Alexandru-Ilie Popa
Pol Casacuberta Gil

CAIM Lab, Session 3: User Relevance Feedback

Document relevance

Hem executat els queries que se'ns donaven a l'enunciat de la pràctica i hem vist com variava la puntuació per a cadascuna, segons fem servir l'operador ^n. Els resultats han estat els següents:

```
['toronto', 'nyc']
ID= dathqIMBpJpGnd_nd-U4 SCORE=4.9770656
PATH= 20_newsgroups/sci.med/0013128
TEXT: Here is a press release from the Natural Resources
------------------------------------------------------------
ID= EKthqIMBpJpGnd_nWKXc SCORE=4.19135
PATH= 20_newsgroups/talk.politics.misc/0018667
TEXT: v140pxgt@ubvmsb.cc.buffalo.edu (Daniel B Case) wri
------------------------------------------------------------
ID= kqthqIMBpJpGnd_nYrMq SCORE=3.6505594
PATH= 20_newsgroups/talk.politics.guns/0015998
TEXT: Jin De Arras (jmd@cube.handheld.com) wrote:
: > La
------------------------------------------------------------
3 Documents
```

Query: toronto nyc

```
['toronto^2', 'nyc']
ID= EKthqIMBpJpGnd_nWKXc SCORE=7.06285
PATH= 20_newsgroups/talk.politics.misc/0018667
TEXT: v140pxgt@ubvmsb.cc.buffalo.edu (Daniel B Case) wri
------------------------------------------------------------
ID= dathqIMBpJpGnd_nd-U4 SCORE=7.0584717
PATH= 20_newsgroups/sci.med/0013128
TEXT: Here is a press release from the Natural Resources
------------------------------------------------------------
ID= kqthqIMBpJpGnd_nYrMq SCORE=5.1772213
PATH= 20_newsgroups/talk.politics.guns/0015998
TEXT: Jin De Arras (jmd@cube.handheld.com) wrote:
: > La
------------------------------------------------------------
3 Documents
```

Query: toronto^2 nyc

```
['toronto', 'nyc^2']
ID= dathqIMBpJpGnd_nd-U4 SCORE=7.872725
PATH= 20_newsgroups/sci.med/0013128
TEXT: Here is a press release from the Natural Resources
------------------------------------------------------------
ID= kqthqIMBpJpGnd_nYrMq SCORE=5.774457
PATH= 20_newsgroups/talk.politics.guns/0015998
TEXT: Jin De Arras (jmd@cube.handheld.com) wrote:
: > La
------------------------------------------------------------
ID= EKthqIMBpJpGnd_nWKXc SCORE=5.5112
PATH= 20_newsgroups/talk.politics.misc/0018667
TEXT: v140pxgt@ubvmsb.cc.buffalo.edu (Daniel B Case) wri
------------------------------------------------------------
3 Documents
```

Query: toronto nyc^2

We will, we will Rocchio you

Per implementar l'algorisme de Rocchio hem seguit el guió de la pràctica. En general tot ha anat bé i no hem tingut cap dificultat en particular per acabar aquest apartat.
Per entendre millor com funciona la regla de Rocchio hem fet preguntes al professor del laboratori i també hem buscat informació a internet, com per exemple hem mirat aquest vídeo de youtube: https://www.youtube.com/watch?v=yPd3vHCG7N4.

Experimentation

Fem servir diverses querys i anem modificant els paràmetres α, β, k, R i nrounds per veure com com van evolucionant els resultats de cada query.

