

Συστήματα Ανάκτησης Πληροφοριών

Φάση 3 – Πιθανοτικό και Γλωσσικό Μοντέλο Ανάκτησης

Στυλιανή Δούκα – p3170042

Βασίλειος Μπάλλας – p170115

Περιγραφή Υλοποίησης

Για τις ανάγκες αυτής της Φάσης της εργασίας αλλάξαμε τον τύπο της κλάσης Ομοιότητας πρώτα σε BM25Similarity και έπειτα σε LMJelinekMercerSimilarity.

```
// Access the index using indexReaderFSDirectory.open(Paths.get(index))
IndexReader indexReader = DirectoryReader.open(FSDirectory.open(Paths.get(index)));
IndexSearcher indexSearcher = new IndexSearcher(indexReader);
indexSearcher.setSimilarity(new BM25Similarity());
indexSearcher.setSimilarity(new LMJelinekMercerSimilarity(0.7f));
```

Στη συνέχεια περάσαμε τα αποτελέσματα από το Trec_eval:

Εκτελούμε trec_eval.exe -m all_trec CISI.REL resultsCISI_BM25_20.txt στα 20 πιο σχετικά κείμενα

```
Microsoft Windows [Version 10.0.19042.985]
(c) Microsoft Corporation. Με επιφύλαξη κάθε νόμιμου δικαιώματος.

C:\Users\Master>cd C:\Users\Master\Downloads\AUEB\ΣΑΠ

C:\Users\Master\Downloads\AUEB\ΣΑΠ>trec_eval.exe -m all_trec CISI.REL resultsCISI_BM25_20.txt
1 [main] trec_eval 13608 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer. Please report this problem to the public mailing list cygwin@cygwin.com

n_com      all      STANDARD
runid      all
num_q      all      76
num_ret    all      1520
num_rel    all      3114
num_rel_ret all      406
map         all      0.1027
gm_map     all      0.0350
Rprec      all      0.1554
bpref      all      0.1877
recip_rank all      0.5979
iprec_at_recall_0.00 all      0.6368
iprec_at_recall_0.10 all      0.3698
iprec_at_recall_0.20 all      0.1781
iprec_at_recall_0.30 all      0.0943
iprec_at_recall_0.40 all      0.0504
iprec_at_recall_0.50 all      0.0359
iprec_at_recall_0.60 all      0.0247
iprec_at_recall_0.70 all      0.0109
iprec_at_recall_0.80 all      0.0008
iprec_at_recall_0.90 all      0.0008
iprec_at_recall_1.00 all      0.0008
p_5         all      0.3868
p_10        all      0.3289
p_15        all      0.2939
p_20        all      0.2671
p_30        all      0.1781
p_100       all      0.0534
p_200       all      0.0267
p_500       all      0.0107
p_1000      all      0.0053
recall_5    all      0.0705
recall_10   all      0.1144
```

Εκτελούμε trec_eval.exe -m all_trec CISI.REL resultsCISI_BM25_30.txt στα 30 πιο σχετικά κείμενα

```
C:\Users\Master\Downloads\AUEB\ΣΑΠ>trec_eval.exe -m all_trec CISI.REL resultsCISI_BM25_30.txt
1 [main] trec_eval 23208 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer. Please report this problem to the public mailing list cygwin@cygwin.com

n_com      all      STANDARD
runid      all
num_q      all      76
num_ret    all      2280
num_rel    all      3114
num_rel_ret all      503
map         all      0.1130
gm_map     all      0.0428
Rprec      all      0.1745
bpref      all      0.2173
recip_rank all      0.5983
iprec_at_recall_0.00 all      0.6368
iprec_at_recall_0.10 all      0.4063
iprec_at_recall_0.20 all      0.2119
iprec_at_recall_0.30 all      0.1097
iprec_at_recall_0.40 all      0.0575
iprec_at_recall_0.50 all      0.0493
iprec_at_recall_0.60 all      0.0303
iprec_at_recall_0.70 all      0.0165
iprec_at_recall_0.80 all      0.0008
iprec_at_recall_0.90 all      0.0008
iprec_at_recall_1.00 all      0.0008
p_5         all      0.3868
p_10        all      0.3289
p_15        all      0.2939
p_20        all      0.2671
p_30        all      0.2206
p_100       all      0.0662
p_200       all      0.0331
p_500       all      0.0132
p_1000      all      0.0066
recall_5    all      0.0705
recall_10   all      0.1144
recall_15   all      0.1500
recall_20   all      0.1877
recall_30   all      0.2173
recall_100  all      0.2173
```

