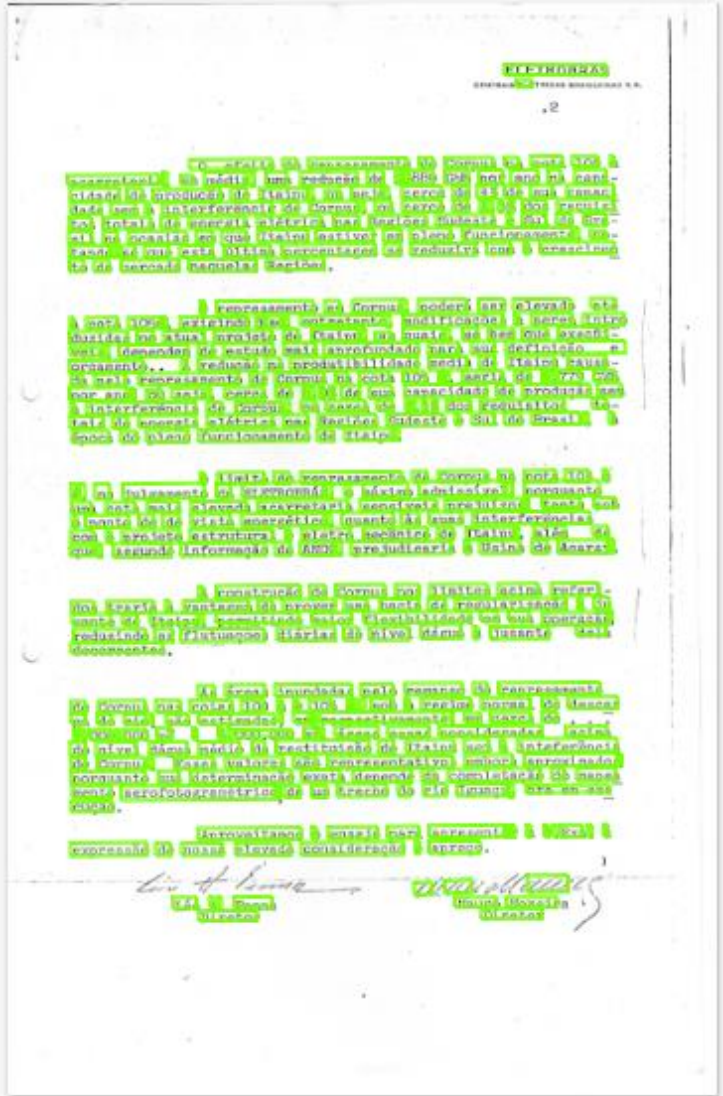
Seu, de sempre

Fary

Computer vision e OCR

- Ferramentas para lidar com imagens de texto:
  - Tesseract
  - Cloud Vision API
  - Transkribus (somente para manuscritos)

