

Συστήματα Ανάκτησης Πληροφοριών

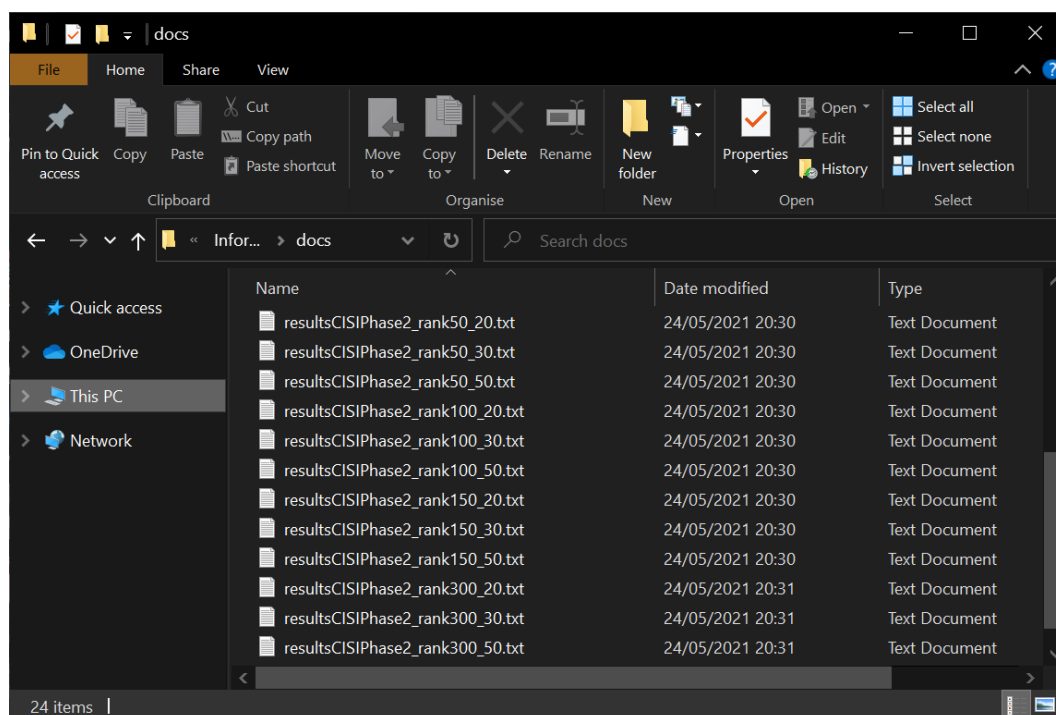
Φάση 2 – Εφαρμογή Λανθάνουσας Σημασιολογικής Ευρετηρίασης στην ανάκτηση

Στυλιανή Δούκα – p3170042

Βασίλειος Μπάλλας – p170115

Περιγραφή Υλοποίησης

1. Για την αποθήκευση των κειμένων στο ευρετήριο χρησιμοποιήσαμε τα πεδία *.I*, *.T*, *.A*, *.W* όπως και στο προηγούμενο στάδιο της εργασίας. Η διαφορά εδώ είναι ότι το πεδίο *.W* αποθηκεύτηκε ως πίνακας συχνότητας (TermXDoc Matrix). Για κάθε ερώτημα (*query*) αποθηκεύσαμε το μοναδικό κωδικό (*id*) του: *.I* και το περιεχόμενο (*content*) του: *.W*. Για την επεξεργασία των κειμένων χρησιμοποιήθηκε ο *EnglishAnalyzer*, τόσο για τα κείμενα που αποθηκεύονται στο ευρετήριο, όσο και για τα ερωτήματα (*queries*) που θα χρησιμοποιηθούν για την αναζήτηση.
2. Για την ανάλυση του πίνακα TermXDocs χρησιμοποιήθηκε η βιβλιοθήκη **Jama**. Με τον αλγόριθμο SVD υπολογίστηκαν οι πίνακες *U*, *S* και *V*. Εφαρμόσαμε την παραλλαγή 2 όπως παρουσιάζεται στο σετ διαφανειών. Συγκεκριμένα, χρησιμοποιήθηκε ο πίνακας *V_k* για την αναπαράσταση των κειμένων, με το *k* να αντιπροσωπεύει τις διαστάσεις του διανύσματος κάθε κειμένου. Εκτελέσαμε για *k* = 50, 100, 150, 300.