Εκτελούμε trec_eval.exe -m all_trec CISI.REL resultsCISI_BM25_50.txt στα 50 πιο σχετικά κείμενα

```
Command Prompt
C:\Users\Master\Downloads\AUEB\IAP>trec_eval.exe -m all_trec CISI.REL resultsCISI_BM25_50.txt
1 [main] trec_eval 190880 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer. Please report this problem to
the public mailing list cygwin@cygwin.com
runid      all      STANDARD
num_q      all      76
num_ret    all      3800
num_rel    all      3114
num_rel_ret all      695
map         all      0.1297
gm_map      all      0.0668
Rprec       all      0.1949
bpref       all      0.3043
recip_rank  all      0.5994
iprec_at_recall_0.00 all 0.6386
iprec_at_recall_0.10 all 0.4351
iprec_at_recall_0.20 all 0.2346
iprec_at_recall_0.30 all 0.1495
iprec_at_recall_0.40 all 0.0852
iprec_at_recall_0.50 all 0.0605
iprec_at_recall_0.60 all 0.0368
iprec_at_recall_0.70 all 0.0212
iprec_at_recall_0.80 all 0.0087
iprec_at_recall_0.90 all 0.0012
iprec_at_recall_1.00 all 0.0012
p_5         all      0.3868
p_10        all      0.3209
p_15        all      0.2939
p_20        all      0.2671
p_30        all      0.2206
p_100       all      0.0914
p_200       all      0.0457
p_500       all      0.0183
p_1000      all      0.0091
recall_5    all      0.0705
recall_10   all      0.1144
recall_15   all      0.1500
recall_20   all      0.1877
recall_30   all      0.2173
recall_100  all      0.3043
```

Εκτελούμε trec_eval.exe -m all_trec CISI.REL resultsCISI_Jelinek_20.txt στα 20 πιο σχετικά κείμενα

```
Command Prompt
C:\Users\Master\Downloads\AUEB\IAP>trec_eval.exe -m all_trec CISI.REL resultsCISI_Jelinek_20.txt
1 [main] trec_eval 9404 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer. Please report this problem to
the public mailing list cygwin@cygwin.com
runid      all      STANDARD
num_q      all      76
num_ret    all      1520
num_rel    all      3114
num_rel_ret all      384
map         all      0.1040
gm_map      all      0.6314
Rprec       all      0.1508
bpref       all      0.1802
recip_rank  all      0.6174
iprec_at_recall_0.00 all 0.6500
iprec_at_recall_0.10 all 0.3625
iprec_at_recall_0.20 all 0.1976
iprec_at_recall_0.30 all 0.0866
iprec_at_recall_0.40 all 0.0536
iprec_at_recall_0.50 all 0.0402
iprec_at_recall_0.60 all 0.0179
iprec_at_recall_0.70 all 0.0102
iprec_at_recall_0.80 all 0.0016
iprec_at_recall_0.90 all 0.0016
iprec_at_recall_1.00 all 0.0016
p_5         all      0.3947
p_10        all      0.3329
p_15        all      0.2825
p_20        all      0.2526
p_30        all      0.1684
p_100       all      0.0505
p_200       all      0.0253
p_500       all      0.0101
p_1000      all      0.0051
recall_5    all      0.0768
recall_10   all      0.1260
recall_15   all      0.1596
recall_20   all      0.1802
recall_30   all      0.1802
recall_100  all      0.1802
```

Εκτελούμε trec_eval.exe -m all_trec CISI.REL resultsCISI_Jelinek_30.txt στα 30 πιο σχετικά κείμενα

```
Command Prompt
C:\Users\Master\Downloads\AUEB\IAP>trec_eval.exe -m all_trec CISI.REL resultsCISI_Jelinek_30.txt
1 [main] trec_eval 3680 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer. Please report this problem to
the public mailing list cygwin@cygwin.com
runid      all      STANDARD
num_q      all      76
num_ret     all      2280
num_rel     all      3114
num_rel_ret all      499
map         all      0.1162
gn_map      all      0.0478
Rprec       all      0.1734
bpref       all      0.2343
recip_rank  all      0.6189
iprec_at_recall_0.00 all      0.6517
iprec_at_recall_0.10 all      0.3998
iprec_at_recall_0.20 all      0.2313
iprec_at_recall_0.30 all      0.1859
iprec_at_recall_0.40 all      0.0698
iprec_at_recall_0.50 all      0.0472
iprec_at_recall_0.60 all      0.0277
iprec_at_recall_0.70 all      0.0178
iprec_at_recall_0.80 all      0.0043
iprec_at_recall_0.90 all      0.0021
iprec_at_recall_1.00 all      0.0021
p_5         all      0.3947
p_10        all      0.3329
p_15        all      0.2825
p_20        all      0.2526
p_30        all      0.2189
p_100       all      0.0657
p_200       all      0.0328
p_500       all      0.0131
p_1000      all      0.0066
recall_5     all      0.0768
recall_10    all      0.1368
recall_15    all      0.1596
recall_20    all      0.1802
recall_30    all      0.2343
recall_100   all      0.2343
```