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

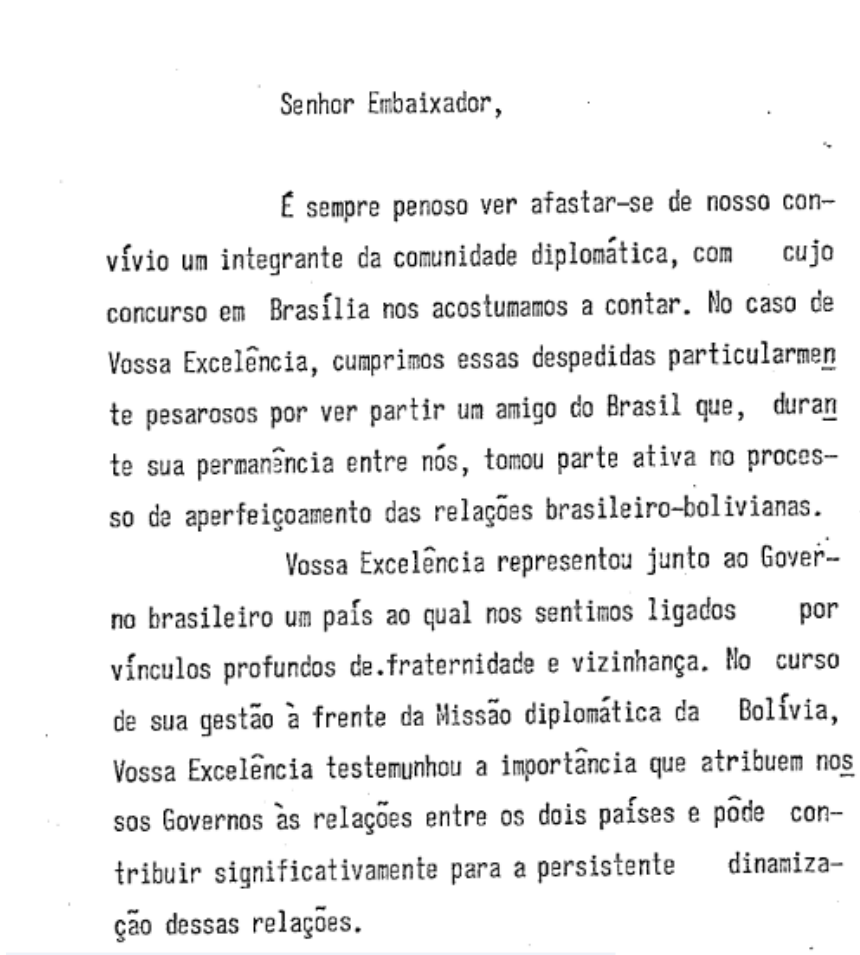


I0016761.JPG

“ 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

- Expressões regulares (regex).
  - docs.python.org/2/library/re.html
  - Regex101
- Soundex
- enchant



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 convivio 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.

. 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.

