Modelagem de tópicos

Utilização de modelagem de tópicos para identificar os assuntos mais discutidos sobre o Haiti nas redes sociais

Jod Fedlet Pierre
fedletpierre15@gmail.com
Autor

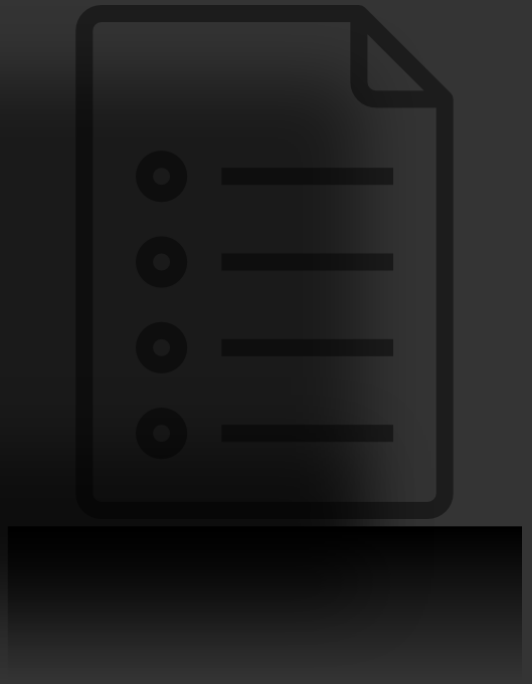
Denio Duarte
duarte@uffs.edu.br
Orientador

Sumário

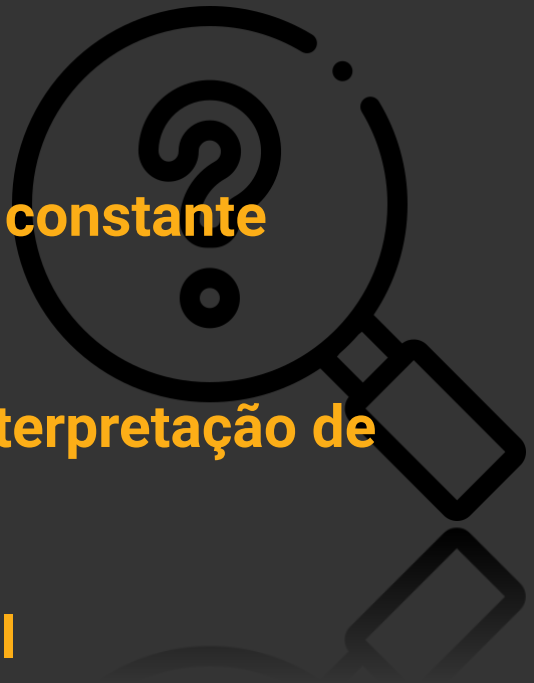
- Problematização
- Justificativa
- Referencial teórico
  - Haiti
  - Twitter
  - Modelagem de tópicos

Sumário

- Objetivos
- Trabalhos relacionados
- Metodologia
- Cronograma



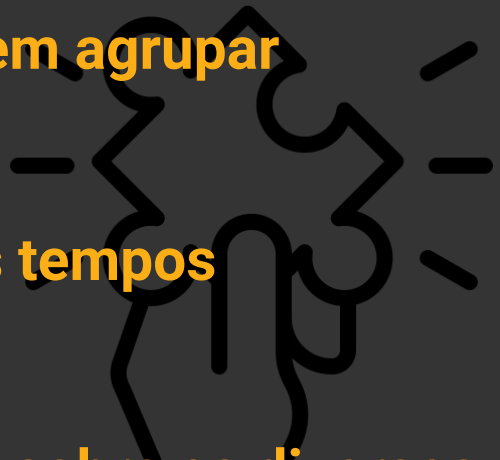
Problematização



- Compartilhamento de dados em tempo constante
- Grandes volumes de dados
- Agrupamento, análise, classificação, interpretação de dados compartilhados
- Tempo enorme de agrupamento manual
- Descoberta de padrões ou assuntos mais abordados

Justificativa

- Facilidade da modelagem de tópicos em agrupar coleção de documentos
- Acontecimentos no Haiti ao longo dos tempos



- Disseminação de informações na web sobre os diversos acontecimentos no Haiti em língua portuguesa

