

Συστήματα Ανάκτησης Πληροφοριών

Φάση 4 - Ανάκτηση χρησιμοποιώντας Ενσωματώσεις Λέξεων (Word Embeddings)

Στυλιανή Δούκα – p3170042

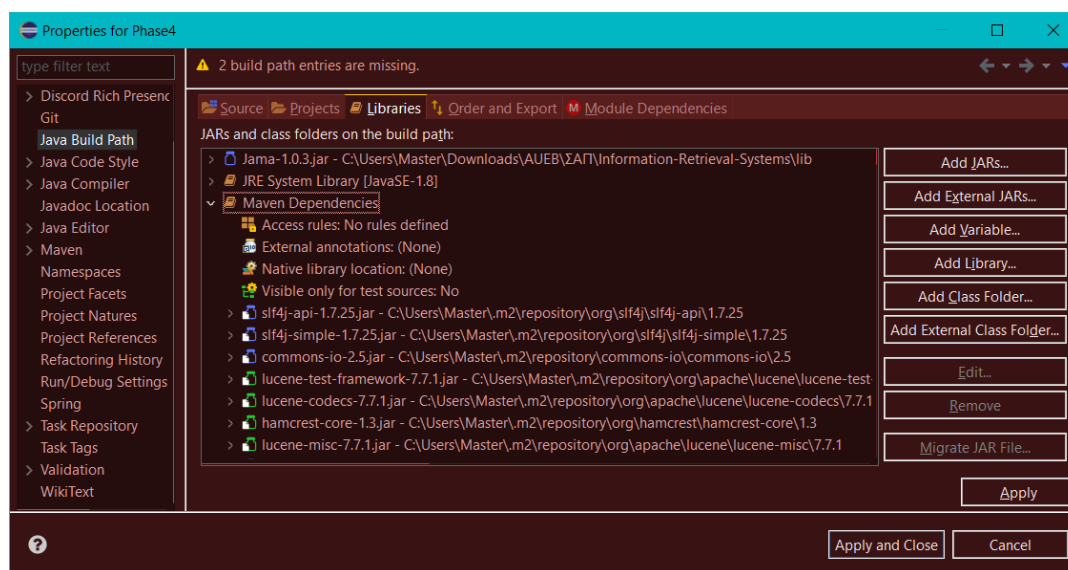
Βασίλειος Μπάλλας – p170115

Περιγραφή Υλοποίησης

Για τις ανάγκες αυτής της Φάσης της εργασίας πραγματοποιήσαμε αρκετές αλλαγές σε επίπεδο κώδικα αλλά και στην δομή του Project.

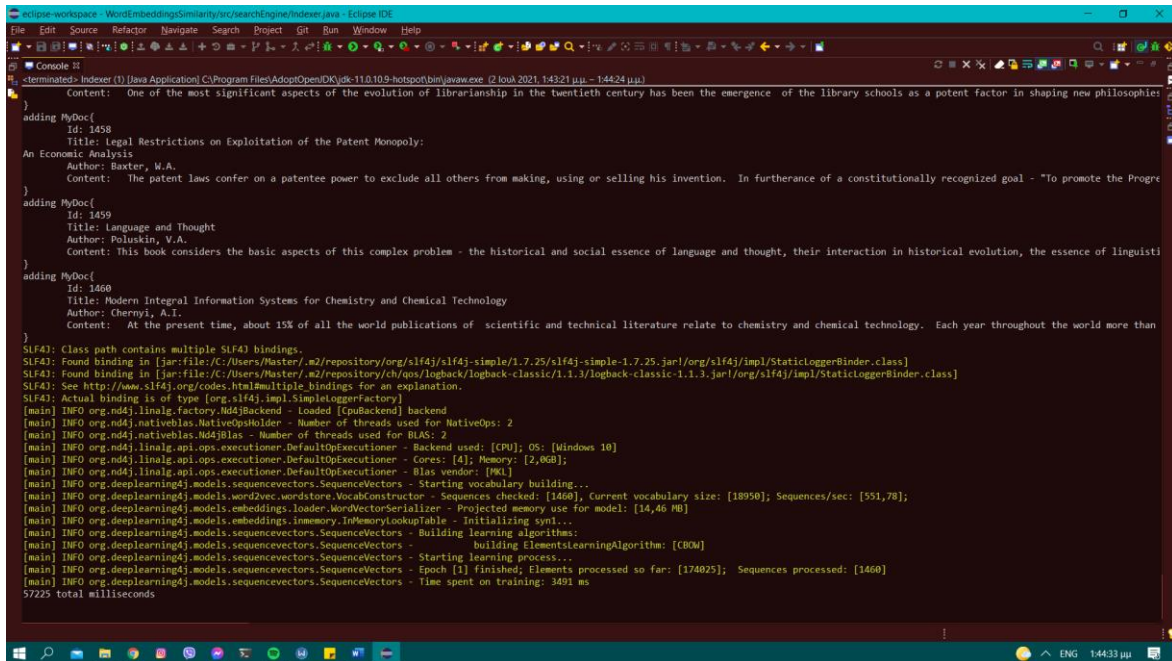
1. Δομή του Project:

Για να πραγματοποιηθεί το ζητούμενο αυτής της Φάσης χρειάστηκε να προσθέσουμε την βιβλιοθήκη `deeplearning4JAVA`. Ο πιο εύκολος τρόπος για να το κάνουμε αυτό ήταν να μετατρέψουμε το project μας σε Maven Project ώστε να φορτώσουμε την βιβλιοθήκη σαν `dependency` στο πρόγραμμα. Χρήσιμο φάνηκε το φροντιστήριο του τελευταίου μαθήματος καθώς είδαμε την δομή του `pom.xml` που χρειαζόμασταν. Για τα υπόλοιπα κομμάτια της εργασίας οι παροχές του Eclipse για Maven Projects φάνηκαν αρκετές.



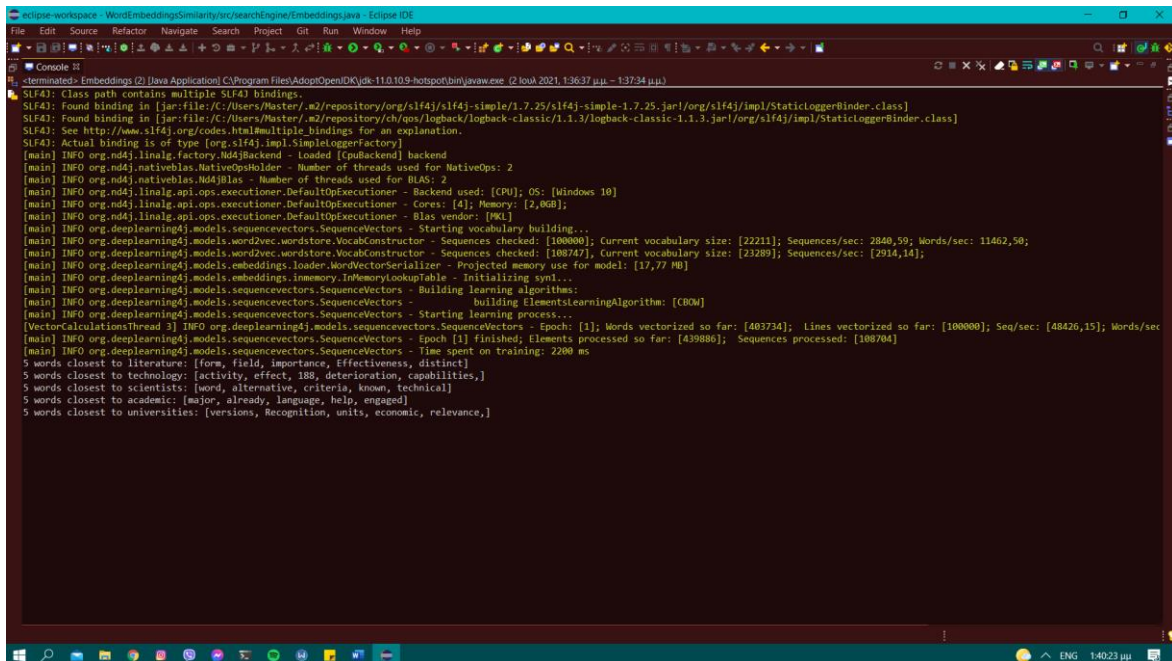
2. Υλοποίηση Κώδικα:

Χρειάστηκαν αρκετές αλλαγές από τις προηγούμενες φάσεις για να λειτουργήσει το πρόγραμμα για τον σκοπό αυτής της φάσης. Ωστόσο, πέραν από τα σταθερά κομμάτια κώδικα που χρησιμοποιούνται σε όλες τις φάσεις, χρήσιμα φάνηκαν σημεία κώδικα από την 2^η Φάση (ο τρόπος που χειριστήκαμε τον SVD...).