Pré-processamento

- Stem
- Lemma
- Stopwords

TF-IDF

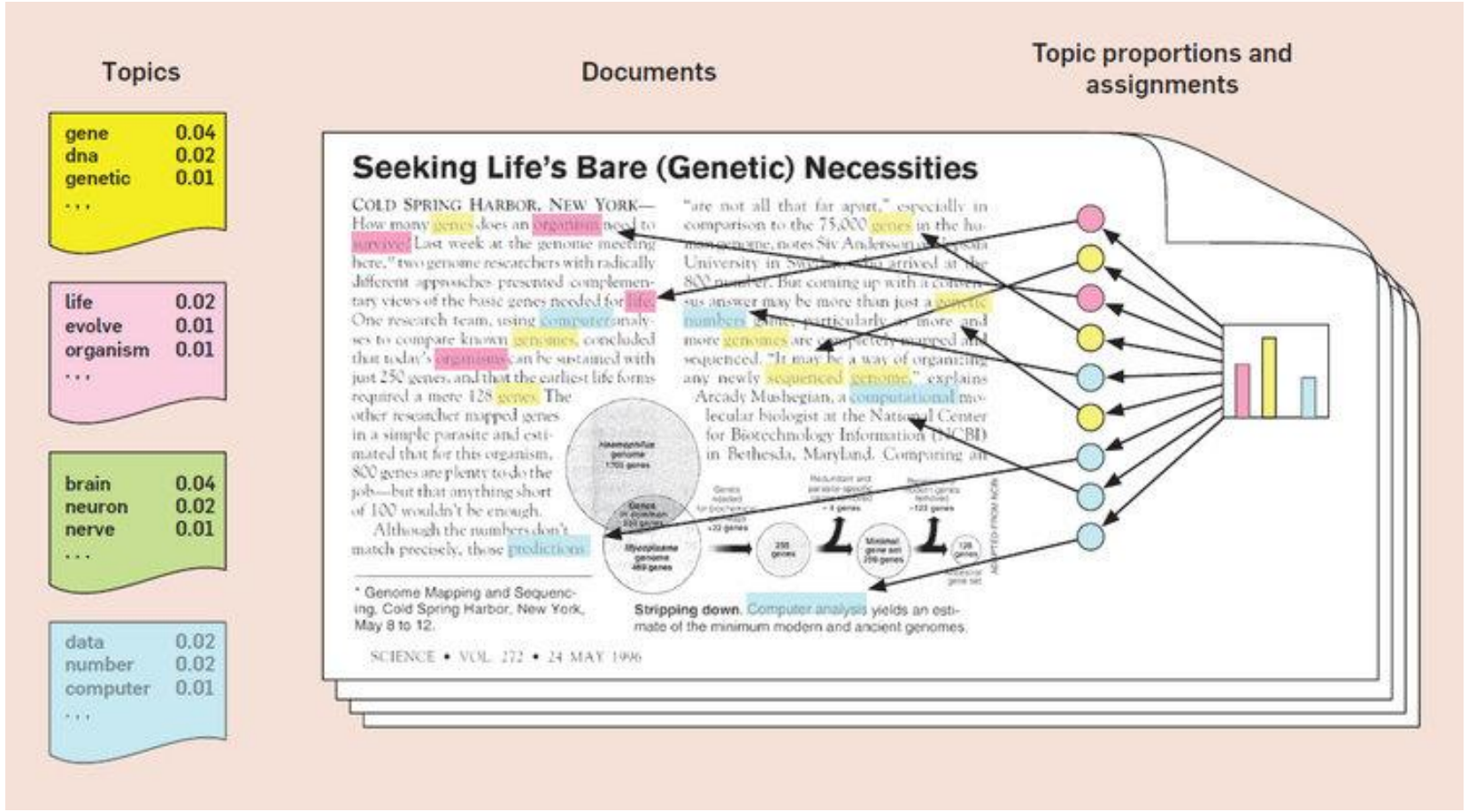
- A fazer

Modelos de similaridade

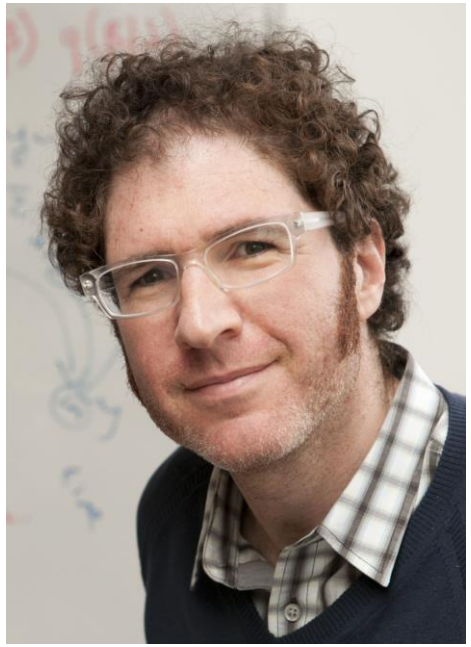
- A fazer
- Cosseno

Modelagem de Tópicos

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

- Pacotes do python: gensim e pyLDAvis
- Modelo mais utilizado: LDA (Latent Dirichlet Allocation)
- Alternativas:
  - HDP (Hierarchical Dirichlet Process)
  - LSI (Latent Semantic Index)

Explicar melhor LDA?