3. Για την αναπαράσταση των ερωτημάτων δημιουργήσαμε αραιούς πίνακες μεγέθους $terms \times X1$ που αντιπροσωπεύουν τα $terms$ που εμφανίζονται στο ερώτημα. Οι πίνακες αυτοί μετασχηματίστηκαν σε πυκνούς αφού πολλαπλασιάστηκαν με τους πίνακες U_k και S_k . Τα ερωτήματα είναι τώρα στη σωστή μορφή ώστε να εφαρμοστεί η συνημιτονοειδής ομοιότητα. Στο στάδιο αυτό δεν χρησιμοποιήθηκε η lucene καθώς θέλαμε η σειρά των $terms$ να είναι η ίδια με αυτή που εφαρμόστηκε στα κείμενα.
4. Για κάθε ερώτημα υπολογίσαμε τη συνημιτονοειδή ομοιότητα της αναπαράστασης του με κάθε στήλη του πίνακα V_k , δηλαδή με την αναπαράσταση κάθε κειμένου στο νέο χώρο. Ο τύπος που χρησιμοποιήθηκε είναι: $\cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||}$. Τα αποτελέσματα ταξινομήθηκαν σε φθίνουσα σειρά και κρατήθηκαν τα πρώτα $k = 20, 30, 50$ πιο σχετικά κείμενα.
5. Για την χρήση του εργαλείου `trec_eval` τροποποιήσαμε κατάλληλα το αρχείο `CISI.REL`. Πιο συγκεκριμένα παρατηρήσαμε ότι η τοποθέτηση των κωδικών των αρχείων ήταν λανθασμένη και πως όλες οι τιμές ομοιότητας ήταν μηδενισμένες.

Εκτελέσαμε την εντολή `trec_eval.exe -m all_trec CISI.REL resultsCISIPhase2_rankX_50.txt` όπου X η τάξη που χρησιμοποιήθηκε για την προσέγγιση των πινάκων.