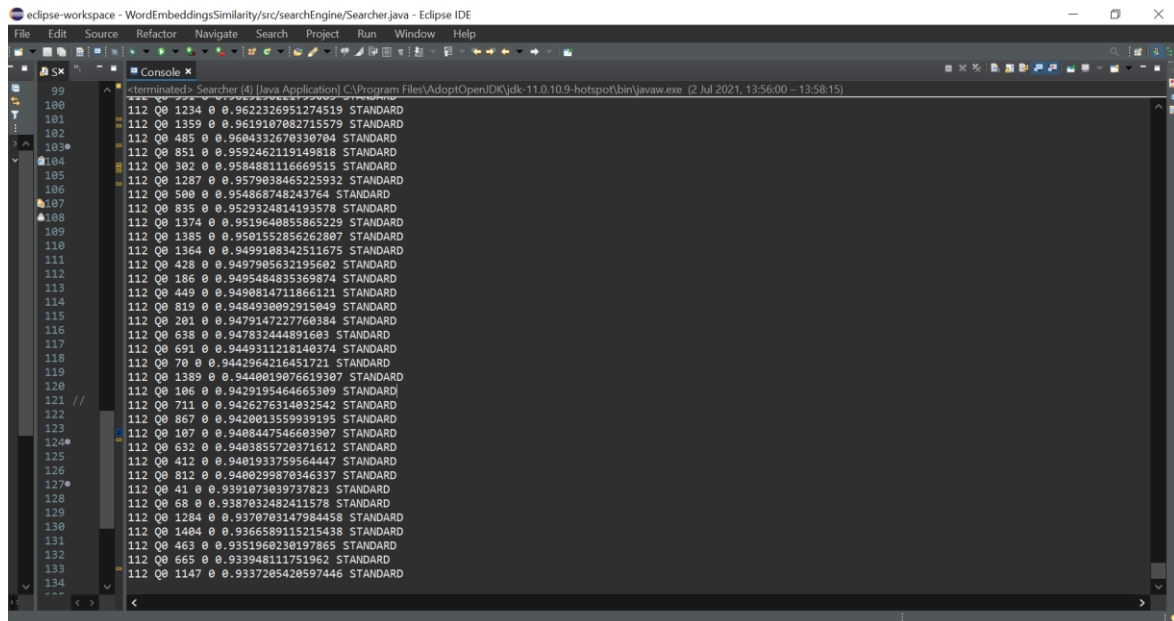
```
.eclipse-workspace WordEmbeddingsSimilarity/src/searchEngine/Indexer.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Git Run Window Help
<terminated> Indexer (1) [Java Application] C:\Program Files\AdoptOpenJDK\jdk-11.0.10-hotspot\bin\javaw.exe (2 Ιου 2021, 14:32 μ.μ. - 14:42 μ.μ.)
Content: One of the most significant aspects of the evolution of librarianship in the twentieth century has been the emergence of the library schools as a potent factor in shaping new philosophies
}
adding MyDoc{
  Id: 1458
  Title: Legal Restrictions on Exploitation of the Patent Monopoly:
  An Economic Analysis
  Author: Baxter, W.A.
  Content: The patent laws confer on a patentee power to exclude all others from making, using or selling his invention. In furtherance of a constitutionally recognized goal - "To promote the Progre
}
adding MyDoc{
  Id: 1459
  Title: Language and Thought
  Author: Poluskin, V.A.
  Content: This book considers the basic aspects of this complex problem - the historical and social essence of language and thought, their interaction in historical evolution, the essence of linguisti
}
adding MyDoc{
  Id: 1460
  Title: Modern Integral Information Systems for Chemistry and Chemical Technology
  Author: Chernys, A.I.
  Content: At the present time, about 15% of all the world publications of scientific and technical literature relate to chemistry and chemical technology. Each year throughout the world more than
}
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/Users/Master/.m2/repository/org/slf4j/slf4j-simple/1.7.25/slf4j-simple-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/Users/Master/.m2/repository/ch/qos/logback/logback-classic/1.1.3/logback-classic-1.1.3.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.SimpleLoggerFactory]
[main] INFO org.nd4j.linalg.factory.Nd4jBackend - Loaded [cpuBackend] backend
[main] INFO org.nd4j.nativeblas.NativeOpsHolder - Number of threads used for NativeOps: 2
[main] INFO org.nd4j.nativeblas.Nd4jBblas - Number of threads used for BLAS: 2
[main] INFO org.nd4j.linalg.api.ops.executioner.DefaultOpExecutioner - Backend used: [CPU]; OS: [Windows 10]
[main] INFO org.nd4j.linalg.api.ops.executioner.DefaultOpExecutioner - Cores: [4]; Memory: [2,060]
[main] INFO org.nd4j.linalg.api.ops.executioner.DefaultOpExecutioner - Blas vendor: [MKL]
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Starting vocabulary building...
[main] INFO org.deeplearning4j.models.word2vec.wordstore.VocabConstructor - Sequences checked: [1460], Current vocabulary size: [18950]; Sequences/sec: [551,78];
[main] INFO org.deeplearning4j.models.embeddings.loader.WordVectorSerializer - Projected memory use for model: [14,46 MB]
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Epoch [1] finished; Elements processed so far: [174025]; Sequences processed: [1460]
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Building learning algorithms:
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - building ElementsLearningAlgorithm: [CBOW]
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Starting learning process...
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Epoch [1] finished; Elements processed so far: [174025]; Sequences processed: [1460]
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Time spent on training: 3491 ms
57225 total milliseconds
```