- Dirichlet Distribution

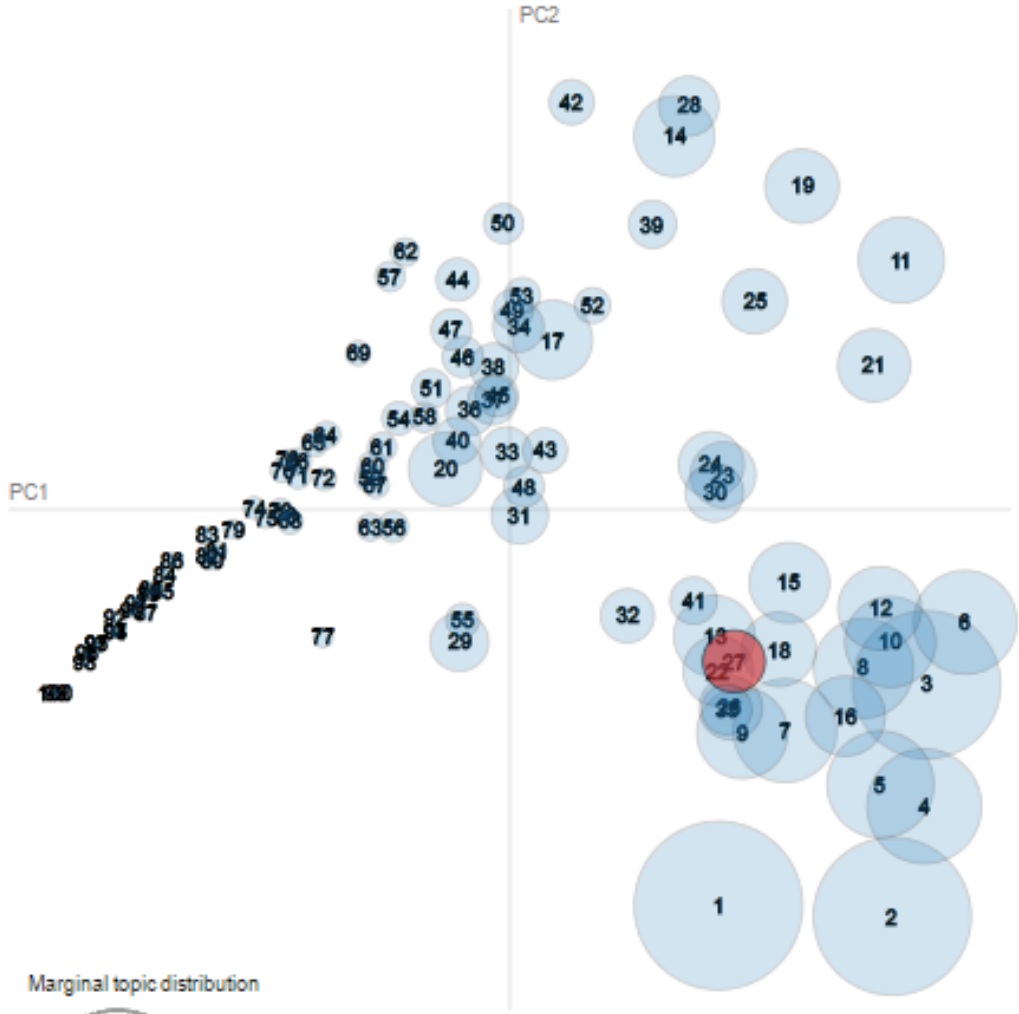
Selected Topic: 27 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)