resultsCISIPhase2_rank50_50.txt

```

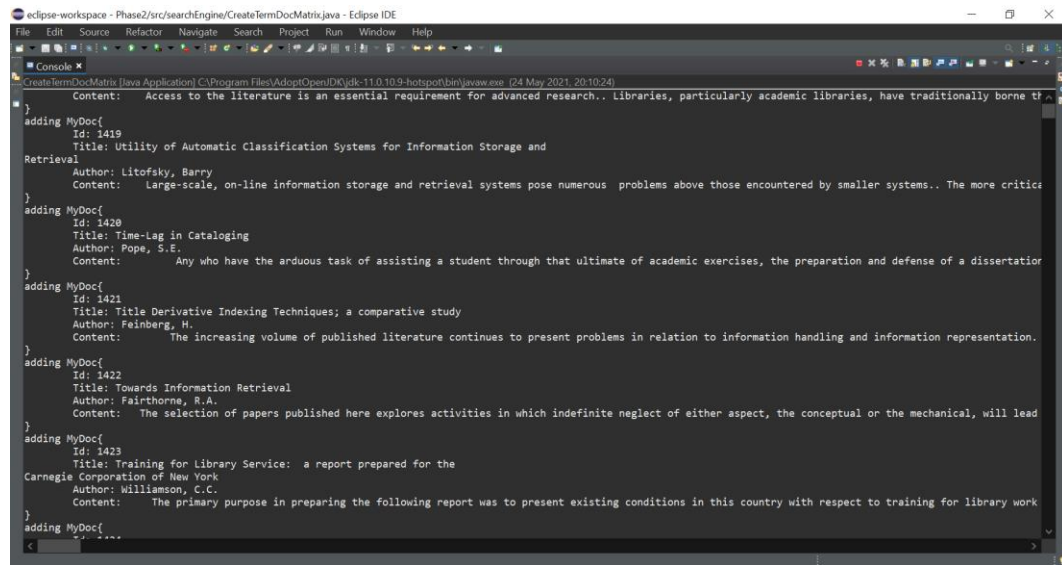
C:\Users\Master\Downloads\AUEB\IAI>trec_eval.exe -m all_trec CISI.REL resultsCISIPhase2_rank50_50.txt
1 [main] trec_eval 22024 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer. Please report this problem to the public mailing list cygwin@cygwin
n_cow          all          STANDARD
runid          all          76
num_q          all          3876
num_ret        all          3114
num_rel        all          164
num_rel_ret    all          0.0051
map            all          0.0008
gm_map         all          0.0284
Rprec          all          0.0586
bpref          all          0.0837
recip_rank     all          0.0966
iprec_at_recall_0.00 all    0.0166
iprec_at_recall_0.20 all    0.0011
iprec_at_recall_0.30 all    0.0003
iprec_at_recall_0.40 all    0.0000
iprec_at_recall_0.50 all    0.0000
iprec_at_recall_0.60 all    0.0000
iprec_at_recall_0.70 all    0.0000
iprec_at_recall_0.80 all    0.0000
iprec_at_recall_0.90 all    0.0000
iprec_at_recall_1.00 all    0.0000
p_5            all          0.0263
p_10           all          0.0342
p_15           all          0.0351
p_20           all          0.0336
p_30           all          0.0333
p_100          all          0.0216
p_200          all          0.0108
p_500          all          0.0043
p_1000         all          0.0022
recall_5       all          0.0023
recall_10      all          0.0057
recall_15      all          0.0116
recall_20      all          0.0139
recall_30      all          0.0230
recall_100     all          0.0586
recall_200     all          0.0586

```

k = 5	k = 10	k = 15	k = 20
0.0263	0.0342	0.0351	0.0336

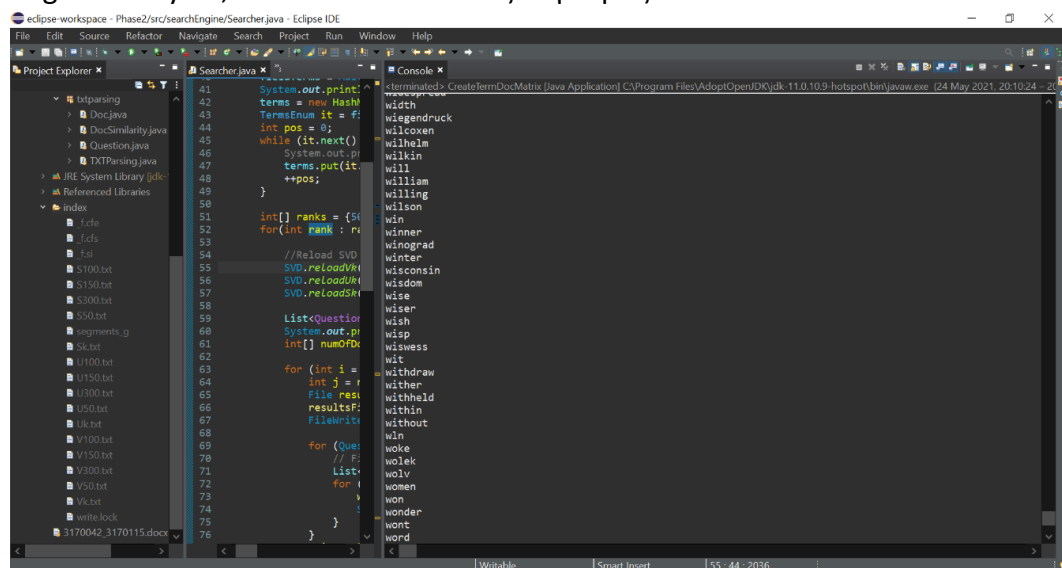
Επιπλέον η τιμή του mean average precision (MAP) για rank = 50 είναι 0.0051.

