INDEORUM

COMO CONVENCER QUE SEUS DADOS SÃO INTERESSANTES

Processamento de Linguagem Natural e Storytelling na prática

MATHEUS EDUARDO RODRIGUES FREITAG
CIENTISTA DE DADOS

INDEORUM

SUMÁRIO

NLP?

O que é isso, de onde veio, e porque eu deveria me importar?

INDEORUM

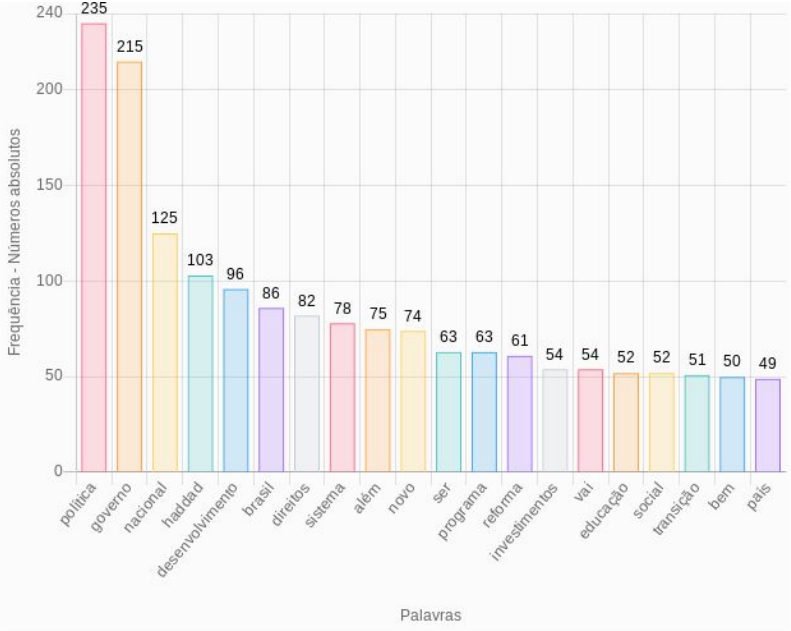
IDENTIFICANDO

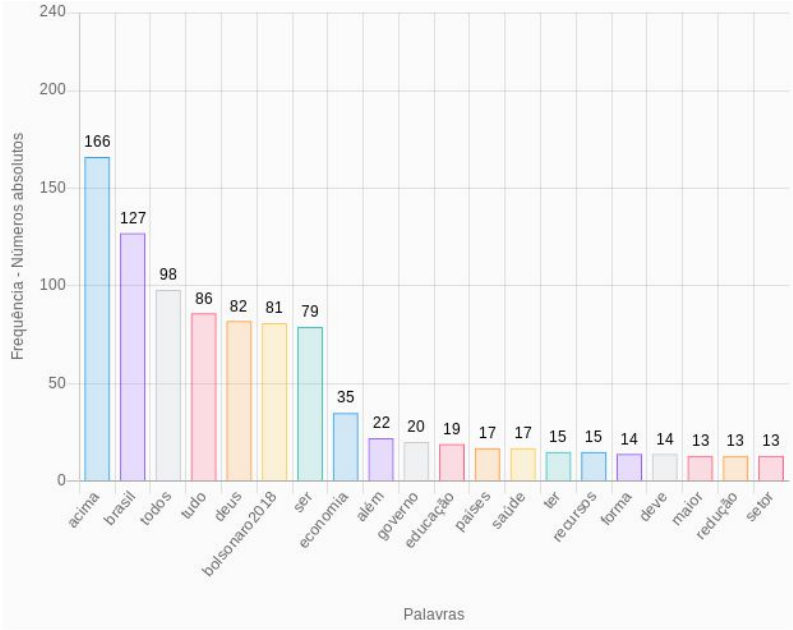
Vamos imaginar que o G1 contratou a Indeorum para categorizar seus artigos

ENTENDENDO

Vamos imaginar um país fictício que esteja em época de eleições.

INDEORUM


Frequência - Números absolutos
240
200
150
100
50
0
235
215
125
103
96
86
82
78
75
74
63
63
61
54
54
52
52
51
50
49
politica
governo
nacional
haddad
desenvolvimento
brasil
direitos
sistema
além
novo
ser
programa
reforma
investimentos
vai
educação
social
transição
bem
país
Palavras


Frequência - Números absolutos
240
200
150
100
50
0
166
127
98
86
82
81
79
35
22
20
19
17
17
15
15
14
14
13
13
13
acima
brasil
todos
tudo
deus
bolsonaro2018
ser
economia
além
governo
educação
países
saúde
ter
recursos
forma
deve
maior
reduçáo
setor
Palavras

INDEORUM

TF-IDF

TF: Term Frequency
Quantas vezes um termo aparece em um texto

TF(t) = (# de vezes que o termo aparece) / (# de termos no texto)

IDF: Inverse Document Frequency
Quão especial é um termo

IDF(t) = log(# de documentos) / (# de documentos que t aparece)

Um documento com 100 palavras, onde a palavra "gato" aparece 3 vezes possui TF igual a (3 / 100) = 0.03

Agora considere que temos 10M de documentos e que "gato" aparece em 1K deles, portanto o IDF é log(10,000,000 / 1,000) = 4.

