Natural Language Processing

Cayan Atreio Portela

UniCEUB

December 1, 2022

Processamento de Linguagem Natural

- Definição

Google Cloud Plataform

"Como um branch de inteligência artificial, o processamento de linguagem natural (PLN) usa machine learning para processar e interpretar texto e dados."

- Definição

Wikipedia

"Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents."

Aplicações

- Exemplos de casos de uso

  - → Tradução e correção automática.
  - → Speech recognition.
  - → Sumarização de textos e documentos.
  - → Análise de sentimentos.
  - → Sistemas de recomendação.
  - → Chatbot.
  - → Jurimetria.

- Desafios
  - → Dados não estruturados
  - → Grande quantidade de informações documentos.
  - → Stemming, Tagging, Lemmatization.
  - → Análise contextual.

Frameworks

- Algoritmos entendem números, não palavras.

Para utilizar informações textuais como *input* em algoritmos e modelos de aprendizado de máquina, deve ser feito um mapeamento de palavras para representações numéricas.

Frameworks

- Algoritmos entendem números, não palavras.

Para utilizar informações textuais como *input* em algoritmos e modelos de aprendizado de máquina, deve ser feito um mapeamento de palavras para representações numéricas.

- Frequencia
  - Bag of Words
  - TF-IDF

→ Contagem

- Embedding
  - Word2Vec
  - BERT

→ Dimensão vetorial

Frameworks

- Algoritmos entendem números, não palavras.

Para utilizar informações textuais como *input* em algoritmos e modelos de aprendizado de máquina, deve ser feito um mapeamento de palavras para representações numéricas.

- Frequencia
  - Bag of Words
  - **TF-IDF**

→ Contagem

- Embedding
  - Word2Vec
  - BERT

→ Dimensão vetorial

TF-IDF

Term Frequency - Inverse Document Frequency

- Termo Frequência - Frequência Inversa do Documento é uma medida usada para quantificar a relevância de representações textuais em um documento, dentre uma coleção de documentos.

TF-IDF

Term Frequency - Inverse Document Frequency

- Termo Frequência - Frequência Inversa do Documento é uma medida usada para quantificar a relevância de representações textuais em um documento, dentre uma coleção de documentos.

- Utilidades
  - → Classificação e similaridade.

TF-IDF

Term Frequency - Inverse Document Frequency

- Termo Frequência - Frequência Inversa do Documento é uma medida usada para quantificar a relevância de representações textuais em um documento, dentre uma coleção de documentos.

- Utilidades
  - → Classificação e similaridade.
  - → Representatividade para *input* em modelos de machine learning.

Term Frequency - Inverse Document Frequency

- Considera a frequência de uma palavra em relação à uma coleção de documentos. Exemplos de medidas:
  - → Número de ocorrências da palavra.
  - → Frequeência ajustada pelo tamanho do documento.
  - → Escala logarítmica (ex. log(1 + contagem)).
  - → Representação binária (1 se ocorre, 0 caso contrário).

Term Frequency - **Inverse Document Frequency**

- Frequência da palavras em relação ao conjunto de documentos. Considerando t a palavra a ser medida em N documentos (d) no conjunto, IDF é dado pela razão entre o número de documentos n sobre o número de documentos que contém a palavra $df(t)$.

$$IDF_{(t)} = log(\frac{n}{df(t)})$$

Logo,

$$\text{TF-IDF}_{(t,d)} = \text{TF} \times IDF_{(t)}$$

TF-IDF

Exemplo:

Notícia A

"Brasil está classificado para as oitavas de final"

8 palavras

Notícia B

"Argentina briga por uma vaga nas oitavas de final"

9 palavras

TF-IDF

Palavra	TF_A	TF_B	IDF	TF-IDF_1	TF-IDF_2
Brasil	$1/8$	0	$\log(2/1) = 0.69$	0.086	0
esta	$1/8$	0	$\log(2/1) = 0.69$	0.086	0
classificado	$1/8$	0	$\log(2/1) = 0.69$	0.086	0
para	$1/8$	0	$\log(2/1) = 0.69$	0.086	0
as	$1/8$	0	$\log(2/1) = 0.69$	0.086	0
oitavas	$1/8$	0	$\log(2/2) = 0.00$	0	0
de	$1/8$	0	$\log(2/2) = 0.00$	0	0
final	$1/8$	0	$\log(2/2) = 0.00$	0	0
Argentina	0	$1/9$	$\log(2/1) = 0.69$	0	0.077
briga	0	$1/9$	$\log(2/1) = 0.69$	0	0.077
por	0	$1/9$	$\log(2/1) = 0.69$	0	0.077
uma	0	$1/9$	$\log(2/1) = 0.69$	0	0.077
vaga	0	$1/9$	$\log(2/1) = 0.69$	0	0.077
nas	0	$1/9$	$\log(2/1) = 0.69$	0	0.077
oitavas	0	$1/9$	$\log(2/2) = 0.00$	0	0
de	0	$1/9$	$\log(2/2) = 0.00$	0	0
final	0	$1/9$	$\log(2/2) = 0.00$	0	0

TF-IDF

- Representação vetorial.

  - Noticia 1

$$\begin{bmatrix} 0.086_1 & 0.086_2 & 0.086_3 & 0.086_4 & 0.086_5 & 0_6 & 0_7 & \cdots & 0_{15} \end{bmatrix}$$

  - Noticia 2

$$\begin{bmatrix} 0_1 & \cdots & 0.077_9 & 0.077_{10} & 0.077_{11} & 0.077_{12} & 0.077_{13} & 0.077_{14} & \cdots & 0_{15} \end{bmatrix}$$

Representação numérica

- A dimensão vetorial fornece uma representação numérica da importância de cada termo em relação ao texto, considerando uma coleção de documentos.

