Lab2: Programming with Elastic Search

Q1 2022-23

Integrants:
Pablo Montón Gimeno
Cristian Sánchez Estapé

Introducció

Primera part

A la primera part de la pràctica, amb l'*script* per indexar i preprocessar els documents, hem experimentat i tractat el resultat d'aplicar els diferents tokenitzadors i filtres als documents indexats. En aquest sentit, s'han fet proves amb el resultat d'aplicar uns certs tokens o filtres a un determinat text, de la mateixa manera que també es combina l'aplicació de diferents filtres entre ells i l'ordre en el qual s'apliquen.

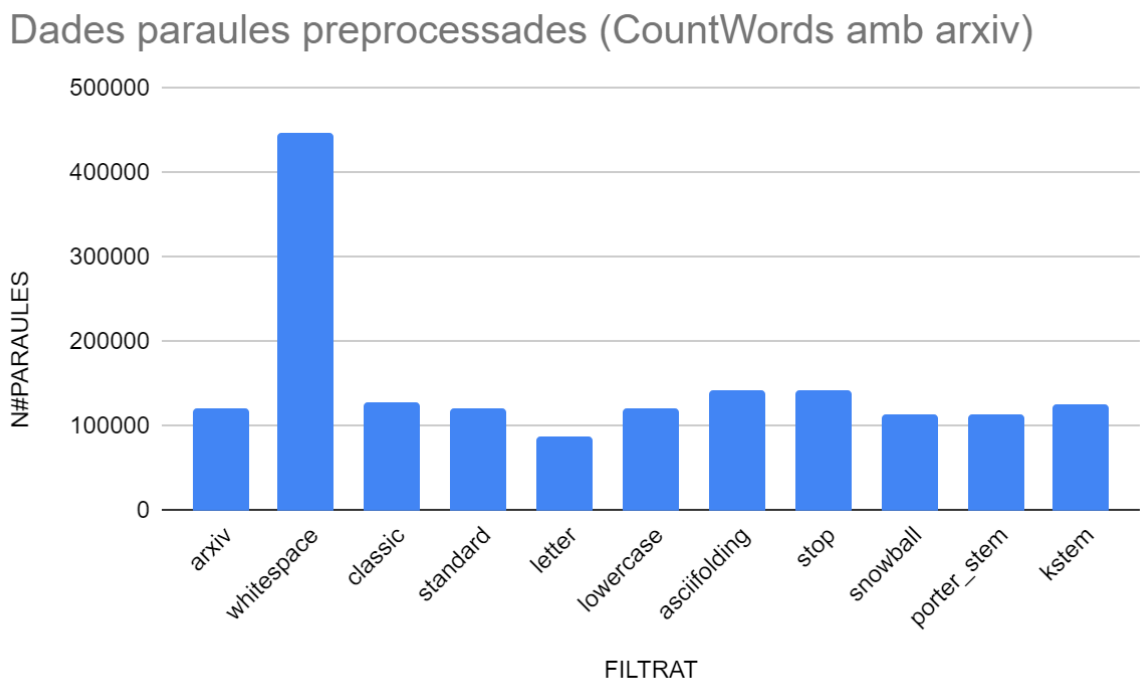

Figura 1: taula representativa del nombre de paraules segons l'opció de filtrat

La figura 1 representa la utilització de l'script sense especificar cap token o filtre, fet del qual podem deduir que, de manera gràfica, s'utilitza el token standard i el filtre lowercase.

A banda d'aquests fets, fent ús de l'script *mostUsedWord* a l'índex novels, pels 4 tipus de *tokenizers* possibles i amb totes les opcions de filtratge existents, les paraules més emprades en anglès són (*The*, *of*, *and*)[1] encara que, pels casos en què s'han filtrat les *stopwords*, canvien pels termes (*I*, *he*, *his*).

Pel que fa a l'índex *arxivs_abs*, el mateix fenomen es pot observar: els filtrats amb *asciifolding stop* i *stop* mostren com a resultats els termes (*we*, *We*, *The*), mentre que el filtratge *kstem* mostra els termes (*The*, *of*, *and*). Per tant, d'acord amb aquests resultats es pot afirmar clarament que els termes més usats en anglès, almenys pels fitxers aquí estudiats, resulten ser articles i conjuncions, tot i que, en absència d'aquests, prenen rellevància els pronoms personals.

[1] L'article *to* en alguns casos li disputa l'importància a l'*and*.

Segona part

Per aquest apartat, se'ns ha proporcionat un script que, tot separant en funcions les diverses operacions involucrades tant en la cerca i recuperació dels fitxers indexats a ElasticSearch com en el càlcul del *TF-IDF* i la *Cosine Similarity*, permet aplicar aquests procediments per tal de comparar la similitud de dos documents (o conjunts dels mateixos) d'acord amb els termes que s'hi troben presents.

