

Processamento de Linguagem Natural e Processamento de Imagens

Aula 2

James Moraes de Almeida



MINISTÉRIO DA
EDUCAÇÃO

MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO



Conteúdo do curso

- Qual o plano?
 - Processamento de linguagem natural (PNL ou NLP do inglês).
 - ~~Mineração de dados.~~
 - ~~Histórico.~~
 - ~~Tokenização.~~
 - ~~Pré-processamento.~~
 - **Similaridade de sentenças.**
 - **Bag-Of-Words.**
 - **TF-IDF.**
 - Modelagem de tópicos.
 - Grandes Modelos de Linguagem para extração de dados.
- Material principal: <https://www.youtube.com/@futuremojo>

Representando palavras com vetores

- É bastante útil representarmos as palavras como vetores.
 - Porque?
 - Já temos muita matemática desenvolvida para vetores.
 - Não precisamos reinventar a roda.

Representando palavras com vetores

- Como fazer isto?
 - Vamos analisar as seguintes sentenças.
 - João gosta muito de estudar python.
 - João gosta de cozinhar miojo.
 - Qual o nosso vocabulário?
 - João
 - gosta
 - muito
 - de
 - estudar
 - python
 - cozinhar
 - miojo

Representando palavras com vetores

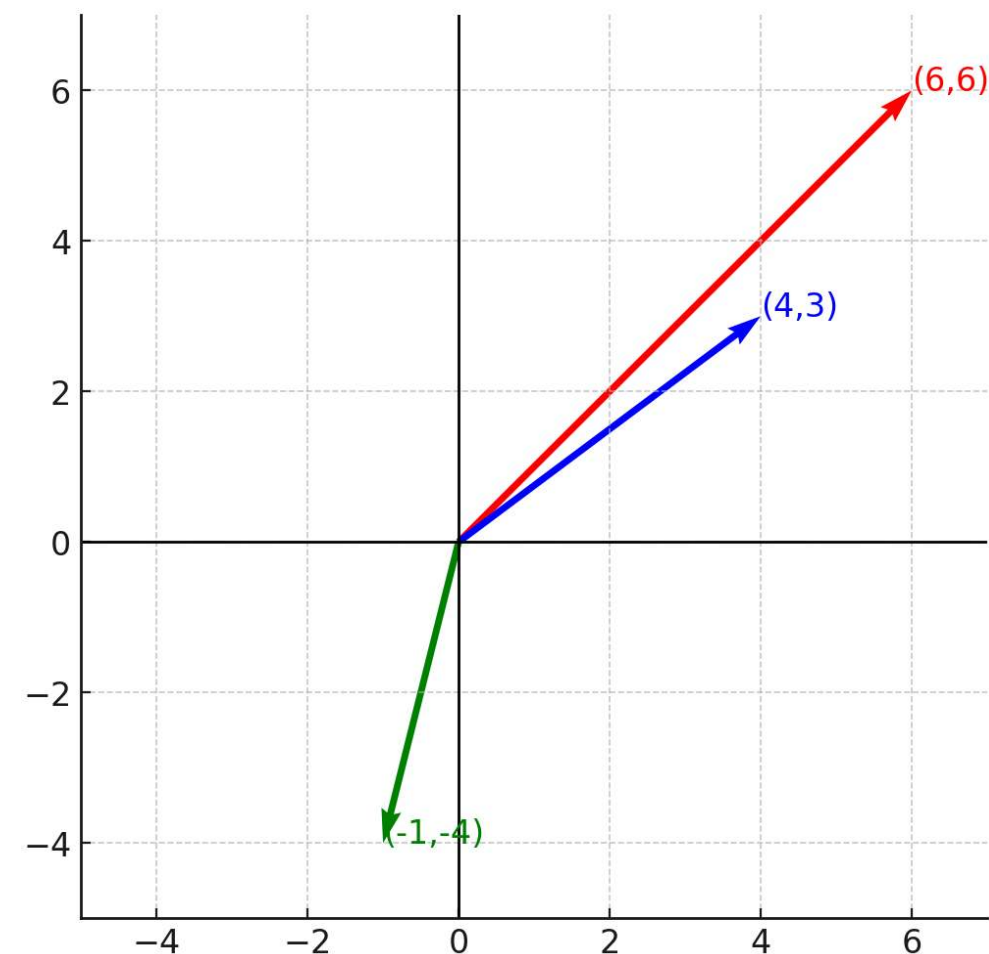
- Como fazer isto?
 - Vamos analisar as seguintes sentenças.
 - João gosta muito de estudar python. —————> Sentença 1
 - João gosta de cozinhar miojo. —————> Sentença 2

	João	gosta	muito	de	estudar	python	cozinhar	miojo	.
Sentença 1 —————>	1	1	1	1	1	1	0	0	1
Sentença 2 —————>	1	1	0	1	0	0	1	1	1

- Isto se chama "sacola de palavras **binária**".
 - *Binary Bag-of-Words (BOW)*.
- Também há a "*Frequency BOW*"

Medindo a semelhança entre sentenças

- Vamos pensar no espaço de vetores.
 - Dos vetores mostrados, quais são os mais "semelhantes"?
 - Definimos matematicamente uma semelhança.