Απόσπασμα από την εκτέλεση του αρχείου CreateTermDocMatrix.java όπου δημιουργείται ο αραιός πίνακας. Βλέπουμε ότι τα documents προστίθενται στο ευρετήριο.



```
eclipse-workspace - Phase2/src/searchEngine/CreateTermDocMatrix.java - Eclipse IDE
CreateTermDocMatrix [Java Application] C:\Program Files\AdoptOpenJDK\jdk-11.0.10-hotspot\bin\javaw.exe (24 May 2021, 20:10:24)
Content: Access to the literature is an essential requirement for advanced research.. Libraries, particularly academic libraries, have traditionally borne the brunt of the...
}
adding MyDoc{
  Id: 1419
  Title: Utility of Automatic Classification Systems for Information Storage and Retrieval
  Author: Litofsky, Barry
  Content: Large-scale, on-line information storage and retrieval systems pose numerous problems above those encountered by smaller systems.. The more critical the...
}
adding MyDoc{
  Id: 1420
  Title: Time-Lag in Cataloging
  Author: Pope, S.E.
  Content: Any who have the arduous task of assisting a student through that ultimate of academic exercises, the preparation and defense of a dissertation...
}
adding MyDoc{
  Id: 1421
  Title: Title Derivative Indexing Techniques; a comparative study
  Author: Feinberg, M.
  Content: The increasing volume of published literature continues to present problems in relation to information handling and information representation.
}
adding MyDoc{
  Id: 1422
  Title: Towards Information Retrieval
  Author: Fairthorne, R.A.
  Content: The selection of papers published here explores activities in which indefinite neglect of either aspect, the conceptual or the mechanical, will lead to...
}
adding MyDoc{
  Id: 1423
  Title: Training for Library Service: a report prepared for the Carnegie Corporation of New York
  Author: Williamson, C.C.
  Content: The primary purpose in preparing the following report was to present existing conditions in this country with respect to training for library work...
}
adding MyDoc{
  Id: 1424
  Title: ...
  Author: ...
  Content: ...
}
```

Μέρος των μοναδικών λέξεων που εντόπισε η Lucene, επεξεργασμένων με τον English Analyser, και αποτελούν το λεξιλόγιο μας.



```
eclipse-workspace - Phase2/src/searchEngine/Searcher.java - Eclipse IDE
Project Explorer:
- src
  - Indexing
  - Doc.java
  - DocSimilarity.java
  - Question.java
  - TXTParsing.java
  - JRE System Library [jdk-11.0.10]
  - Referenced Libraries
  - index
    - fcd
    - fcds
    - fsi
    - S100.txt
    - S150.txt
    - S300.txt
    - S50.txt
    - segments.g
    - Sk.txt
    - U100.txt
    - U150.txt
    - U300.txt
    - U50.txt
    - Uk.txt
    - V100.txt
    - V150.txt
    - V300.txt
    - V50.txt
    - V.txt
    - write.lock
    - 3170042_3170115.docx
  - ...

Searcher.java:
41 System.out.println("Terms:");
42 terms = new HashSet<>();
43 TermsEnum it = f.getTermsEnum();
44 int pos = 0;
45 while (it.next()) {
46   System.out.println(it.toString());
47   terms.put(it.toString());
48   ++pos;
49 }
50
51 int[] ranks = {50, 40, 30, 20, 10, 5, 4, 3, 2, 1};
52 for (int rank : ranks) {
53   //Reload SVD
54   SVD.reloadVr(rank);
55   SVD.reloadUk(rank);
56   SVD.reloadSk(rank);
57 }
58
59 List<Question> questions = new ArrayList<>();
60 System.out.println("Questions:");
61 int[] numOfDocs = new int[terms.size()];
62
63 for (int i = 0; i < terms.size(); i++) {
64   int j = 0;
65   File res = new File("results/" + terms.get(i) + ".txt");
66   results = new ArrayList<>();
67   FileWrt wrt = new FileWrt(res);
68
69   for (Question q : questions) {
70     // F:
71     List<String> list = new ArrayList<>();
72     for (String s : terms) {
73       if (s.equals(terms.get(i))) {
74         list.add(s);
75       }
76     }
77   }
78 }

Console:
CreateTermDocMatrix [Java Application] C:\Program Files\AdoptOpenJDK\jdk-11.0.10-hotspot\bin\javaw.exe (24 May 2021, 20:10:24)
Terms:
width
wiegendruck
wilcoxon
wilhelm
wilkin
will
william
willing
wilson
win
winner
winograd
winter
wisconsin
wisdom
wise
wiser
wish
wisp
wisness
wit
withdraw
wither
withheld
within
without
win
woke
wolek
wolv
women
won
wonder
wont
word
```