Εικόνα 1: Indexer.java



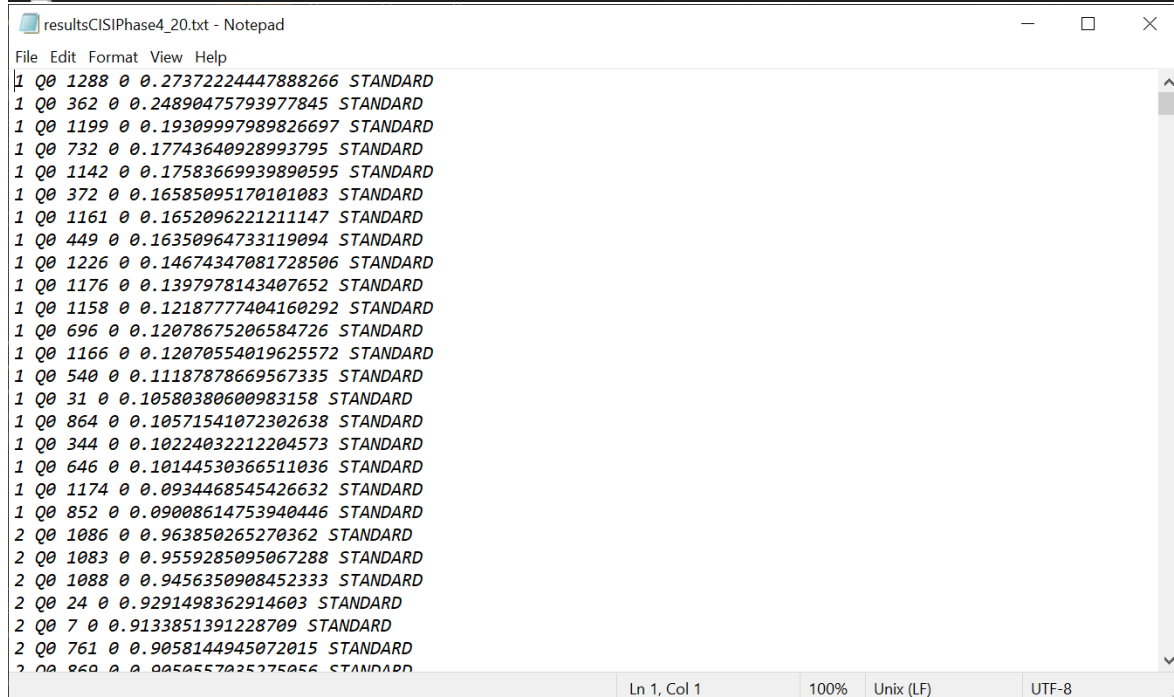
```
.eclipse-workspace WordEmbeddingsSimilarity/src/searchEngine/Embeddings.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Git Run Window Help
<terminated> Embeddings (2) [Java Application] C:\Program Files\AdoptOpenJDK\jdk-11.0.10-hotspot\bin\javaw.exe (2 Ιου 2021, 13:37 μ.μ. - 13:34 μ.μ.)
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/Users/Master/.m2/repository/org/slf4j/slf4j-simple/1.7.25/slf4j-simple-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/Users/Master/.m2/repository/ch/qos/logback/logback-classic/1.1.3/logback-classic-1.1.3.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.SimpleLoggerFactory]
[main] INFO org.nd4j.linalg.factory.Nd4jBackend - Loaded [cpuBackend] backend
[main] INFO org.nd4j.nativeblas.NativeOpsHolder - Number of threads used for NativeOps: 2
[main] INFO org.nd4j.nativeblas.Nd4jBblas - Number of threads used for BLAS: 2
[main] INFO org.nd4j.linalg.api.ops.executioner.DefaultOpExecutioner - Backend used: [CPU]; OS: [Windows 10]
[main] INFO org.nd4j.linalg.api.ops.executioner.DefaultOpExecutioner - Cores: [4]; Memory: [2,060]
[main] INFO org.nd4j.linalg.api.ops.executioner.DefaultOpExecutioner - Blas vendor: [MKL]
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Starting vocabulary building...
[main] INFO org.deeplearning4j.models.word2vec.wordstore.VocabConstructor - Sequences checked: [100000]; Current vocabulary size: [22211]; Sequences/sec: 2840,59; Words/sec: 11462,50;
[main] INFO org.deeplearning4j.models.word2vec.wordstore.VocabConstructor - Sequences checked: [108747], Current vocabulary size: [23289]; Sequences/sec: [2914,14];
[main] INFO org.deeplearning4j.models.embeddings.loader.WordVectorSerializer - Projected memory use for model: [17,77 MB]
[main] INFO org.deeplearning4j.models.embeddings.inmemory.InMemoryLookupTable - Initializing synl...
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Building learning algorithms:
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - building ElementsLearningAlgorithm: [CBOW]
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Starting learning process...
[VectorCalculationsThread 3] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Epoch: [1]; Words vectorized so far: [403734]; Lines vectorized so far: [100000]; Seq/sec: [48426,15]; Words/sec:
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Epoch [1] finished; Elements processed so far: [439880]; Sequences processed: [108704]
[main] INFO org.deeplearning4j.models.sequencevectors.SequenceVectors - Time spent on training: 2200 ms
5 words closest to literature: [form, field, importance, effectiveness, distinct]
5 words closest to technology: [activity, effect, 180, deterioration, capabilities,]
5 words closest to scientists: [word, alternative, criteria, known, technical]
5 words closest to academic: [major, already, language, help, engaged]
5 words closest to universities: [versions, Recognition, units, economic, relevance,]
```

