1. NLP

El procesamiento de lenguaje natural (NLP - natural language processing) es un campo de las ciencias de la computación, de la inteligencia artificial y de la lingüística que estudia las interacciones entre las computadoras y el lenguaje humano.

1.1. Similitud coseno

La similitud coseno (Cosine similarity) es una medida de la similitud existente entre dos vectores en un espacio que posee un producto interior con el que se evalúa el valor del coseno del ángulo comprendido entre ellos. El valor de esta métrica se encuentra entre -1 y 1, es decir en el intervalo cerrado [-1,1].

Definición 1.1. Para vectores u y v, en un espacio euclídeo, distintos de cero en $\mathbb{R}^n$,

$$\cos(\theta) = \frac{u \cdot v}{\|u\|\|v\|} \tag{1}$$

- Dos vectores u y v en $\mathbb{R}^n$ son mutuamente ortogonales si y solo si $u \cdot v = 0$.
- Dos vectores u y v en $\mathbb{R}^n$ son paralelos si y solo si $u \cdot v = \|u\|\|v\|$.

1.2. Bag of Words (BOW)

El modelo bolsa de palabras (del inglés, Bag of Words) es un método que se utiliza en el procesado del lenguaje para representar documentos ignorando el orden de las palabras. Con este modelo podemos tener una representación de cada documento, en función de las palabras que este contiene.

1.3. Term Frequency (Count Vectorizer)