Απόσπασμα από την εκτέλεση του Searcher για την αναζήτηση των ερωτημάτων. Βλέπουμε ότι αναζητούνται κείμενα σχετικά με τα ερωτήματα 70 και 71 και τυπώνονται τα κείμενα και ο βαθμός ομοιότητας τους σύμφωνα με τη μορφοποίηση που δέχεται το εργαλείο trec_eval.

```

<terminated> Searcher (1) [Java Application] C:\Program Files\AdoptOpenJDK\jdk-11.0.109-hotspot\bin\javaw.exe (24 May 2021, 20:30:17 - 20:31:33)
70 Q0 29 0 0.1722647745228282 STANDARD
70 Q0 30 0 0.14146764795783967 STANDARD
70 Q0 31 0 0.15290125945194616 STANDARD
70 Q0 32 0 0.17721513812006204 STANDARD
70 Q0 33 0 0.19636856413837983 STANDARD
70 Q0 34 0 0.16139765890611634 STANDARD
70 Q0 35 0 0.19074681348661288 STANDARD
70 Q0 36 0 0.14404757070136767 STANDARD
70 Q0 37 0 0.17843212643024614 STANDARD
70 Q0 38 0 0.15851887729003236 STANDARD
70 Q0 39 0 0.18201058472895643 STANDARD
70 Q0 40 0 0.16276795467415645 STANDARD
70 Q0 41 0 0.2017579276382637 STANDARD
70 Q0 42 0 0.15010580278861008 STANDARD
70 Q0 43 0 0.1651592891308024 STANDARD
70 Q0 44 0 0.2005016662211466 STANDARD
70 Q0 45 0 0.17556727715777343 STANDARD
70 Q0 46 0 0.1632532948596444 STANDARD
70 Q0 47 0 0.1574273208895885 STANDARD
70 Q0 48 0 0.20406319930928055 STANDARD
70 Q0 49 0 0.1728387394493794 STANDARD
70 Q0 50 0 0.19954084355362914 STANDARD
71 Q0 0 0 0.20259670379557848 STANDARD
71 Q0 1 0 0.13891950751077248 STANDARD
71 Q0 2 0 0.1342763717236754 STANDARD
71 Q0 3 0 0.1826876980251235 STANDARD
71 Q0 4 0 0.23412144795393644 STANDARD
71 Q0 5 0 0.12952431323183117 STANDARD
71 Q0 6 0 0.17878756362228734 STANDARD
71 Q0 7 0 0.1680020397157488 STANDARD
71 Q0 8 0 0.17236876188144848 STANDARD
71 Q0 9 0 0.15704478974923433 STANDARD
71 Q0 10 0 0.19252906356653934 STANDARD
71 Q0 11 0 0.1485751806656329 STANDARD
71 Q0 12 0 0.17283723226759523 STANDARD
  
