Miquel Gotanegra Estañol
Oscar Daniel Sans Pallarés

INTRODUCCIÓ

En aquesta segona pràctica hem après a filtrar les paraules dels documents segons diferents criteris, per a així poder eliminar tots aquells mots que no ens interesenen o que tenen del mateix significat (mateixa stem word).

També hem après que comporta aquests canvis en les cerques finals i a observar les similituds de dos documents amb la mesura del cosinus.

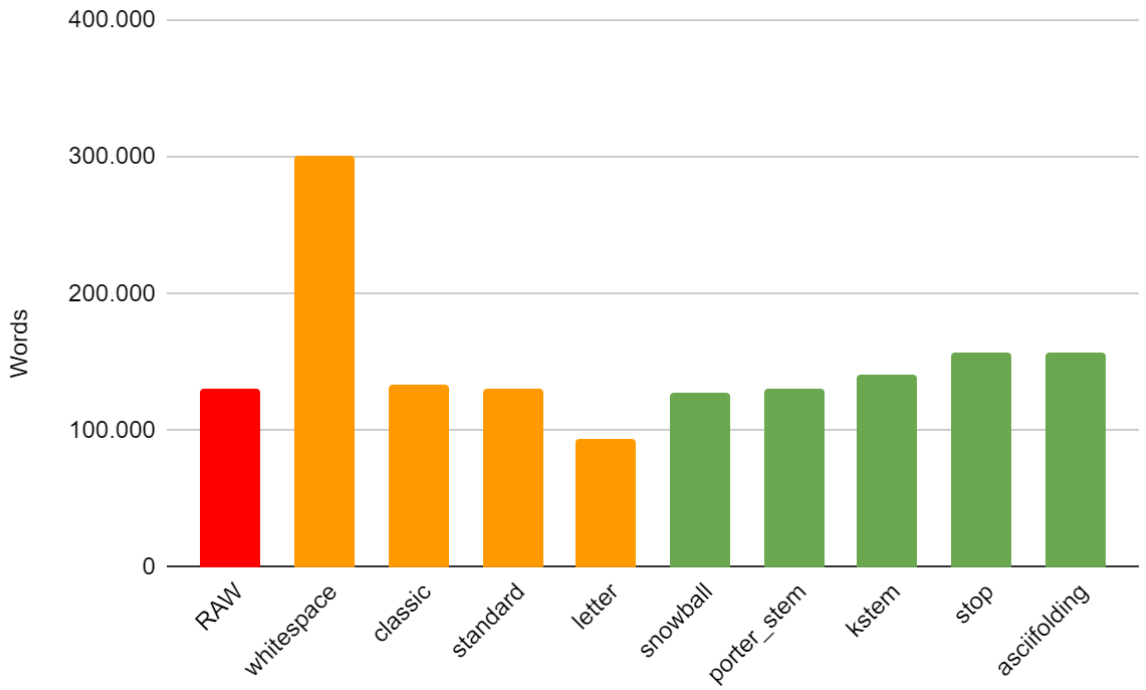
Finalment hem utilitzat tf-idf i similaritat per cosinus per determinar la semblança entre dos documents.

ALTERANT L'ÍNDEX DEL ELASTICSEARCH

Primer de tot hem volgut analitzar com afecten els diferents flags al script **IndexFilesPreprocess.py**, on tenim el **--token** que determina el que és una paraula durant el processament del text , entre els quals tenim **whitespace, classic, standard i letter**. Per exemple el whitespace determina que tot el que tingui un espai en blanc a davant i a darrera es una paraula, independentment dels caràcters que contingui.

També podem aplicar un (o varis) filtre de postprocessat amb la flag **--filtre**, entre els quals tenim **lowercase, asciifolding, stop** i els mecanismes d' steming **snowball, porter_stem** i **kstem**.
Una cosa de la qual ens hem adonat es que l'ordre en el que s'apliquen els filtres afecta el resultat (entre els filtres normals i els de steming), ja que una paraula que pot ser eliminada pel filtre també se'n pot ajuntar a una altra pel significat.

