CAIM

(Laboratorio)

Práctica 2

Documentación

Curso 2021/2022 Q1

- Jesús Benítez Díaz
- Javier Rivera Hernández

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Índice:

1 Modifying ElasticSearch index behavior

En la primera parte de esta sesión estudiaremos cómo funcionan y cómo se comportan los flags *-token* y *-filter* a la hora de crear un índice. A continuación mostraremos los resultados obtenidos al probar los diferentes tokens.

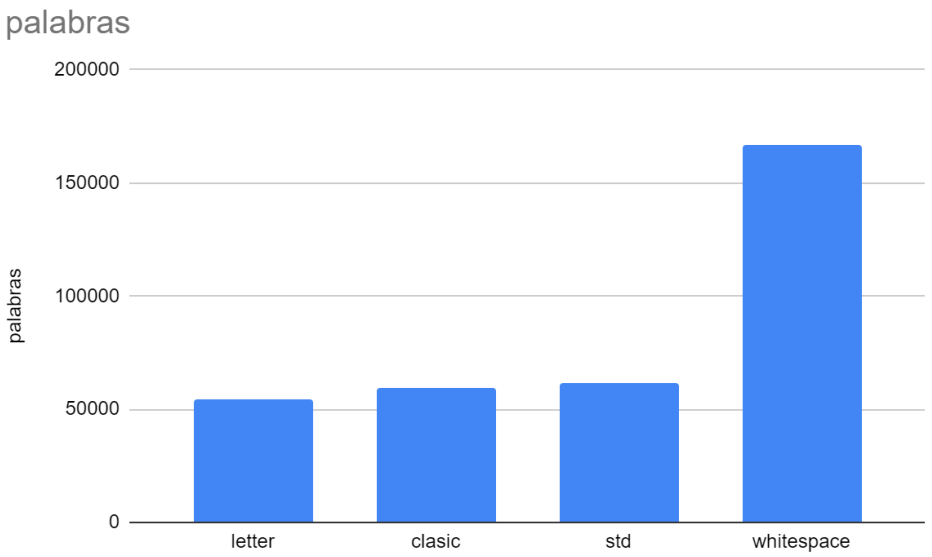

Figura 1

Podemos apreciar con claridad (Figura 1) que el token más agresivo es letter, siendo este el que elegiremos.

Por otro lado, seleccionamos los siguientes filtros: lowercase y asciifolding para librarnos de caracteres ASCII innecesarios y no utilizados en inglés.

Además como stemmer utilizaremos kstem, que es el menos agresivo de ellos, ya que no queremos que nos agrupe palabras que a priori no son del mismo tema (ej. university - universe -> univers). Cuando decimos que un stemmer es más agresivo nos referimos a que deja la raíz de la palabra lo más corta posible.

Aun así hemos probado los otros dos y hemos observado que snowball es más agresivo que porter y a su vez estos dos son más agresivos que kstem.

2 Computing tf-idf 's and cosine similarity

Una vez hemos experimentado con los diferentes filtros y tokens, nos tocará jugar con los diferentes documentos viendo que similitud hay entre ellos. Para ello debemos completar una serie de funciones del scirpt *TFIDFViewer.py*.

3. Experimentación

Nuestra experimentación se desarrollará en tres escenarios diferentes. Primero de todo, probaremos a comparar un texto consigo mismo, donde lógicamente esperamos un 100% de similitud.

Para nuestro segundo experimento, compararemos textos de un mismo tema, en este caso serán del espacio. De esta prueba esperamos conseguir una similitud considerable, ya que tratamos textos del mismo ámbito.

Para nuestro tercer y último experimento, buscaremos comparar textos de diferentes temas, trabajaremos con textos sobre hockey y textos religiosos. Nuestra hipótesis en este caso será que obtendremos peores porcentajes de similitud.

Para realizar estos experimentos, usaremos un script que comparara los 100 primeros textos entre ellos, y mostraremos la distribución de las probabilidades obtenidas.