Experiment 1

α = 1.0, β = 0.7, k = 5, R = 5, nrounds = 5

Query: computer windows
New query:['gillard^14.765333869979344', 'boot^9.84675583726732', 'window^9.035262625609569', 'dos^8.684170124027348', 'comput^6.425586930879628']
2 Documents

Query: computer^2 windows
New query:['mb^8.55650649615901', 'comput^6.875066700203778', 'window^5.389664481139722', 'computer^2.0', 'windows^1.0']
64 Documents

Query: computer windows^2
New query:['sharma^15.633882921154601', 'window^11.993433580390004', 'deptt^8.004706152930762', 'miniva^7.444706152930761', 'windows^2.0']
1 Documents

Experiment 2

Hem decidit disminuir la R, ja que era massa estricte i trobàvem molt pocs resultats
β = 0.7, k = 5, R = 3, nrounds = 5
Query: computer windows
Variem alfa:
α = 0.2 : 1 Documents found; α = 0.5 2 Documents found; α = 0.6 91 Documents found
α = 0.9 91 Documents found; α = 1.0: 293 Documents found; α = 2.0: 293 Documents found

Experiment 3

Del experiment 2 agafem α = 0.6, ja que hem vist que té en compte la nostre query inicial peró també fa servir els nous termes que ens aporta Rocchio.
Query: computer windows
α = 0.6, k = 5, R = 3, nrounds = 5
Variem beta:

β = 0.2 : 1 Documents found ; β = 0.5 2 Documents found ; β = 0.6 91 Documents found
β = 0.9 91 Documents found; β = 1.0: 91 Documents found; β = 2.0: 2 Documents found
β = 100.0: 2 Documents found

Experiment 4

Query: computer windows
α = 0.6, β = 0.6, R = 3, nrounds = 5
Variem K:
K = 2: 2 Documents found
K = 5: 91 Documents found
K = 10: 131 Documents found
K = 20: 131 Documents found
K = 40: 131 Documents found

Experiment 5

Query: computer windows
α = 0.6, β = 0.6, k = 10, nrounds = 5
Variem R:
R = 2: 293 Documents found; R = 5: 2 Documents found; R = 10: 1 Documents found
R = 20: 0 Documents found; R = 40: 0 Documents found

Experiment 6

Query: computer windows
α = 0.6, β = 0.6, k = 10, R = 3
Variem nrounds:
nrounds = 2 293 Documents found ; nrounds = 5 131 Documents found
nrounds = 10 131 Documents found; nrounds = 20 131 Documents found
nrounds = 40 131 Documents found; nrounds = 100 131 Documents found

Experiment 7

Query: computer windows amb la base de dades arxiv_abs
α = 0.6, β = 0.6, k = 10, R = 3, nrounds = 10
New query:['window^0.020160534962734147', 'brownian^0.009139449895355683', 'stft^0.007012273277936855']
1 Document found

Experiment 8

Query: computer windows amb la base de dades novels
α = 1.0, β = 0.7, k = 5, R = 2, nrounds = 5
Usar database novels
New query:['scroog^5.616316296527309', 'cratchit^0.8934796071625117', 'marley^0.576820196053']
0 Documents found

Conclusion

Variant la variable alfa, veiem que si augmentem el valor de la variable ens augmenta els documents trobats, això és degut al fet que sabem que les paraules escollides es troben en bastants documents i si el valor d'alfa és baix, la query que hem realitzat no es té tant en compte i s'utilitzen paraules que ha trobat l'algoritme de Rocchio, en comptes de les que li donem nosaltres.

A mesura que variem la variable beta, ens donem conta d'una particularitat i és que augmentar beta té efectes similars a reduir alfa, cosa que té sentit coneixent com és l'equació de Rocchio.

Augmentant els valors de K el que fem és tenir en compte més documents a l'hora d'escollir els termes per la nova query això ens porta a escollir termes de query que segurament són compartides per més documents i per això les nostres querys troben més documents a mesura que augmentem K.

Hem observat que augmentant el valor de R, tot i que la precisió del que estem buscant, millora, també reduïm els documents que es troben és a dir disminuïm el recall.

Emprant valors baixos de NRounds podem veure que la query es podria refinar una mica més amb valors més alts de la variable i que a partir d'un cert punt afegir més rounds no té cap efecte en la query.

En els experiments 7 i 8 podem veure que pel fet que el caràcter dels documents és diferent. El de l'experiment 7 és científic i el 8 literari, no troba gaires o cap document que contingui els termes de la query en aquests índexs.