Haiti

- **Período pré-colonial**

  Marcado por: Indígenas
- **Período colonial**

  Marcado por: Espanhóis e franceses
- **Período pós-colonial**

  Marcado por: Ato da independência

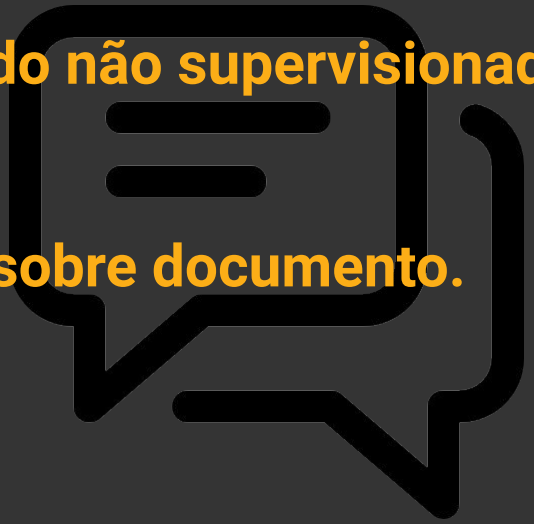
Posição geográfica do Haiti em relação às Américas

Twitter

- Lançado em 2006
- Disseminação de informação
- Postagens ou tweets, principalmente, no formato de textos de até 280 caracteres
- Uso do # para destacar assuntos principais do tweet

Modelagem de tópicos

- Conjunto de algoritmos de aprendizado não supervisionado
- Documentos são mistura de tópicos
- Tópicos são distribuição de palavras sobre documento.



Topics

gene	0.04
dna	0.02
genetic	0.01
. . .	

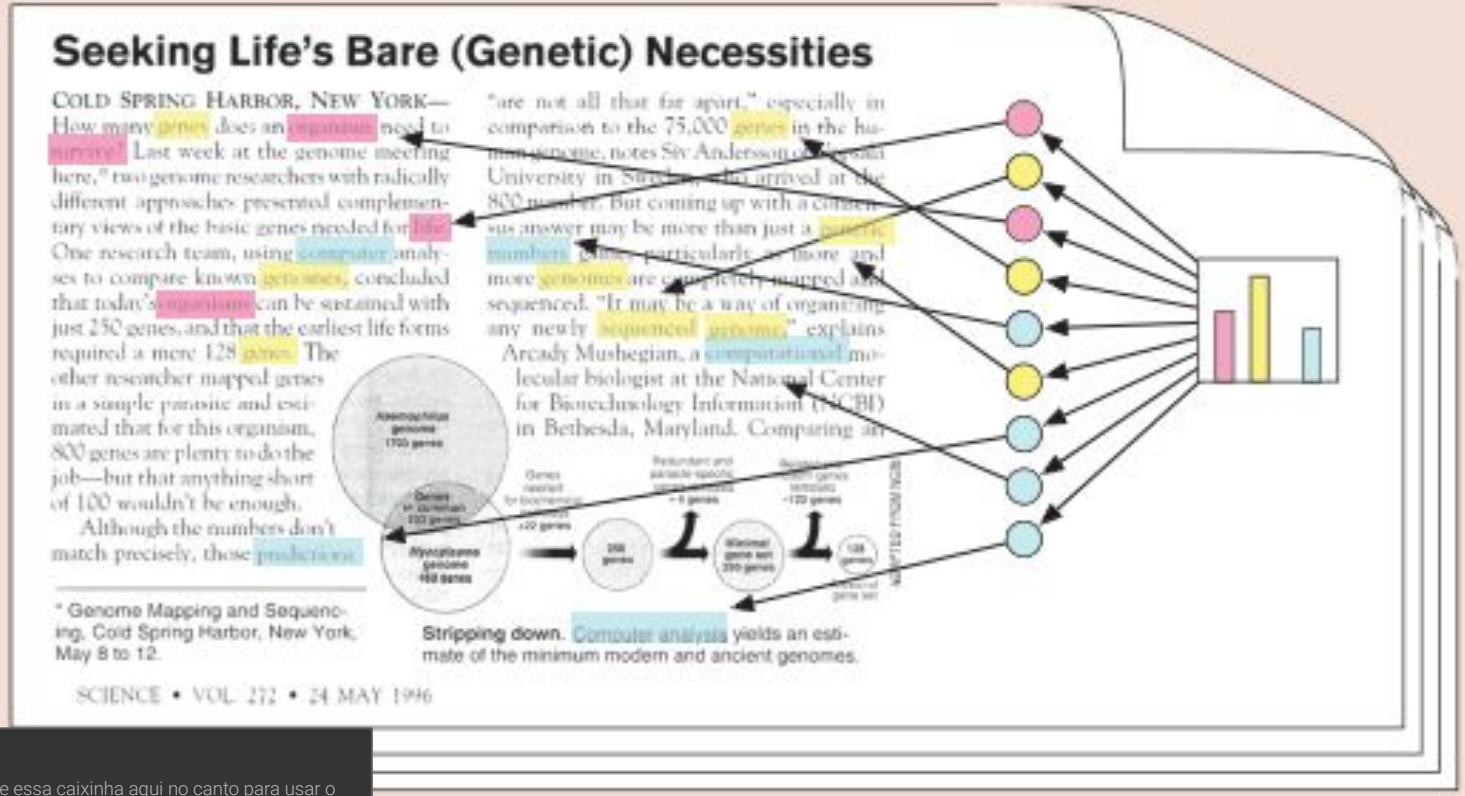
life	0.02
evolve	0.01
organism	0.01
. . .	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Foco total na imagem

