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 $u \cdot v = \|u\|\|v\|$.

Entonces un documento es similar a otro, si y solo si $\cos(\theta) \overset{tiende}{\longrightarrow} 1$.

1.2. Tokenización

La tokenización es el proceso de dividir una cadena de caracteres en unidades más pequeñas, llamadas tokens.

¿Por que tokenizar, y no usar simplente hacer un split? Porque la tokenización `nltk` es un proceso más complejo, mientrar que `split` solo separa de acuerdo a la key que se le pasa como parametro.

1.3. 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.

1.4. Term Frequency (Count Vectorizer)

Hace lo mismo que BOW, pero además cuenta la cantidad de veces que aparece cada palabra en cada documento.

1.5. Term Frequency x Inverse Document Frequency (TF-IDF)

Definición 1.2.

$$\text{TF-IDF}(palabra) = \log\left(\frac{N+1}{frecuencia}\right) \tag{2}$$

- N Cantidad total de documentos.
- $frecuencia$ Cantidad de documentos en los que aparece la palabra.

Entonces, si una palabra aparece en todos los documentos, su valor de TF-IDF será 0.

1.6. Normalización

1.6.1. Stemming

Es un proceso en donde se elimina la última parte de las palabras por algún proceso (hay steammers de varios tipos). El objetivo es llegar a la raíz (stem) de la palabra por medio de su prefijo.

- caballo, caballería, caballero → caball
- biblioteca, bibliotecario, bibliotecología → bibliotec
- canto, cantar, cantante → cant
- cantidad, cantar, cantante → cant

1.6.2. Lemmatization

Devuelve la base de la pala (lemma), muchas veces por medio de un **diccionario de lemas** para cada palabra.

- caballo, caballería, caballero → caballo
- biblioteca, bibliotecario → biblioteca
- canto, canta, cantamos, cantan → canto
- cantidad, cantidades → cantidad

1.6.3. Stopwords

Las stopwords o "palabras vacías"son palabras demaciado comunes o sin significado como artículos, pronombres, preposiciones, etc. Estas palabras no aportan información relevante para el análisis de texto, por lo que se suelen eliminar.

A veces (por ejemplo usando BOW o TF-IDF) queremos removerlas.