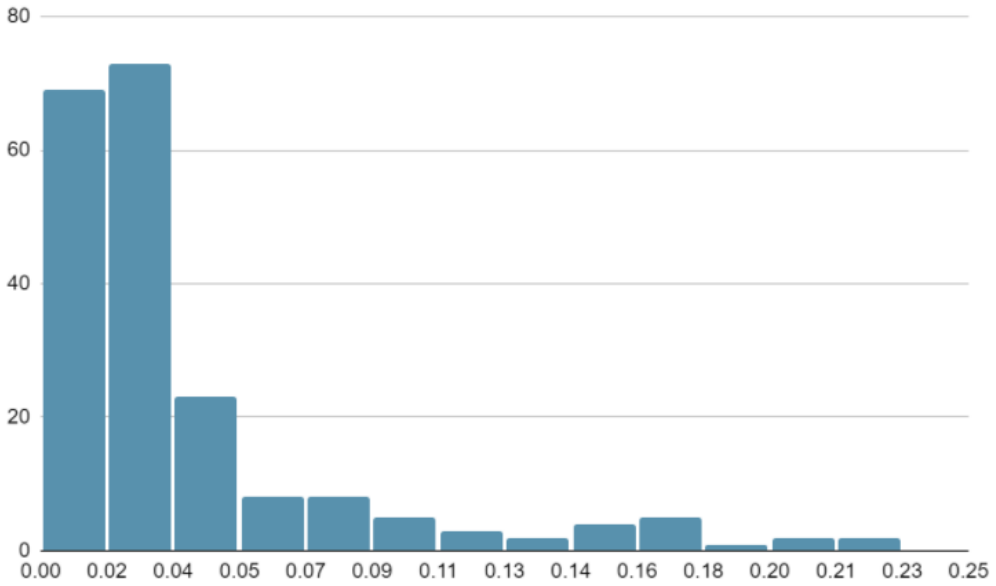

Figura 2

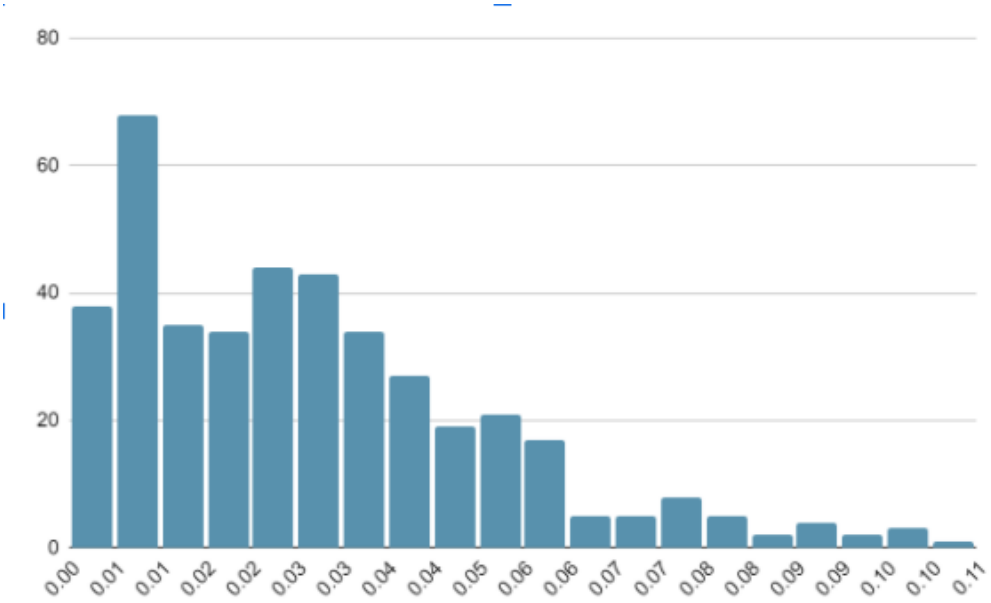

Figura 3

En la Figura 2 tenemos la comparación de textos del mismo tema. En la Figura 3 tenemos la comparación de los textos de diferentes temas.

4 Conclusión

Nuestra implementación de la similaridad del coseno parece funcionar, puesto que al comparar un texto consigo mismo da similaridad de 1. Aun así, obtenemos unos valores muy bajos por lo general, creemos que podría deberse a la elección del stemmer.

La hipótesis de que los textos científicos se parecen entre sí, se valida teniendo en cuenta las bajas similaridades que obtienen en general al compararse con el experimento de hockey vs cristianos. Observamos que a pesar de que hay bastantes textos que no se parecen entre sí, tenemos una cola en la distribución de bastantes textos que tienen similitudes superiores al 10%, cosa que no sucede en la distribución de hockey vs cristianos, donde todas las probabilidades se acumulan entre 0-6% y la cola solo llega al 11%.

Por lo tanto podemos concluir que ambas hipótesis se confirman. Aun así, en el caso de la primera, si no se obtienen valores de similitud tan bajos por la elección de stemmer o otro motivo de implementación, no se han obtenido unas similitudes tan altas como se podría esperar a priori.