λ = 0.48

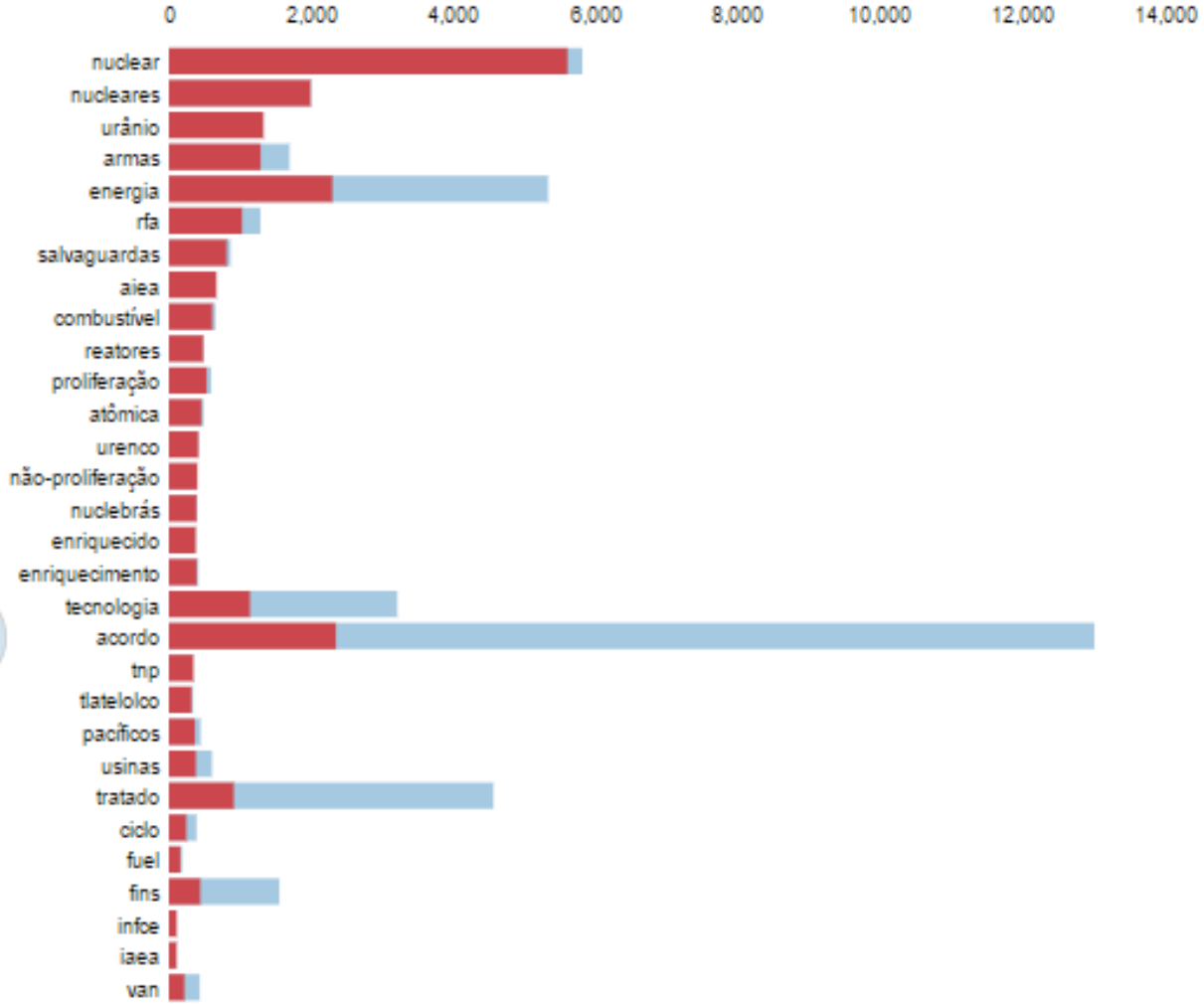
0.0 0.2 0.4 0.6 0.8 1.0

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

- O processo de geração de tópicos em geral demanda a colaboração de um especialista,
- Análise exaustiva de tópicos, tokens e documentos.
- Formulação interna de índices de coesão.
- É possível automatizar parte do processo. Ver, por exemplo, o paper "Exploring the Space of Topic Coherence Measures"

Named-entity recognition

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	Espaço reservado a autenticação mecânica
HORA DA TRANSMISSÃO	
INICIAIS DO OPERADOR	
INDICAÇÕES DE SERVIÇOS TAXADOS	URGENTE