	TOKEN				FILTER				
RAW	whitespace	classic	standard	letter	snowball	porter_stem	kstem	stop	asciifolding
130.290	300.604	132.956	130.290	94.003	127.280	130.413	140.593	156.809	156.831



A les gràfiques, fetes amb la totalitat de la carpeta **20_newgroups,** hi podem veure el reconte de les paraules total a l'índex amb els diferents filtres i tokens; a destacar el token **whitespace** per ser el mes permissiu, cosa que ja ens esperàvem perquè és l'únic que accepta lletres soltes, números i i combinacions rares de caràcters.

També hem trobat que la paraula ‘the’ és la més utilitzada en el vocabulari anglès, el qual no ens ha sorprès ja que és l'únic article que es fa servir en anglès i per tant serà molt més freqüent del que ho seria un “el” o “la” en castellà. També cal esmentar que quan utilitzem el flag stop aquesta paraula ja no sortirà ja que era filtrada.

TF-IDF I EL COSINUS DE SIMILITUD

Primer de tot, se'ns demana completar unes funcions que estan incompletes en el fitxer TFIDPViewer.py.

search_file_by_path: Busca el fitxer utilitzan el seu path.
document_term_vector: Retorna el terme del vector del document i les seves estadistiques.
toTFIDF: Retorna el pes dels termes del documents.
normalize: Normalitza els pesos.
print_term_weigth_vector: Imprimeix els termes dels vectors i els seus pesos.
cosine_similarity: Calcula la similitud entre dos vectors.

Per completar aquestes funcions hem fet servir bàsicament la llibreria numpy que incorpora funcions per a vectors com el normalize.
 Per fer la la **cosine_similarity** en temps polinòmic hem recorregut els dos vectors d'un en un comparant els elements i avançant la posició del que tingués la paraula més petita en cas de que no coincidicin (cost 2n).

COMPARACIÓ AMB EL MATEIX DOCUMENT

El primer que hem fet ha sigut comparar un fitxer amb ell mateix per a veure quin resultat ens donava cosinus. Com era d'esperar el resultat de similitud es igual a 1 ja que eren els mateixos documents.

```
miquel@miquel-VirtualBox:~/Desktop/CAIM/session2ESprogramming$ python3 TFIDFViewer.py --index news --files /tmp/2
0_newsgroups/alt.atheism/0000001 /tmp/20_newsgroups/alt.atheism/0000001
/home/miquel/.local/lib/python3.8/site-packages/elasticsearch/connection/base.py:209: ElasticsearchWarning: Elast
icsearch built-in security features are not enabled. Without authentication, your cluster could be accessible to
anyone. See https://www.elastic.co/guide/en/elasticsearch/reference/7.13/security-minimal-setup.html to enable se
curity.
  warnings.warn(message, category=ElasticsearchWarning)
Similarity = 1.00000
```

COMPARACIÓ AMB UN DOCUMENT DIFERENT

Quan ens hem possat a comparar el document **alt.atheism/0000001** amb altres documents semblants (mateixa carpeta) ens hem adonat que, tot hi a tractar el mateix tema i per tant fer servir paraules semblants, la similaritat entre documents era d'al voltant del 10%. A més dins al mateix tema hi ha documents amb els que és relativament més semblant (5 => 22%) i d'altres amb els que té una similaritat molt baixa (4 => >1%).

	ateismo 1	ateismo 2	ateismo 3	ateismo 4	ateismo 5
ateismo 1	1,0000	0,1104	0,1008	0,0097	0,2201

També hem comparat el mateix document amb documents d'una altre temàtica i el resultat obtingut ha estat el següent:

	sci_space 14001	sci_space 14002	sci_space 14003	sci_space 14004	sci_space 14005
ateismo 1	0,0183	0,0085	0,0154	0,0402	0,0512

Com podem observar, quan comparem una tematica amb una altra els resultats de les similituds del cosinus son molt més baixes.

CONCLUSIONS

Les conclusions que hem pogut treure del primer apartat han sigut les que es esperàvem, el filtre de whitespace és molt més permissiu que tots els altres ja que no té en conte si el la "word" és realment una paraula o no, mentre que tots els altres ho tenen en compte d'una manera o altre.

Per altre part, el que sí que ens ha sorprès a l'hora del segon experiment es que els documents de temàtiques similars no tienen una similitud de cosinus molt elevada (~10%), cosa que no ens esperàvem. Suposem que es deu a que, tot-hi a tenir termes comuns, cada article ha estat escrit per una persona diferent i que la similaritat seria més alta entre dos articles de la mateixa temàtica i escrits per mateix autor.