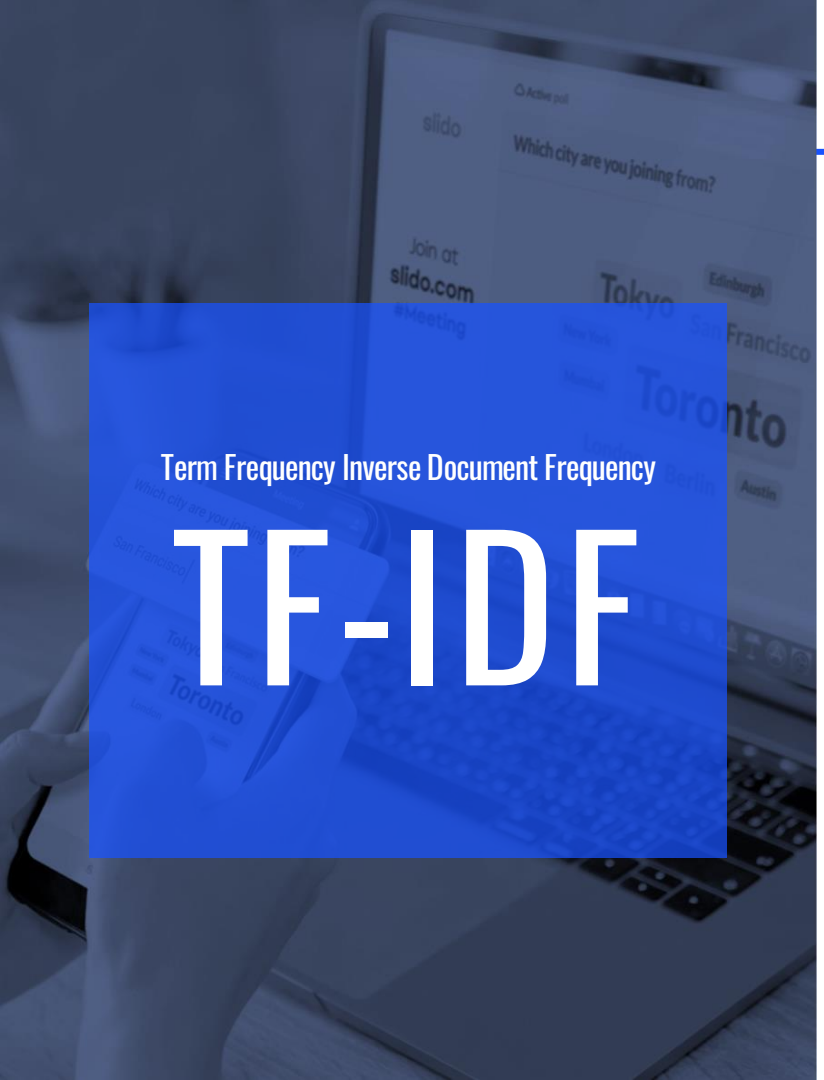
PUC Minas Virtual

PROCESSAMENTO DE LINGUAGEM NATURAL

TF-IDF



Term Frequency Inverse Document Frequency

TF-IDF

Frequência relativa de uma palavra em um documento em comparação a sua frequência em todos os documentos

$$W_{x,y} = tf_{x,y} \times \log \frac{N}{df_x}$$

TF-IDF

Termo x no documento y

$tf_{x,y}$ = frequência de x em y

df_x = número de documentos que contém x

N = número total de documentos

Exemplo de TF-IDF

C1 Este filme é muito assustador e longo

C2 Este filme não é assustador e é lento, muito lento

C3 Este filme é assustador e bom

Exemplo de TF-IDF

		Tokens
C1	~~Este~~ filme ~~é muito~~ assustador ~~e~~ longo	3
C2	~~Este~~ filme não ~~é~~ assustador ~~e é~~ lento ~~, muito~~ lento	5
C3	~~Este~~ filme ~~é~~ assustador ~~e~~ bom	3

Term Frequency

$$tf_{x,y} = \frac{df_{x,y}}{\sum_k df_{k,y}}$$

Exemplo de TF

Termo	C1	C2	C3	TF C1	TF C2	TF C3
filme	1	1	1	1/3	1/5	1/3
assustador	1	1	1	1/3	1/5	1/3
longo	1	0	0	1/3	0/5	0/3
não	0	1	0	0/3	1/5	0/3
lento	0	2	0	0/3	2/5	0/3
bom	0	0	1	0/3	0/5	1/3

Inverse Document Frequency

$$IDF_x = \log \frac{N}{df_x}$$

Exemplo de IDF

Termo	C1	C2	C3	IDF
filme	1	1	1	log(3/3) = 0,000
assustador	1	1	1	log(3/3) = 0,000
longo	1	0	0	log(3/1) = 0,477
não	0	1	0	log(3/1) = 0,477
lento	0	2	0	log(3/2) = 0,176
bom	0	0	1	log(3/1) = 0,477

Exemplo de TF-IDF

Termo	C1	C2	C3	TF C1	TF C2	TF C3	IDF	TF-IDF C1	TF-IDF C2	TF-IDF C3
filme	1	1	1	1/3	1/5	1/3	log(3/3) = 0,000	0,000	0,000	0,000
assustador	1	1	1	1/3	1/5	1/3	log(3/3) = 0,000	0,000	0,000	0,000
longo	1	0	0	1/3	0/5	0/3	log(3/1) = 0,477	0,159	0,000	0,000
não	0	1	0	0/3	1/5	0/3	log(3/1) = 0,477	0,000	0,095	0,000
lento	0	2	0	0/3	1/5	0/3	log(3/2) = 0,176	0,000	0,035	0,000
bom	0	0	1	0/3	0/5	1/3	log(3/1) = 0,477	0,000	0,000	0,159

Vantagens do TF-IDF

Quantifica a importância da palavra

Desvantagens do TF-IDF

- Tamanho do vetor
- Não entende novas palavras
- Nenhuma informação semântica

Desvantagens do TF-IDF

- Tamanho do vetor
- Não entende novas palavras
- Nenhuma informação semântica

Desvantagens do TF-IDF

- Tamanho do vetor
- Não entende novas palavras
- Nenhuma informação semântica

```
1. from sklearn.feature_extraction.text import TFidfVectorizer


2. # cria a instância do TFidfVectorizer
3. vectorizer = TFidfVectorizer()


4. # transforma o corpus em uma matriz numérica
5. tfidf = vectorizer.fit_transform(corpus)
```

Implementação do TF-IDF

PUC Minas
Virtual