Medindo a semelhança entre sentenças CNPEM

- Vamos pensar no espaço de vetores.
 - Dos vetores mostrados, quais são os mais "semelhantes"?
 - Definimos matematicamente uma semelhança.
 - Vamos definir a norma do vetor:

$$||\vec{v}|| = \sqrt{\sum_{i=1}^n \vec{v}^2}$$

- E a norma de um produto interno de vetores:

$$\frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| ||\vec{b}||}$$

Medindo a semelhança entre sentenças CNPEM

- Vamos pensar no espaço de vetores.
 - Dos vetores mostrados, quais são os mais "semelhantes"?
 - Definimos matematicamente uma semelhança.
 - Vamos definir a norma do vetor:

$$||\vec{v}|| = \sqrt{\sum_{i=1}^n \vec{v}^2}$$

- E a norma de um produto interno de vetores:

$$\frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| ||\vec{b}||} = \cos\theta$$

Medindo a semelhança entre sentenças CNPEM

- Agora vamos calcular as similaridades para alguns exemplos.
 - João gosta muito de estudar python.
 - João gosta de cozinhar miojo.
 - Maria vai viajar amanhã.

	João	gosta	muito	de	estudar	python	cozinhar	miojo	Maria	vai	viajar	amanhã
s_1	1	1	1	1	1	1	0	0	0	0	0	0
s_2	1	1	0	1	0	0	1	1	0	0	0	0
s_3	0	0	0	0	0	0	0	0	1	1	1	1

$$\cos(\vec{s}_1, \vec{s}_2) = 0.452$$

$$\cos(\vec{s}_1, \vec{s}_3) = 0.000$$

$$\cos(\vec{s}_2, \vec{s}_3) = 0.000$$

Medindo a semelhança entre sentenças CNPq

- Muito legal, mas é limitado a sentença de vocabulário.
 - Meu cachorro me mordeu.
 - O cão que pertence a mim abocanhhou-me.
 - O cosseno entre essas duas sentenças é zero, mas elas dizem o mesmo.
- Não captura semelhança entre sinônimos.
- Comparações estão limitadas ao vocabulário.
- Acaba com a ordem das palavras.
- Precisamos de algo mais refinado.

Medindo a semelhança entre sentenças CNPEM

- Podemos medir a frequência relativa de uma palavra.
 - Uma palavra pode ser comum em todo *corpus*.
 - Comparamos a frequência de um *token* no documento com no *corpus*.
 - Documento pode ser uma sentença, ou todo um abstract, ...
- Vamos primeiramente para a frequência em um documento.

$$TF('prato', d_1) = 10$$

$$TF('prato', d_2) = 1000$$

- Mais interessante utilizarmos o logaritmo:

$$TF(t, d) = \log(f_{t,d} + 1)$$

Medindo a semelhança entre sentenças

- Agora vamos definir a IDF (*Inverse Document Frequency*).

$$IDF(t, D) = \log \left(\frac{N}{n_t} \right)$$

Diagram illustrating the components of the IDF formula:

- N : Número de documentos
- n_t : Número de documentos que o termo t aparece.
- D : Corpus

- Se um termo t , aparece em todos documentos, sua IDF é zero.
- Caso um termo apareça em poucos documentos, sua IDF é maior.

Medindo a semelhança entre sentenças CNPEM

- Agora vamos definir a TF-IDF (*Term Frequency - Inverse Document Frequency*).
 - $TF - IDF(t, d) = TF(t, d) \times IDF(t, d)$
- Somente multiplicamos a TF e a IDF.
 - Quanto mais frequente em um documento e menos frequente no *corpus* maior será o valor da TF-IDF.
- Muito bom para tirar a importância de stop-words.
- Há variações da TF-IDF
 - Por exemplo para não zerar as palavras frequentes.
 - Variações em geral da expressão matemática.
 - Normalização.
 - Entre outras.

Medindo a semelhança entre sentenças

- Podemos calcular a semelhança com a TF-IDF.
 - Ela é igual a uma BOW, mas com os valores da TF-IDF.

		Vocabulário				
Documentos						
	TF-IDF(0,0)	TF-IDF(1,0)	TF-IDF(2,0)	TF-IDF(3,0)	...	TF-IDF(t,0)
	TF-IDF(0,1)	TF-IDF(1,1)	...	...	...	...
	TF-IDF(0,2)	...	...	...	...	...
	...	...	...	...	...	...
	...	...	...	...	...	...
	TF-IDF(0,d)	...	...	...	...	TF-IDF(t,d)

- Com isso calculamos a similaridade via cosseno.
 - Agora com um peso relativo a frequência das palavras.

Tarefa

- Pronto, agora podemos utilizar o notebook da aula 2 do NLP Demystified

OBRIGADO

James Moraes de Almeida
james.almeida@ilum.cnpem.br

ilum.cnpem.br



MINISTÉRIO DA
EDUCAÇÃO

MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO