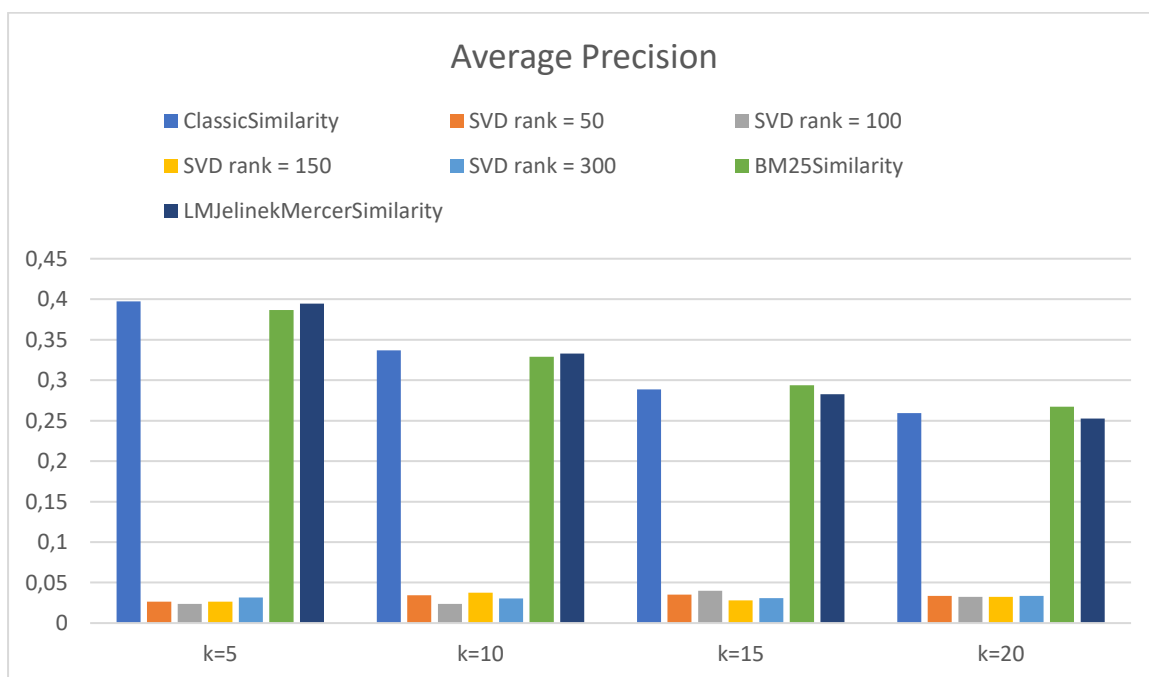
Εκτελούμε trec_eval.exe -m all_trec CISI.REL resultsCISI_Jelinek_50.txt στα 50 πιο σχετικά κείμενα

```
Command Prompt
C:\Users\Master\Downloads\AUEB\IAP>trec_eval.exe -m all_trec CISI.REL resultsCISI_Jelinek_50.txt
1 [main] trec_eval 8624 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer. Please report this problem to
the public mailing list cygwin@cygwin.com
runid      all      STANDARD
num_q      all      76
num_ret     all      3800
num_rel     all      3114
num_rel_ret all      673
map         all      0.1310
gn_map      all      0.0636
Rprec       all      0.1924
bpref       all      0.3035
recip_rank  all      0.6193
iprec_at_recall_0.00 all      0.6531
iprec_at_recall_0.10 all      0.4243
iprec_at_recall_0.20 all      0.2515
iprec_at_recall_0.30 all      0.1430
iprec_at_recall_0.40 all      0.0858
iprec_at_recall_0.50 all      0.0588
iprec_at_recall_0.60 all      0.0334
iprec_at_recall_0.70 all      0.0212
iprec_at_recall_0.80 all      0.0147
iprec_at_recall_0.90 all      0.0021
iprec_at_recall_1.00 all      0.0021
p_5         all      0.3947
p_10        all      0.3329
p_15        all      0.2825
p_20        all      0.2526
p_30        all      0.2189
p_100       all      0.0886
p_200       all      0.0443
p_500       all      0.0177
p_1000      all      0.0080
recall_5     all      0.0768
recall_10    all      0.1368
recall_15    all      0.1596
recall_20    all      0.1802
recall_30    all      0.2343
recall_100   all      0.3035
```

