Lab 3

User Relevance Feedback

Bartomeu Perelló Comas

5/11/2020

CAIM - GEI FIB

Introduction

In this session we have to implement Rochhio's rule so we can experiment with different parameters and understand how we can enrich user's queries.

In our case, we just followed the scheme that was presented in the task, the most difficult part we faced was how to operate with Python's dictionaries.

Experimenting

For experimenting we have chosen the 20newsgruop in order to do the experiments, our goal is to see how the different parameters affect the retrieval of information.

Experimenting with 20newsgruop

Experiment 1

The elements of the query we have used for se the relationship between variables are the following:

query = science human, a*lpha* = 2, *beta* = 1, *R* = 4, ***k* = 10** and *nrounds* = 10

The result were 19 documents being the last query the following one:

['science^0.6116516881559172','value^0.568463399693776',
'human^0.41812810362222874', 'values^0.3576317443207319']

The k top documents in this case it's a mix between the *talk.religion.mix* and *alt.atheism* also the scores are located from 86 to 40 with a mean score of 68

Now we try with ***k* = 8**, now we retrieve 12 Documents and the texts come from the same mix as before but the score now is comprehended between 87 and 57 with a mean score of 75.

Last query:
['value^0.5942532622416337','science^0.5930819197441739',
'values^0.3852418772695071', 'human^0.3830216609874795']

Up next we tried the same query but changing the amount of documents we consider relevant to 6 so now ***k* = 6**.

Now we have only obtained 7 documents and the scores are way lower compared to the previous experiment, the top document has a score of 81 but the other 6 have a score between 40 and 15 with a mean score of 36
The last query is the next one:

['science^0.6636448692275375','human^0.6334016037112427',
'objective^0.3631468163583422', 'nature^0.1627952263820751']

Experiment 2

Now we switched the alpha and beta parameters so *alpha* = 1 and *beta* = 2, all other variables are the same as the first experiment *R* = 4, *k* = 10 and *nrounds* = 10.

Now we are giving new parameters higher weight than the original ones, so the final query and the number of documents could change.

['value^0.703126521987238', 'values^0.4533984560315351',
'truth^0.38758868638333577', 'science^0.3870632304081472']

As we can see in the query above as we give more relevance to new terms the human term has been deleted from the query, also because the content of it it's different we have obtained 13 documents.

Experiment 3

Now we will be trying different values for R, the number of parameters we want to take into account on the following queries, the following parameters aren't changing throughout the experiment.

alpha=2, *beta*=1, *k* = 10 and *nrounds* = 10
R = 2

Final query:
['science^0.8552752824073461', 'human^0.5181739006386122']

Total documents: 106

Mean score: 64.97

R = 4

Final query:
['science^0.6116516881559172', 'value^0.568463399693776', 'human^0.41812810362222874', 'values^0.3576317443207319']
Total documents: 19
Mean score: 68.14

In this case it seems that there’s not such a difference but the deviation between the scores is higher because previous almost all scores were between 68 and 60 and now are comprehended between 80 and 40

R = 6

Final query:
['science^0.5698132885467524', 'human^0.4208313267232695', 'des^0.4134890158347772', 'value^0.4091503246053865', 'truth^0.28478472827577667', 'values^0.28059635572126523']

Total documents: 5
Mean score: 95.62

R = 7

Final query:
['science^0.7267007812510945', 'human^0.6841189025282604', 'values^0.030651620219933614', 'value^0.028222776145738874', 'objective^0.027355003655341426', 'truth^0.026725352147794264', 'origins^0.026243363821939727']

In this case we have used too many parameters so the system cannot find any document with the whole elements in the query.

Experiment 4

Finally we will have a look at the effect of the *nrounds* variable.

nrounds = 5

Final query:
['science^0.7078784899413202','human^0.5744897954632076','value^0.34498948423933556', 'values^0.22327510866936187']

Number of documents: 19
Mean score: 30

nrounds = 10

Last query:
['science^0.6116516881559172', 'value^0.568463399693776',
'human^0.41812810362222874', 'values^0.3576317443207319']

Number of documents: 19
Mean score: 68

nrounds = 15

Last query:
['value^0.6718229170803434','science^0.527448638283895','values^0.4190382032754849', 'human^0.3079916983520436']

Number of documents: 19
Mean score: 110,4

Conclusions

Through experimentation we learned that if we assign higher values to beta, original terms from the query are more likely to be removed from it, otherwise, higher values on alpha will make them more relevant leaving the other ones with a lower weight overall.

If we take a look at the *k* experiment results we can appreciate that lowering it or raising it too much may cause the system not be able to retrieve any document, so probably we could hit the sourspot by making some tries with different values on it.

An interesting result has been the differences on the change of the *R* variable, it seems as the R increases the recall gets lower while the precision/score increases, but if we end up with a higher enough value we cannot find any document.

Finally the nrounds seems to affect the weight of the terms, so by more iterations we do we increase the precision of the search without losing recall, probably with enough iteration the weights stabilize or they get changed too much and we start losing precision or recall.