Práctica 5

Programen un algoritmo que mida la similitud semántica entre los documentos de una colección.

1. Matriz término-documento

Calculen la matriz término documento de un conjunto de documentos $D = \{d_1, d_2, \ldots, d_N\}$. Para ello:

1. Normalicen **todos** los documentos.
   - Tokenicen, lematicen y remuevan las palabras funcionales de cada texto $d_i \in D$, incluyendo puntuación.
     - Se generará un vocabulario $V = \{tipo_1, tipo_2, \ldots tipo_n\}$
     - Los tipos se expresarán en minúsculas
2. Programen una función que calcule el *score* tf-idf siguiendo la siguiente fórmula:

$$tfidf(t, d, D) = \frac{C(t, d)}{\sum_{t' \in d} C(t', d)} \cdot log\left(\frac{|D|}{|\{d' \in D \mid t \in d'\}|}\right)$$

   Donde:
   - t denota un término
   - d denota un documento
   - D denota un conjunto de documentos
   - $C(t, d)$ denota la frecuencia de t en d
3. Construyan la matriz usando la función programada, de modo que:

	$tipo_1$	$tipo_2$	$\ldots$	$tipo_n$
d_1	$tfidf(tipo_1, d_1, D)$	$tfidf(tipo_2, d_1, D)$	$\ldots$	$tfidf(tipo_n, d_1, D)$
d_2	$tfidf(tipo_1, d_2, D)$	$tfidf(tipo_2, d_2, D)$	$\ldots$	$tfidf(tipo_n, d_2, D)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
d_N	$tfidf(tipo_1, d_N, D)$	$tfidf(tipo_2, d_N, D)$	$\ldots$	$tfidf(tipo_n, d_N, D)$

En Python, la función seguirá el siguiente patrón:

```
def td_matrix(D):
        '''D es una lista de textos de diferente longitud'''
    normalized_D, vocabulary = normalize(D)
    vocabulary.sort()
    collection = []
```

```
for d in normalized_D:
    d_vector = []
    for t in vocabulary:
        d_vector.append(tfidf(t, d, normalized_D))
    collection.append(d_vector)
return collection, vocabulary
```

2. Similitud coseno

Programe una función que calcule el coseno del ángulo entre dos vectores A y B, dado por la siguiente fórmula:

$$\cos(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

2.1. Matriz de similitud

Sea v_i el vector correspondiente a la fila del documento d_i, tal y como aparece en la matriz término-documento. Utilice la función de coseno para construir la matriz de similitud como se muestra a continuación:

$$\begin{array}{cccccc} & d_1 & d_2 & \dots & d_N \\ d_1 & \cos(v_1, v_1) & \cos(v_1, v_2) & \dots & \cos(v_1, v_N) \\ d_2 & & \cos(v_2, v_2) & \dots & \cos(v_2, v_N) \\ \vdots & & & \ddots & \vdots \\ d_N & & & \dots & \cos(v_N, v_N) \end{array}$$

3. Ejercicio Definiciones

Cree la matriz de similitud para todas las definiciones que encontraron en el ejercicio que vimos en clase. Esto incluye tanto las definiciones de “aparato” cómo las de los términos que usaron en su vocabulario.

4. Repositorio

Guarden su código en GitHub. En el README del repositorio:

1. Impriman la matriz de similitud.
2. Pongan las definiciones que utilizaron.

- Cada definición debe tener un número asignado, para saber cómo buscarla en la matriz de similitud.