En aquest sentit, el codi proveït és bastant clar: la modularitat de les funcions permet diferenciar exactament quines operacions hi intervenen tot i la relació existent entre aquestes. És per això que, per exemple, podem identificar perfectament el càlcul del *TF-IDF* (que pertany a la funció *toTFIDF*), la cerca del(s) document(s) a tractar (*search_file_by_path*), la normalització del vector de pesos associat a un document (*normalize*) o el càlcul de la simili tud cosinus (*cosine_similarity*), entre d'altres.

És degut a aquest fet precisament que la implementació requerida per les funcions de *toTFIDF*, *normalize*, *cosine_similarity* i *print_term_weight_vector* ha estat relativament senzilla, posat que els computs que s'han d'implementar no resulten complexos. Tot i això, val la pena esmentar una dificultat que acompanyava al format del codi: posat que el vector de pesos es crea a la funció *toTFIDF*, i donat que l'estructura de dades en què (originalment) s'emmagatzemava aquesta informació era un vector, no hi havia manera de conservar informació respecte als termes presents a cadascun dels documents. Val afegir que, tot i que aquests pesos arriben ordenats a la funció *cosine_similarity* segons l'ordre (alfabètic) dels termes presents al document, si només s'emmagatzemen aquests valors numèrics, no hi ha manera de saber a quins termes es troben associats i, per tant l'aplicació de la similitud cosinus seria generalment errònia. Per tal de resoldre aquest problema, hem optat per transformar aquest vector en un diccionari: el cost de consultar un dels seus elements idealment es mantindrà a $O(1)$ i ens permet emmagatzemar el valor lligat al seu terme. Això, posteriorment, permet aplicar la similitud cosinus correctament.

I sabem que és correcta precisament perquè, en aplicar el nostre script en un document sobre si mateix, veiem que el resultat de la similitud cosinus és 1. Sabent que aquest fet se sosté per tots els documents, hem muntat una matriu de la similitud de cada document tant amb si mateix com amb els restants (per un mateix índex), una informació amb la qual hem manufacturat les següents taules:

	Mitjana	**Max**	**Min**	**Fitxer associat al max**	**Fitxers associat al min**
comp.windows.x	0,01181	0,03616	0,00371	comp.os.ms-windows.misc	alt.atheism
talk.politics.misc	0,01458684211	0,06109	0,00011	rec.autos	rec.sport.baseball
sci.electronics	0,02237894737	0,04798	0,00493	comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware
misc.forsale	0,009583684211	0,04531	0,00092	sci.electronics	comp.graphics
soc.religion.christian	0,01736684211	0,03949	0,00134	sci.med	rec.sport.baseball
comp.sys.mac.hardware	0,01664210526	0,04465	0,00421	rec.autos	rec.sport.baseball
rec.sport.baseball	0,003182631579	0,01756	0,00011	comp.windows.x	talk.politics.misc
rec.sport.hockey	0,01359789474	0,02736	0,00146	rec.motorcycles	rec.sport.baseball
talk.politics.mideast	0,01352684211	0,03786	0,00082	soc.religion.christian	rec.sport.baseball
sci.med	0,02261894737	0,05222	0,00136	talk.politics.guns	rec.sport.baseball

talk.religion.misc	0,01610526316	0,03377	0,00016	soc.religion.christian	rec.sport.baseball
alt.atheism	0,01683894737	0,05688	0,00155	rec.autos	rec.sport.baseball
sci.crypt	0,01059105263	0,02659	0,00017	rec.autos	rec.sport.baseball
rec.autos	0,03556842105	0,10211	0,00329	talk.politics.guns	rec.sport.baseball
comp.sys.ibm.pc.hardware	0,01041684211	0,02246	0,00261	rec.motorcycles	soc.religion.christian
rec.motorcycles	0,02748421053	0,07749	0,00112	rec.autos	rec.sport.baseball
comp.graphics	0,008474736842	0,01809	0,00092	comp.sys.mac.hardware	misc.forsale
sci.space	0,01526473684	0,04216	0,00018	talk.politics.guns	rec.sport.baseball
comp.os.ms-windows.misc	0,01637789474	0,04798	0,00021	sci.electronics	rec.sport.baseball
talk.politics.guns	0,03199	0,10211	0,00191	rec.autos	rec.sport.baseball