Então o TF-IDF se torna o produto desses dois valores: 0.03 * 4 = 0.12

INDEORUM

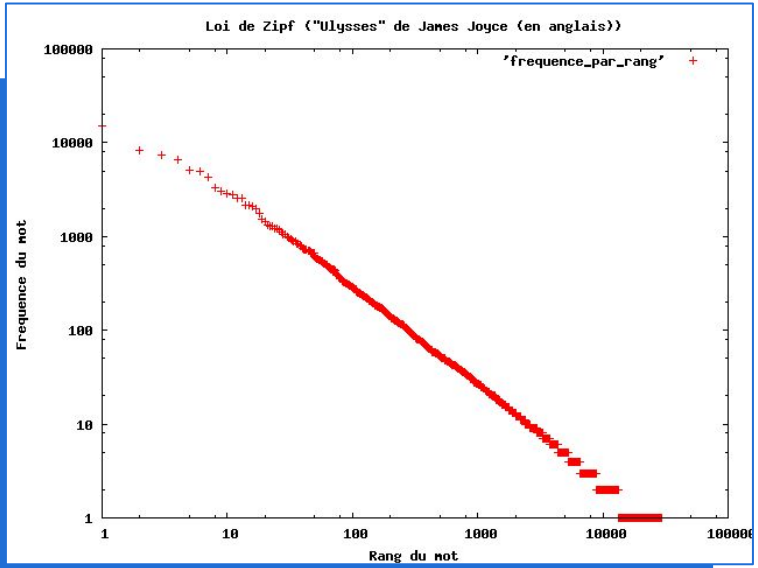
MATEMÁTICA????

- Se uma palavra aparece muito em um documento, ela é importante. Score sobe.

- Mas se essa palavra aparece em muitos documentos, ela não serve de identificador único. O Score cai.

  Então palavras sem muito significado e que aparecem muito, como artigos e preposições, recebem score baixo, e palavras que aparecem muito em um único documento recebem score alto

LEI DE ZIPF



Em Ulysses:

- A palavra mais comum surgia 8000 vezes
- A décima, 800 vezes
- A centésima, 80 vezes
- A milésima, 8 vezes.

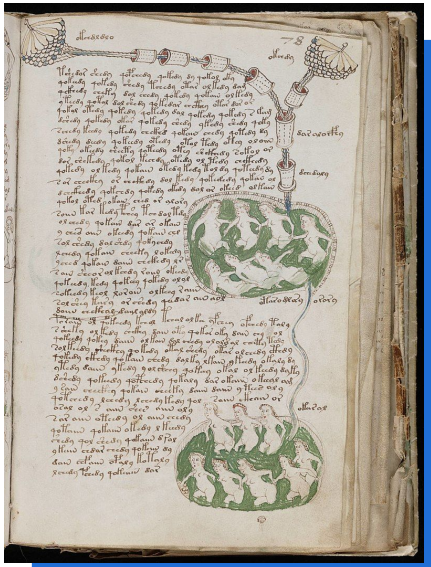
Análise Presidencial

~18k palavras

~11k palavras (sem stopwords)

~4.5k palavras únicas

Manuscrito Voynich



COMPARANDO

Como fazemos para comparar dois corpus de texto? E como quantificar as semelhanças e diferenças?

QUÃO PARECIDOS SÃO ESSES TEXTOS?

“[...] A limpeza e secagem correta dos pés após o banho é um fator decisivo para evitar a proliferação de fungos que podem causar odores fortes[...]”

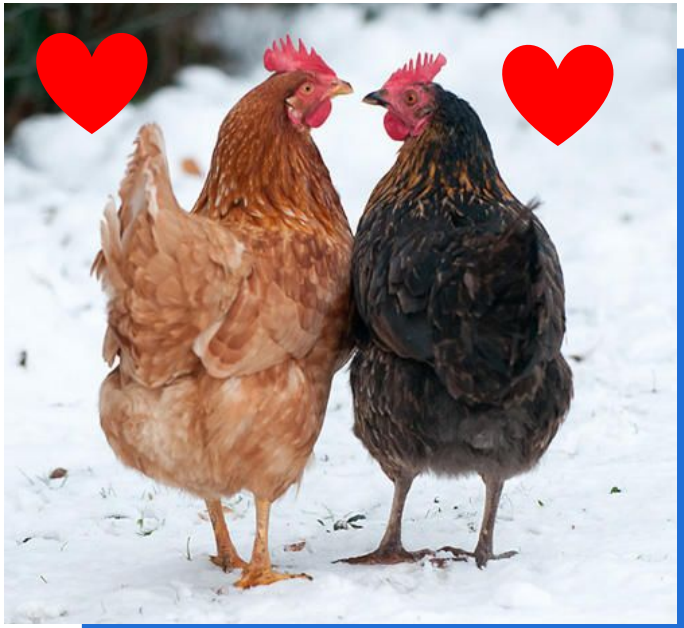
“[...] A limpeza e secagem dos pés é um fator decisivo para evitar a proliferação de fungos que causam odores fortes[...]”

“[...] A limpeza e secagem correta dos pés após o banho auxilia no combate a proliferação de fungos e do mau odor nos pés [...]”

QUÃO PARECIDOS SÃO ESSES TEXTOS?

Matheus Freitag	Mateus Freitag	Mateus Freitas

DISTÂNCIA DE LEVENSHTEIN

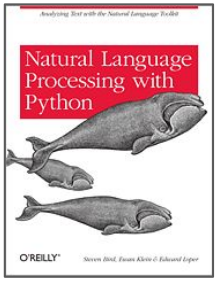


SENTINDO

Como máquinas podem compreender emoções inerentemente humanas?

EXPANDINDO

- **Applied Data Science with Python Specialization** https://bit.ly/2danP4n
- **Natural Language Processing with Python**



- **Recurrent Neural Networks**
- **Kaggle Competitions**

FUTURO

Palavras sofrem muitas variações linguísticas (especialmente línguas românticas)

STEMMING

LEMMATIZAÇÃO

INDEORUM

OBRIGADO!

Alguma pergunta?

Matheus Freitag

/matheusfreitag