Σύγκριση με προηγούμενες Φάσεις

Average Precision

	k=5	k=10	k=15	k=20
ClassicSimilarity	0,3974	0,3368	0,2886	0,2592
SVD rank = 50	0,0263	0,0342	0,0351	0,0336
SVD rank = 100	0,0237	0,0237	0,0398	0,0322
SVD rank = 150	0,0263	0,0376	0,0281	0,0322
SVD rank = 300	0,0316	0,0303	0,0307	0,0336
BM25Similarity	0,3868	0,3289	0,2939	0,2671
LMJelinekMercerSimilarity	0,3947	0,3329	0,2825	0,2526

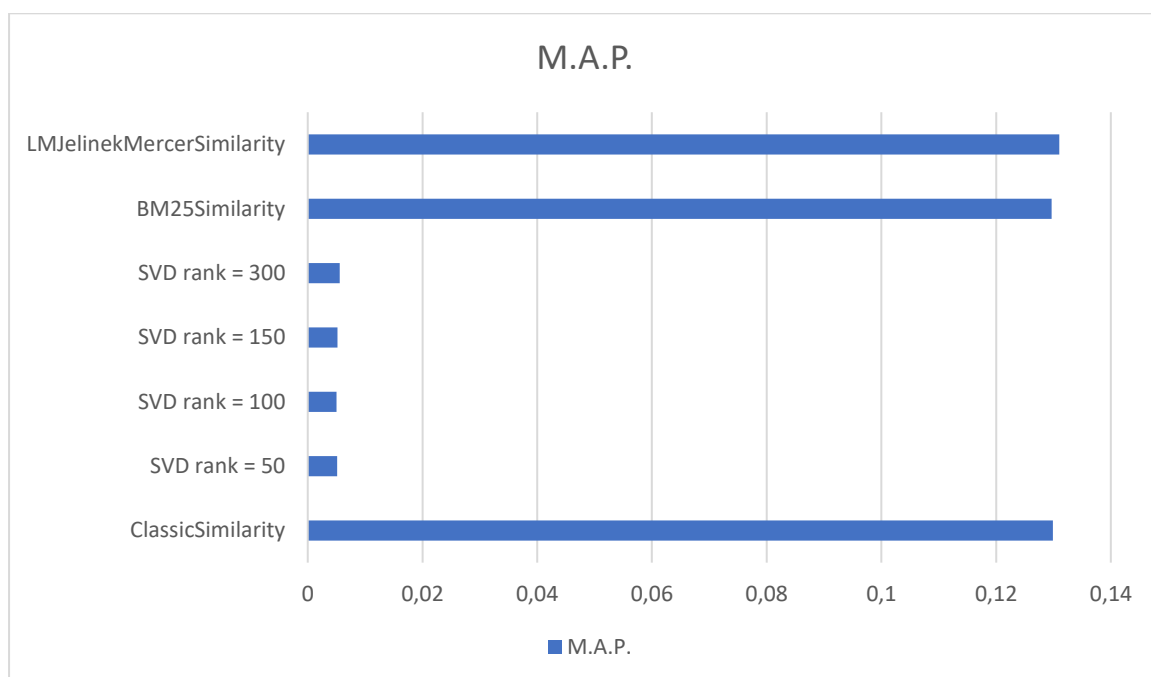


Παρατηρούμε ότι ο SVD βγαζει πολύ χειρότερα αποτελέσματα από τις κλάσεις ομοιότητας. Οι κλάσεις που χρησιμοποιήσαμε στη Φάση 3 έχουν παρόμοια αποτελέσματα με την ClassicSimilarity που χρησιμοποιήσαμε στην Φάση 1.

Παρόμοια είναι και η κατάσταση στο M.A.P.:

Mean Average Precision

	M.A.P.
ClassicSimilarity	0,1299
SVD rank = 50	0,0051
SVD rank = 100	0,0050
SVD rank = 150	0,0052
SVD rank = 300	0,0056
BM25Similarity	0,1297
LMJelinekMercerSimilarity	0,1310



Πηγές:

https://lucene.apache.org/core/8_0_0/core/org/apache/lucene/search/similarities/LMJelinekMercerSimilarity.html