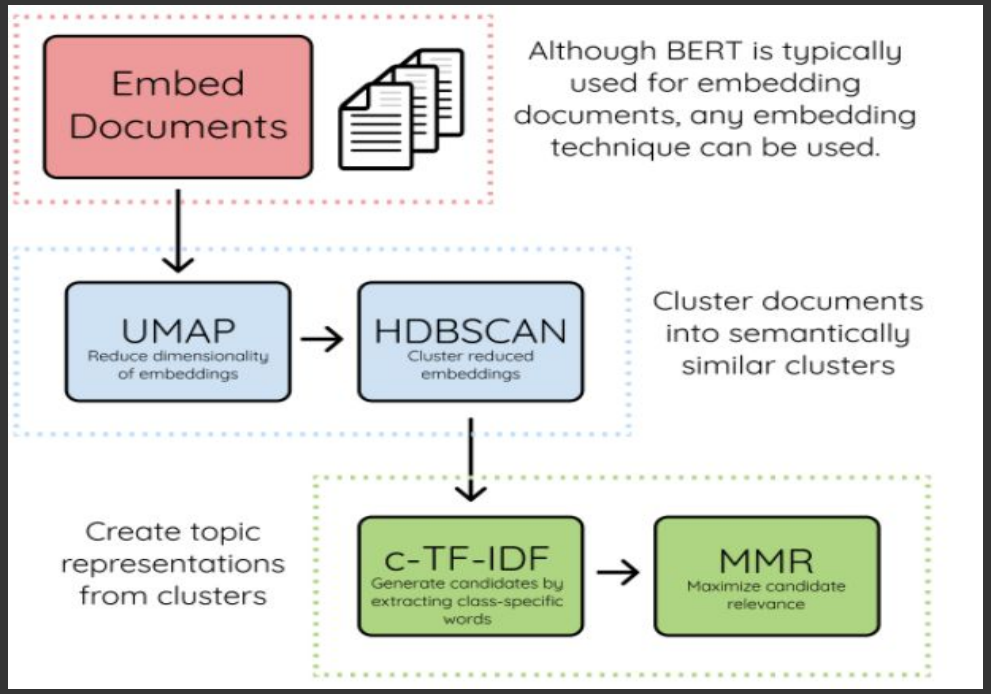
Se você quer destacar a imagem, utilize essa caixinha aqui no canto para usar o mínimo de espaço possível.

Modelagem de tópicos

BERTOPIC

- **Baseado no BERT (Bidirectional Representations from Transformers)**
  - BERT -> arquitetura neural;
    - Transformador: unidade básica
      - codificador: leitura do texto
      - decodificador: compreensão do texto

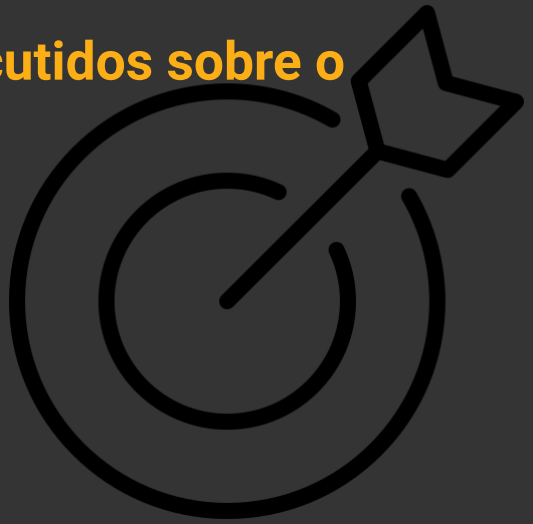
- **TF-IDF (term frequency - inverse document frequency)**