- Considerando uma coleção de documentos, é possível calcular similaridades através dos vetores mapeados.

Representação numérica

- A dimensão vetorial fornece uma representação numérica da importância de cada termo em relação ao texto, considerando uma coleção de documentos.

- Considerando uma coleção de documentos, é possível calcular similaridades através dos vetores mapeados.

→ Similaridade de cosseno.

Similaridade

1. *"Dois dos melhores jogadores das últimas décadas correm o risco de se despedir definitivamente da Copa do Mundo nesta quarta-feira ... Polônia, de Robert Lewandowski, e a Argentina, de Lionel Messi, se enfrentam no estádio 974 pela última rodada do Grupo C."*
2. "A modelo Jessica Turini, apontada em agosto deste ano como o novo affair de Neymar ... Bruna Marquezine, ex-namorada do craque brasileiro e seu relacionamento mais duradouro, conhecido e assumido."
3. "A CBF atualizou nesta terça-feira o boletim médico do lateral-direito Danilo e do atacante Neymar ... lesões de tornozelo e que não vão enfrentar Camarões na sexta-feira."
4. "Dos 26 jogadores convocados pelo técnico Tite para defender a Seleção na Copa do Mundo, sete aguardam ansiosamente pela oportunidade de estrear no Catar..."

Similaridade

- Matriz de similaridade

$$
\begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \\ n_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} n_1 & n_2 & n_3 & n_4 \end{array} \\
\begin{bmatrix} 1 & 0.02 & 0.03 & 0.09 \\ 0.02 & 1 & 0.03 & 0.015 \\ 0.03 & 0.03 & 1 & 0.02 \\ 0.09 & 0.015 & 0.02 & 1 \end{bmatrix}
\end{array}
$$

Similaridade

- Matriz de similaridade

$$
\begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \\ n_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} n_1 & n_2 & n_3 & n_4 \end{array} \\
\left[\begin{array}{cccc}
1 & 0.02 & 0.03 & 0.09 \\
0.02 & 1 & 0.03 & 0.015 \\
0.03 & 0.03 & 1 & 0.02 \\
0.09 & 0.015 & 0.02 & 1
\end{array}\right]
\end{array}
$$

- Consumo em notícia 1, levaria à recomendação para notícia 2

Similaridade

- Matriz de similaridade

$$
\begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \\ n_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} n_1 & n_2 & n_3 & n_4 \end{array} \\
\begin{bmatrix} 1 & 0.02 & 0.03 & 0.09 \\ 0.02 & 1 & 0.03 & 0.015 \\ 0.03 & 0.03 & 1 & 0.02 \\ 0.09 & 0.015 & 0.02 & 1 \end{bmatrix}
\end{array}
$$

- Consumo em notícia 1, levaria à recomendação para notícia 2

https://github.com/Cayan-Portela/TCF/ceub_nlp.R

Exercício

Calcular os vetores TF-IDF$_a$ e TF-IDF$_B$ considerando os dois textos abaixo.

Notícia A

"O gato perseguiu o rato"

Notícia B

"O rato pegou o queijo e fugiu do gato"

Exercício

Palavra	TF_A	TF_B	IDF	TF-IDF_1	TF-IDF_2
o	.	0	log(/) = .	.	.
gato	.	0	log(/) = .	.	.
perseguiu	.	0	log(/) = .	.	.
o	.	0	log(/) = .	.	.
rato	.	0	log(/) = .	.	.
o	0	.	log(/) = .	.	.
rato	0	.	log(/) = .	.	.
pegou	0	.	log(/) = .	.	.
o	0	.	log(/) = .	.	.
queijo	0	.	log(/) = .	.	.
e	0	.	log(/) = .	.	.
fugiu	0	.	log(/) = .	.	.
do	0	.	log(/) = .	.	.
gato	0	.	log(/) = .	.	.

.

Exercício

Palavra	TF_A	TF_B	IDF	TF-IDF_1	TF-IDF_2
o	$1/5$	0	$\log(2/2) = 0.00$	0	0
gato	$1/5$	0	$\log(2/2) = 0.00$	0	0
perseguiu	$1/5$	0	$\log(2/1) = 0.69$	0.14	0
o	$1/5$	0	$\log(2/2) = 0.00$	0	0
rato	$1/5$	0	$\log(2/2) = 0.00$	0	0
o	0	$1/9$	$\log(2/2) = 0.00$	0	0
rato	0	$1/9$	$\log(2/2) = 0.00$	0	0
pegou	0	$1/9$	$\log(2/1) = 0.69$	0	0.08
o	0	$1/9$	$\log(2/2) = 0.00$	0	0
queijo	0	$1/9$	$\log(2/1) = 0.69$	0	0.08
e	0	$1/9$	$\log(2/1) = 0.69$	0	0.08
fugiu	0	$1/9$	$\log(2/1) = 0.69$	0	0.08
do	0	$1/9$	$\log(2/1) = 0.69$	0	0.08
gato	0	$1/9$	$\log(2/2) = 0.00$	0	0

.

Bibliografía

1. Christian, H., Agus, M. P., Suhartono, D. (2016). *Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)*. ComTech: Computer, Mathematics and Engineering Applications, 7(4), 285-294.
2. Cloud, T. P. U. (2022). Google cloud. URL https://cloud. google. com/products/ai.