```

6. Αν επαναλάβουμε τα προηγούμενα βήματα για τάξη = 100, 150 και 300 θα λάβουμε διαφορετικά αποτελέσματα τα οποία φαίνονται τις επόμενες εικόνες:

resultsCISIPhase2_rank100_50.txt

```

C:\Users\Master\Downloads\AUEB\IAN>trec_eval.exe -m all trec CISI.REL resultsCISIPhase2_rank100_50.txt
1 [main] trec_eval 2448 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      3876
num_rel    all      3114
num_rel_ret all      164
map        all      0.0050
gm_map     all      0.0000
Rprec      all      0.0278
bpref      all      0.0506
recip_rank all      0.0746
iprec_at_recall_0.00 all 0.0915
iprec_at_recall_0.10 all 0.0174
iprec_at_recall_0.20 all 0.0012
iprec_at_recall_0.30 all 0.0003
iprec_at_recall_0.40 all 0.0000
iprec_at_recall_0.50 all 0.0000
iprec_at_recall_0.60 all 0.0000
iprec_at_recall_0.70 all 0.0000
iprec_at_recall_0.80 all 0.0000
iprec_at_recall_0.90 all 0.0000
iprec_at_recall_1.00 all 0.0000
P_5        all      0.0237
P_10       all      0.0237
P_15       all      0.0298
P_20       all      0.0322
P_30       all      0.0351
P_100      all      0.0216
P_200      all      0.0108
P_500      all      0.0043
P_1000     all      0.0022
recall_5   all      0.0029
recall_10  all      0.0048
recall_15  all      0.0091
recall_20  all      0.0134
recall_30  all      0.0239
recall_100 all      0.0506
  
```

k = 5	k = 10	k = 15	k = 20
0.0237	0.0237	0.0398	0.0322

Map = 0.0050

resultsCISIPhase2_rank150_50.txt

```

C:\Users\Master\Downloads\AUEB\IAN>trc_eval.exe -m all_trc CISI.REL resultsCISIPhase2_rank150_50.txt
1 [main] trc_eval 8948 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      3876
num_rel        all      3114
num_rel_ret    all      164
map            all      0.0052
gm_map         all      0.0008
Rprec          all      0.0283
bpref          all      0.0586
recip_rank     all      0.0002
iprec_at_recall_0.00 all    0.0948
iprec_at_recall_0.10 all    0.0173
iprec_at_recall_0.20 all    0.0012
iprec_at_recall_0.30 all    0.0004
iprec_at_recall_0.40 all    0.0000
iprec_at_recall_0.50 all    0.0000
iprec_at_recall_0.60 all    0.0000
iprec_at_recall_0.70 all    0.0000
iprec_at_recall_0.80 all    0.0000
iprec_at_recall_0.90 all    0.0000
iprec_at_recall_1.00 all    0.0000
P_5            all      0.0263
P_10           all      0.0276
P_15           all      0.0231
P_20           all      0.0322
P_30           all      0.0360
P_100          all      0.0216
P_200          all      0.0108
P_500          all      0.0043
P_1000         all      0.0022
recall_5       all      0.0030
recall_10      all      0.0057
recall_15      all      0.0084
recall_20      all      0.0131
recall_30      all      0.0247
recall_100     all      0.0506

