Introdução ao Processamento de Linguagem Natural (Natural Language Processing)

Tarefa de Leitura

Leia o artigo da Wikipédia sobre processamento de linguagem natural

NLP

Imagine que você trabalha para o Google Notícias e deseja agrupar artigos de notícias por tópico.

Ou você trabalha para um escritório de advocacia e precisa vasculhar milhares de páginas de documentos legais para encontrar os relevantes.

É aqui que a NLP pode ajudar!

NLP

Vamos querer:

- Compilar documentos
- Caracterizá-los
- Comparar suas características

NLP

Exemplo simples:

- Você possuí 2 documentos:
  - “Blue House”
  - “Red House”

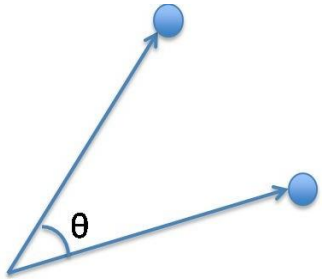
- Caracterize com base na contagem de palavras:
  - “Blue House” -> (red,blue,house) -> (0,1,1)
  - “Red House” -> (red,blue,house) -> (1,0,1)

NLP

- Um documento representado como um vetor de contagem de palavras é chamado de “Saco de Palavras (Bag of Words)”
  - “Blue House” -> (red,blue,house) -> (0,1,1)
  - “Red House” -> (red,blue,house) -> (1,0,1)

- Você pode usar a similaridade do cosseno nos vetores obtidos para determinar a similaridade:

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



NLP

- Podemos melhorar o Bag of Words ajustando a contagem de palavras com base em sua frequência no corpus (o grupo de todos os documentos)

- Podemos usar TF-IDF (Term Frequency - Inverse Document Frequência)

NLP

- Term Frequency - Importância do termo dentro desse documento
  - TF(d,t) = Número de ocorrências do termo t no documento d

- Inverse Document Frequency - Importância do termo no corpus
  - IDF(t) = log(D/t) onde
    - D = número total de documentos
    - t = número de documentos com o termo

NLP

- Matematicamente, TF-IDF é então expresso por:

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents