Coincidencia de documentos por su similitud calculada mediante la función coseno

Adrián Gallego Sánchez
4/2/2018
Sistemas inteligentes y representación del conocimiento

Indice

Lógica del programa

Preprocesado

Para poder realizar correctamente la similitud entre los documentos o el documento y la query debemos procesar el documento asi unicamente nos quedaremos con las palabras más significativas del texto. Esto se realiza mediante una lista de parada, la cual nos permite añadir palabras para su eliminación posterior en los documentos.

Count Vectorice

Una vez limpiado tanto los documentos y la query realizamos un Count Vectorice, Convierte una colección documentos en una matriz de tokens con las veces que este aparece en el documento

	0	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	1	0	0	0	1	0
1	1	0	0	0	0	0	0	1	0	0	1
2	0	1	0	1	0	0	0	0	1	0	0
3	0	0	1	0	1	1	1	0	0	0	0

Columnas serian los documentos/query y en las filas las palabras que aparecen en todos los documentos

*Importante la Query y los documentos tienen que tener el mismo número de filas con lo que el count vectorize de la query se realizará con las palabras de los documentos.

TF-IDF

Una vez que disponemos de una matriz con pensos de las palabras debemos realizar el tf idf para poder calcular los pesos que dichos términos tienen en el documento y en la lista de documentos.

	0	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0.486934	0	0	0	0.617614	0
1	0.57735	0	0	0	0	0	0	0.57735	0	0	0.57735
2	0	0.447214	0	0.447214	0	0	0	0	0.447214	0	0
3	0	0	0.525473	0	0.525473	0.414289	0.525473	0	0	0	0

Función Coseno

Una vez que disponemos de la matriz de pesos de los documentos como anteriormente he mostrado, utilizaremos la formula

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{n}(w_{ij} \times w_{iq})}{\sqrt{\sum_{i=1}^{n}(w_{ij})^2} \times \sqrt{\sum_{i=1}^{n}(w_{iq})^2}}$$

para el cálculo de la similitud entre la query y cada documento. El resultado será un numero entre 0.0 y 1. 0 seria coincidencia nula y el 1 serían documentos idénticos o muy parecidos.

Mostrar los resultados

Para mostrar los resultados he desarrollado una pequeña interfaz gráfica donde podremos visualizar una tabla con los documentos que disponíamos, y el grado de semejanza con la query

	Numero	Documento	Coincidencia
1	0	Hola mi nombre es Carlitos	0.356579823338
2	1	Coches motos autocares	0.0
3	2	Caballo pony vaca gallina cabra	0.516397779494
4	3	Chupoptero cabezon carlitos ca...	0.0

Manual de Usuario

Para interactuar con la interfaz gráfica únicamente debera introducir la query manualmente, clicar en el botón ‘Buscar’, automáticamente se verán los resultados en la tabla.

Para poder introducir más documentos en la carpeta del proyecto, encontrara otra carpeta con el nombre Documentos dicha carpeta contiene .txt, unicamente copiar más .txt. La aplicación no devolverá los resultados esperados si esos .txt contiene unicamente el carácter 0 o xxx0.txt