```

k = 5	k = 10	k = 15	k = 20
0.0263	0.0376	0.0281	0.0322

Map = 0.0052

resultsCISIPhase2_rank300_50.txt

```

C:\Users\Master\Downloads\AUEB\IAN>trc_eval.exe -m all_trc CISI.REL resultsCISIPhase2_rank300_50.txt
1 [main] trc_eval 7676 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      3876
num_rel        all      3114
num_rel_ret    all      164
map            all      0.0056
gm_map         all      0.0008
Rprec          all      0.0267
bpref          all      0.0586
recip_rank     all      0.0952
iprec_at_recall_0.00 all    0.1114
iprec_at_recall_0.10 all    0.0171
iprec_at_recall_0.20 all    0.0012
iprec_at_recall_0.30 all    0.0003
iprec_at_recall_0.40 all    0.0000
iprec_at_recall_0.50 all    0.0000
iprec_at_recall_0.60 all    0.0000
iprec_at_recall_0.70 all    0.0000
iprec_at_recall_0.80 all    0.0000
iprec_at_recall_0.90 all    0.0000
iprec_at_recall_1.00 all    0.0000
P_5            all      0.0316
P_10           all      0.0303
P_15           all      0.0307
P_20           all      0.0336
P_30           all      0.0342
P_100          all      0.0216
P_200          all      0.0108
P_500          all      0.0043
P_1000         all      0.0022
recall_5       all      0.0035
recall_10      all      0.0060
recall_15      all      0.0089
recall_20      all      0.0146
recall_30      all      0.0251
recall_100     all      0.0586

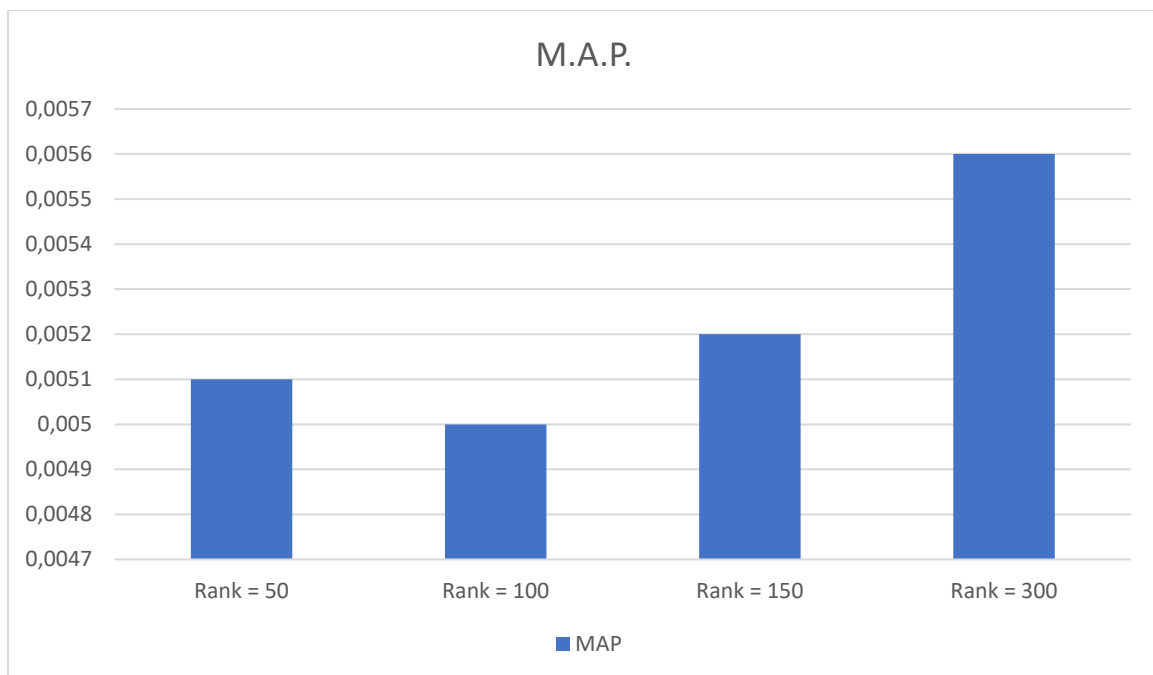
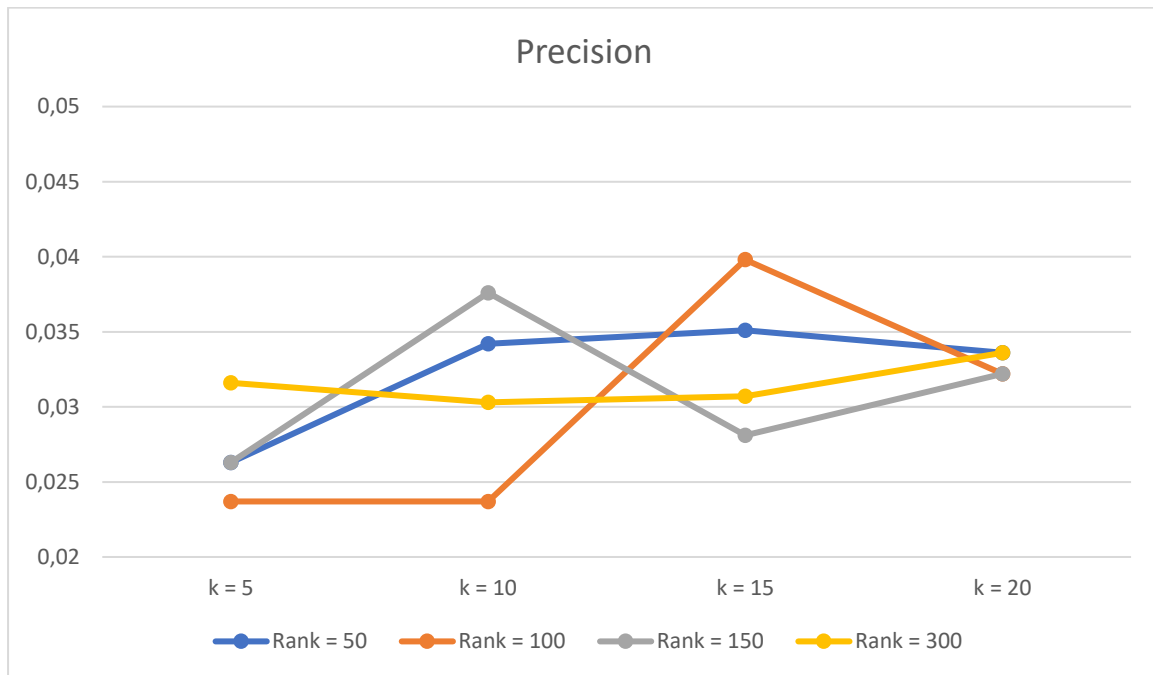
```