DESTINATÁRIO: Embaixador Azeredo da Silveira

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

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

TEXTO E ASSINATURA — ENDEREÇO

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

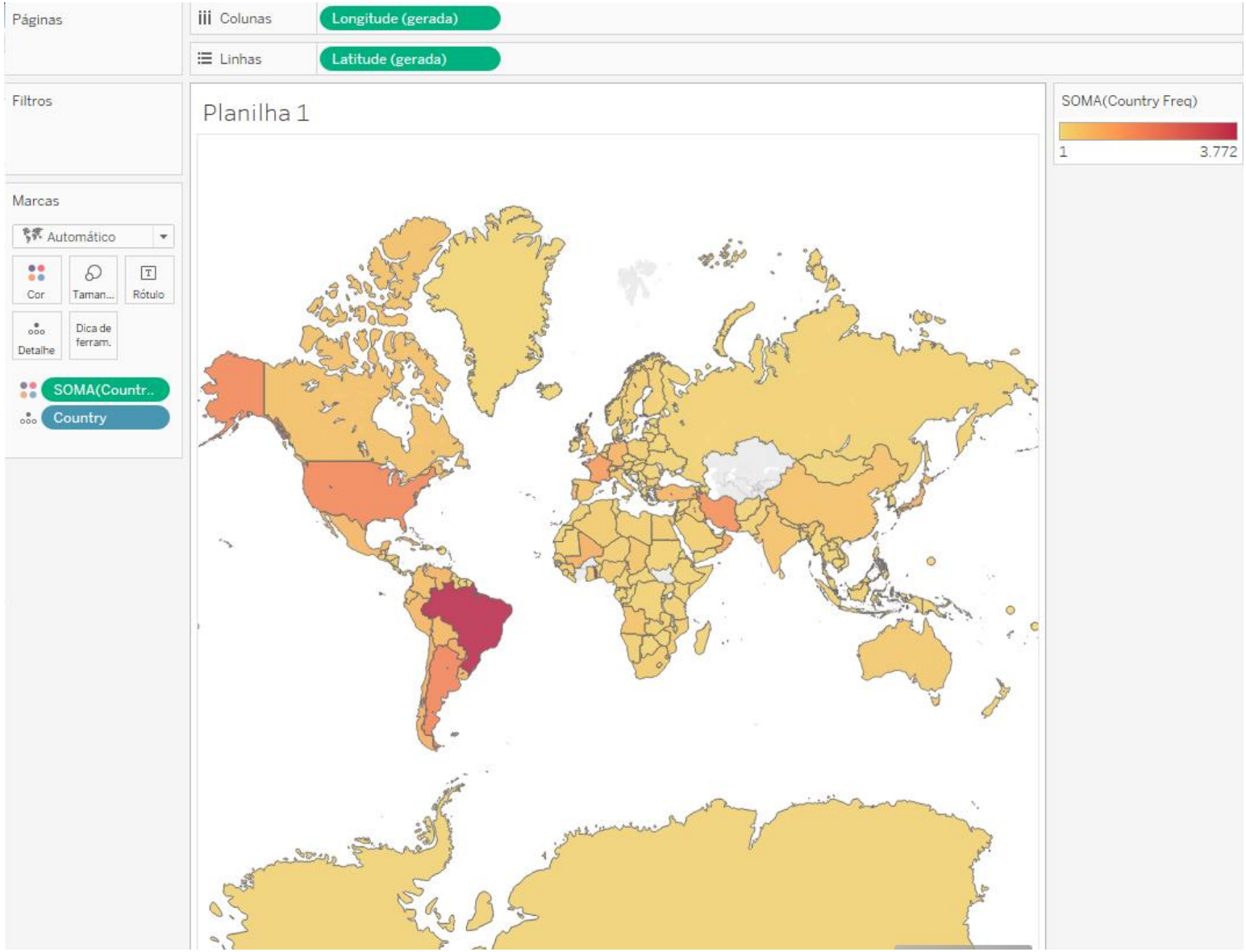
2530-007-0051 162 x 229 mm.

I - 46 Au

Named-entity recognition

- Ferramentas:
  - palavras
  - Stanford NLP

Exemplo de aplicação



Obrigado

Hugo Barreto

•

Marcelo Barata Ribeiro
marcelobbribeiro@gmail.com