*Figura 2: taula associada a l'índex **news***

	Mitjana	**Max**	**Min**	**Fitxer del max**	**Fitxers del min**
cs.updates.on.arXiv.org	0,01085714286	0,01995	0,00101	math.updates.on.arXiv.org	cond-mat.updates.on.arXiv.org
quant-ph.updates.on.arXiv.org	0,009274285714	0,01475	0,00517	astro-ph.updates.on.arXiv.org	hep-th.updates.on.arXiv.org
cond-mat.updates.on.arXiv.org	0,006387142857	0,01514	0,00101	physics.updates.on.arXiv.org	cs.updates.on.arXiv.org
hep-th.updates.on.arXiv.org	0,006745714286	0,01371	0,0017	hep-ph.updates.on.arXiv.org	cond-mat.updates.on.arXiv.org
physics.updates.on.arXiv.org	0,01718285714	0,03177	0,00487	astro-ph.updates.on.arXiv.org	hep-th.updates.on.arXiv.org
hep-ph.updates.on.arXiv.org	0,01330285714	0,02899	0,00463	physics.updates.on.arXiv.org	cs.updates.on.arXiv.org
math.updates.on.arXiv.org	0,01055428571	0,01995	0,00412	cs.updates.on.arXiv.org	cond-mat.updates.on.arXiv.org
astro-ph.updates.on.arXiv.org	0,01652428571	0,03177	0,00748	physics.updates.on.arXiv.org	cond-mat.updates.on.arXiv.org

*Figura 3: taula associada a l'índex **arxivs***

Amb aquestes dues taules, diverses coses es poden apreciar: en primer lloc, la relació que existeix entre els documents d'un mateix índex és un fenomen que pot veure's manifestat a partir de la mitjana. D'aquesta manera, podem observar que, pel cas de l'índex **news**, el document *rec.autos* resultaria ser el que més relació té amb els altres (la qual cosa podria voler dir que inclou un vocabulari prou variat per a poder-se relacionar amb la gran majoria de documents restants), mentre que el document *rec.sport.basketball* és el que menys relació té amb els altres, fenomen que també es pot apreciar a la columna de *fitxers associats al min* (des de la qual no tan sols es reafirma, sinó que posa en relleu diverses implicacions: potser el llenguatge d'aquests documents és molt tècnic, o

potser es fa ús d'un llenguatge barroer o associat a un determinat dialecte de l'anglès). Pel que fa a l'índex **arxivs**, no aplica el mateix que en el cas previ (o no del tot almenys): es pot assumir que el llenguatge tècnic present en aquests documents és el que dona lloc a unes mitjanes tan reduïdes i que, per tant, fa que la diferència entre la mitjana màxima i mínima no sigui tan gran com a l'índex anterior (0,0108 envers 0,0235).

En segon lloc, tenim les columnes de màxims i mínims (i documents associats a aquests), que ens diuen, per cada document, quin li resulta més (i menys) semblant. Altre cop, pel que fa a l'índex **news**, tenim algunes relacions interessants, tals com la que existeix entre *misc.forsale* i *sci.electronics* o *rec.motorcycles* i *rec.autos*, que posen de manifest relacions més o menys esperables (podria ser que els objectes que es venen son predominantment electrònics, o que tot usuari familiar amb les motocicletes també ho fos amb els automòbils); i també hi apareixen relacions curioses, com *soc.religion.christian* i *sci.med* o *talk.politics.mideast* i *soc.religion.christian* (podria ser que la salvació cristiana fos una qüestió molt discutida, i que pogués tenir alguna vinculació més eminent amb la ciència mèdica del que hom podria pensar, de la mateixa manera que semblaria ser que la política a la regió *mideast* dels EUA es trobés altament influenciada per la religió). De la mateixa manera, semblaria ser que l'hockey no té gaire a veure amb el beisbol, de la mateixa manera el hardware dels ordinadors d'IBM no té relació amb el cristianisme.

Ara, pivotant a l'índex **arxivs**, també podem constatar relacions igualment interessants i previsibles: les matemàtiques tenen més relació amb les ciències de la computació que no pas amb la física, alhora que física i l'astrofísica tenen una estreta relació. Val a dir que, en aquests casos, es produeix la casualitat que la relació d'importància és recíproca: per les ciències de la computació, les matemàtiques són les més importants i viceversa. Nogensmenys, també podem veure que (assumint que cond-mat correspon al camp de la física de *condensed matter*), la física quàntica sembla tenir més a veure amb l'astrofísica, mentre que el camp d'estudi de la matèria condensada depèn més de la física com a camp general de la ciència. Ara bé, pel que fa als valors mínims, no es veuen relacions eminentment rellevants, tot i que si ens diuen coses: per exemple, la ciència de la computació no és gaire valuosa pels camps de la *condensed matter* (cond-mat) o *high energy physics* (*hep-ph*).

Finalment, la similitud cosinus ens permet derivar relacions indirectes entre els dos índexs: si hom aplica el script proveït sobre els documents *sci.space* de **news** i *astro-ph* d'**arxivs**, es veu que la relació existent entre ambdues és d'un valor de 0,00381. Tot i que s'entén la magnitud del resultat obtingut, ens diu que, indirectament, tindrà una relació (prou interessant) amb *talk.politics.guns*. Els motius exactes del perquè d'aquestes relacions concretes escapa l'abast d'aquest entregable, però el que sí ens permet afirmar tot aquest entramat és que la similitud cosinus i, per tant, el tf-idf resulten ser eines atractives d'anàlisi i permeten posar en relleu les semblances existents tant entre camps de coneixement pel que fa al llenguatge com en quant a de tòpics i temes tractats.