

Natural language processing - Procesamiento del lenguaje natural

NLP es un tipo de modelo de aprendizaje supervisado, el cual utiliza como entrada (inputs), texto inestructurado, realiza transformaciones matemáticas y las trabaja como vectores de palabras para obtener resultados a partir de ello.

Un ejemplo práctico de lo anterior puede ser el análisis de 2 documentos:

- a. "Red house"
- b. "Blue house"

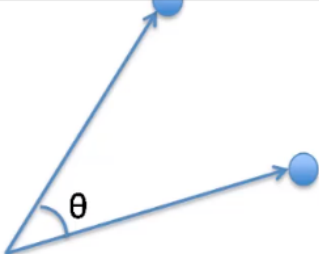
Ambos documentos se componen netamente de 2 palabras, pero para poder generar un modelo a partir de ellos, debemos empezar con lo llamado "bag of words", que es el conjunto vectorizado (a través de conteo de palabras, en donde se cuenta cuantas veces una palabra aparece en un documento):

"blue house" -> (red, blue, house) -> (0, 1, 1)

"red house" -> (red, blue, house) -> (1, 0, 1)

Un documento representado como un vector de conteo de palabras es llamado "Bag of words".

Luego puedo utilizar la similitud coseno en los vectores para determinar cierta similitud entre documentos:

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


Una manera de mejorar lo anterior es incorporar dentro del conteo de palabras basado en su frecuencia en el corpus (el grupo de todos los documentos), de esta manera podemos usar TF-IDF (Term frequency - Inverse document frequency)

Definiciones

Term frequency: Importancia del término dentro de un documento.

- $TF(d, t) = \text{Número de ocurrencias del término } t \text{ en el documento } d.$

Inverse Document Frequency: Importancia del término en el corpus.

- $IDF(t) = \log(D/t)$ donde:

- D = número total de documentos
- t = número total de documentos con el término

Matemáticamente podemos expresarlo como la siguiente ecuación:

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

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

Así además de tener una certeza de la frecuencia, también tendremos cierta notación de qué tan importante es esa palabra para todo el corpus.