Embed Documents
Although BERT is typically used for embedding documents, any embedding technique can be used.
UMAP
Reduce dimensionality of embeddings
HDBSCAN
Cluster reduced embeddings
Cluster documents into semantically similar clusters
Create topic representations from clusters
c-TF-IDF
Generate candidates by extracting class-specific words
MMR
Maximize candidate relevance

Objetivos

Objetivo geral

- **Explorar e analisar os assuntos mais discutidos sobre o Haiti na rede social Twitter.**



Objetivos

Objetivos específicos

- Definir a string para filtrar tweets;
- Coletar os dados no período de datas a ser definido;
- Realizar pré-processamento dos dados;
- Codificar e treinar o modelo;

Objetivos

Objetivos específicos

- **Encontrar o melhor número de tópicos a serem extraídos;**
- **Extrair os tópicos extraídos;**
- **Analisar e classificar os resultados**

Trabalhos relacionados

PEREIRA, Mariana. Análise exploratória de tweets utilizando modelagem de tópicos para textos curtos: caso Olimpıadas Rio 2016. Universidade Federal da Fronteira Sul, 2019.

- Extração dos dados no Twitter;

  Strings:"rio2016","olimpiadas", "olimpiada","ceremoniaderura", "ceremoniadeencerramento", "jogosolimpicos", "olympics", "olympicgames", "openingceremony", "closingceremony"

- Pré-processamento dos tweets;

- BTM (Biterm Topic Model);

Trabalhos relacionados

PEREIRA, Mariana. Análise exploratória de tweets utilizando modelagem de tópicos para textos curtos: caso Olimpıadas Rio 2016. Universidade Federal da Fronteira Sul, 2019.

- **Quantidades de tópicos: 5, 10, 15, 20 e 30**

- **Análise dos resultados**

Rótulo	Palavras
Cerimônia de Abertura	pais, lindo, copa, mundo, brasileiro, povo, dinheiro, bonito, mal, festa, deus, gisele, demais, cerimonia, orgulho, maravilhosa, amo, parabens, pais, melhor, hoje, momento, atleta, anitta, caetano, gil, cantando, cantar, mc, fernanda, karol, montenegro, gilberto, jogo, medalha, ouro, futebol, esporte, selecao, olimpico, delegacao, bandeira, alemanha, mulher, vem, porta, grecia, aquecimento, indios, portugueses, global, parte, mostrar, historia, hino, musica, paulinho, viola, nacional, zeca, flamengo, pira, olimpica, acender, vanderlei, tocha, guga, cordeiro, pele, lima, chama, maracana, janeiro, cidade, estado, paises, samba, aula, carnaval, geografia, escola, falar, regina, ingles, case, falando, portugues, tremendo, santos, fala, homem.
Mídia	galvao, boca, cala, gloria, maria, globo, ouvir, bueno, falando, falar.
Político	temer, vaia, dilma, presidente, vaiado, golpista, michel, lula, medo, povo.
Transmissão das Olimpíadas	pokemon, vendo, hoje, cerimonia, mundo, casa, assistir, assistindo, estar, tv.

Trabalhos relacionados

HIDAYATULLAH, Ahmad Fathan et al. Twitter topic modeling on football news. In: IEEE. 2018 3rd International Conference on Computer and Communication Systems (ICCCS). [S.l.: s.n.], 2018. p. 467–471.

- Extração dos dados em contas oficiais como: @VIVAbola, @panditfootball, @detiksport, @Bolanet, @GOAL_ID;
- Pré-processamento dos tweets;
- LDA (Latent dirichlet Allocation);
- Quantidade de tópicos: 10
- Análise dos resultados


madrid
barcelona
real
ronaldo
messi
neymar
viva
atletico
main
zidane
borneo
barca
enrique
psg
kontrak
bintang
kukar
panjang
rp
mitra
anyar
vivacoidbolaread
jersey
sevilla
debut
enggan
pique
gabung
bikin
kalah
comeback

Trabalhos relacionados

ASGHARI, Mohsen; SIERRA-SOSA, Daniel; ELMAGHRABY, Adel. Trends on health in social media: Analysis using twitter topic modeling. In: IEEE. 2018 IEEE international symposium on signal processing and information technology (ISSPIT). [S.l.: s.n.], 2018. p. 558–563.

- **Extração dos dados no Twitter;**

  Palavras-chave: healthcare, health, doctors, homecare, digitalhealth and digital health.

- **Pré-processamento dos tweets;**
  - ❑ Lematização -> análise morfológica das palavras
  - ❑ Stemming -> análise flexional

Trabalhos relacionados

ASGHARI, Mohsen; SIERRA-SOSA, Daniel; ELMAGHRABY, Adel. Trends on health in social media: Analysis using twitter topic modeling. In: IEEE. 2018 IEEE international symposium on signal processing and information technology (ISSPIT). [S.l.: s.n.], 2018. p. 558–563.

- LDA e Word2Vect;
- Quantidade de tópicos: 5
- Análise dos resultados

  Tópico: Care

  Estados mais interessado: NY e California (80%)

  Estado menos interessado: Virgínia (41%)

Trabalhos relacionados

OH, Onook; KWON, Kyounghee Hazel; RAO, H Raghav. An exploration of social media in extreme events: Rumor theory and Twitter during the Haiti earthquake 2010, 2010.

- **Extração dos dados em contas oficiais como: #haitiearthquake;**
- **Pré-processamento dos tweets;**
- **ANOVA de Friedman;**

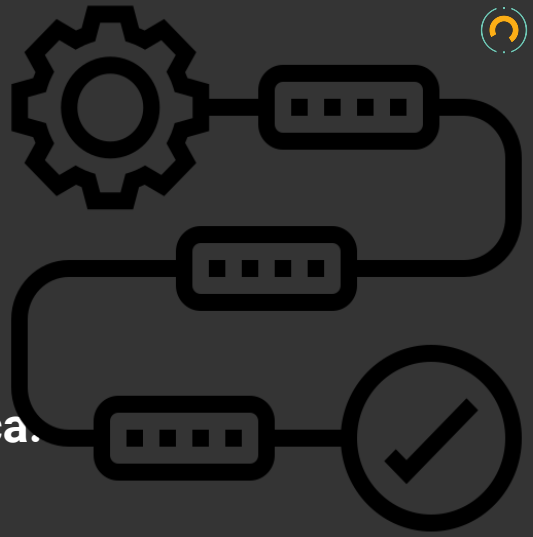
Trabalhos relacionados

OH, Onook; KWON, Kyounghee Hazel; RAO, H Raghav. An exploration of social media in extreme events: Rumor theory and Twitter during the Haiti earthquake 2010, 2010.

- **Análise dos resultados**

  palavras como: blog, CNN, picture, list, info, report e update representam declarações de autenticação.

Metodologia



- **Início dos estudos**

  -> Definição do tema e pesquisa bibliográfica.

- **Coleta dos dados**

  -> Extração de tweets com a palavra-chave ‘#haiti’ na API do Twitter.

- **Pré-processamento dos dados**

  -> Limpeza dos tweets extrádos

- **Aplicação do Bertopic**

  -> Codificação e treino do modelo

Metodologia

- **Experimento**

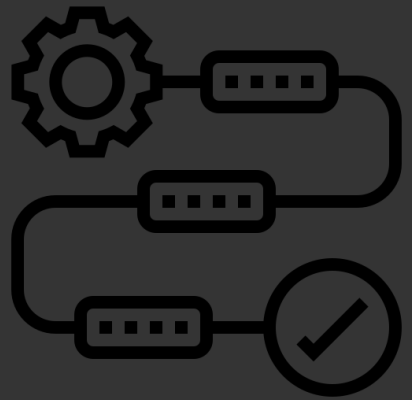
  -> Definição do número de tópicos a serem extraídos;

  -> Testes do modelo;

  -> Rotulação dos tópicos.

- **Análise dos resultados**

  -> Interpretação dos tópicos



Cronograma

Atividades	Nov		Dez		Jan		Fev		Mar		Abr		Mai		Jun		Jul		Ago		Set	
Início dos estudos	X	X	X	X	X	X	X	X	X	X												
Coleta dos dados											X	X										
Pré-processamento dos dados													X	X								
Aplicação do Bertopic														X	X							
Experimentação															X	X	X					
Análise dos resultados																	X	X	X	X		
Redação da monografia												X	X	X	X	X	X	X	X	X	X	X

Modelagem de tópicos

Utilização de modelagem de tópicos para identificar os assuntos mais discutidos sobre o Haiti nas redes sociais

Obrigado!

Créditos

Orientação

Denio Duarte
duarte@uffs.edu.br

Ícones

mangsaabguru
flaticon.com/authors/mangsaabguru

Freepik
flaticon.com/authors/freepik

Smashicons
flaticon.com/authors/smashicons

Icons made by mangsaabguru, Freepik and Smashicons
from www.flaticon.com

CIÊNCIA DA
COMPUTAÇÃO UFFS