k = 5	k = 10	k = 15	k = 20
0.0316	0.0303	0.0307	0.0336

Map = 0.0056

Σύγκριση:

	k = 5	k = 10	k = 15	k = 20	Map
Rank = 50	0.0263	0.0342	0.0351	0.0336	0.0051
Rank = 100	0.0237	0.0237	0.0398	0.0322	0.0050
Rank = 150	0.0263	0.0376	0.0281	0.0322	0.0052
Rank = 300	0.0316	0.0303	0.0307	0.0336	0.0056



Παρατηρούμε ότι δεν υπάρχει δεν υπάρχει κάποια σταθερά με τη χρήση της οποίας οι ακρίβεια βελτιώνεται. Συνεπώς είναι στο χέρι μας να δοκιμάσουμε αρκετές υπερπαραμέτρους ώστε να βρούμε αυτές που ταιριάζουν καλύτερα στο πρόβλημα και τα δεδομένα μας.

Ωστόσο, είναι γεγονός ότι το Mean Average Precision (MAP) βελτιώνεται όταν χρησιμοποιούμε προσεγγίσεις μεγαλύτερης τάξης.

Σε αντίθεση με τα ευρήματα της προηγούμενης φάσης, η Ακρίβεια **δεν** μειώνεται όσο αυξάνεται το k .

Πηγές:

<https://introcs.cs.princeton.edu/java/95linear/SVD.java.html>

<https://math.nist.gov/javanumerics/jama/doc/>

<https://stackoverflow.com/questions/520241/how-do-i-calculate-the-cosine-similarity-of-two-vectors>