Εικόνα 2: Embeddings.java



Αποθηκεύουμε τα αποτελέσματα σε αρχεία .txt

resultsCISIPhase4_20.txt	02/07/2021 13:56	Text Document
resultsCISIPhase4_30.txt	02/07/2021 13:57	Text Document
resultsCISIPhase4_50.txt	02/07/2021 13:58	Text Document



Στη συνέχεια περάσαμε τα αποτελέσματα από το Trec_eval:

```

C:\Users\Master\Downloads\AUEB\ΣΑΠ>trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_20.txt
1 [main] trec_eval 16476 find_fast_cwd: WARNING: Couldn't compute FAST_CWD pointer. Please report this problem to
n.com
runid          all      STANDARD
num_q          all      76
num_ret        all      1520
num_rel        all      3114
num_rel_ret    all      43
map            all      0.0029
gm_map         all      0.0001
Rprec          all      0.0119
bpref          all      0.0175
recip_rank     all      0.0727
iprec_at_recall_0.00 all    0.0763
iprec_at_recall_0.10 all    0.0075
iprec_at_recall_0.20 all    0.0009
iprec_at_recall_0.30 all    0.0009
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      0.0184
P_10           all      0.0250
P_15           all      0.0246
P_20           all      0.0283
P_30           all      0.0189
P_100          all      0.0057
P_200          all      0.0028
P_500          all      0.0011
P_1000         all      0.0006
recall_5       all      0.0027
recall_10      all      0.0071
recall_15      all      0.0131
recall_20      all      0.0175
recall_30      all      0.0175
recall_100     all      0.0175
recall_200     all      0.0175
recall_500     all      0.0175
recall_1000    all      0.0175
infAP          all      0.0029
gm_bpref       all      0.0003

```

1. trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_20.txt

```

C:\Users\Master\Downloads\AUEB\ΣΑΠ>trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_30.txt
1 [main] trec_eval 3736 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      60
map            all      0.0032
gm_map         all      0.0002
Rprec          all      0.0146
bpref          all      0.0224
recip_rank     all      0.0754
iprec_at_recall_0.00 all    0.0801
iprec_at_recall_0.10 all    0.0075
iprec_at_recall_0.20 all    0.0009
iprec_at_recall_0.30 all    0.0009
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      0.0184
P_10           all      0.0250
P_15           all      0.0246
P_20           all      0.0283
P_30           all      0.0263
P_100          all      0.0079
P_200          all      0.0039
P_500          all      0.0016
P_1000         all      0.0008
recall_5       all      0.0027
recall_10      all      0.0071
recall_15      all      0.0131
recall_20      all      0.0175
recall_30      all      0.0224
recall_100     all      0.0224
recall_200     all      0.0224
recall_500     all      0.0224
recall_1000    all      0.0224
infAP          all      0.0032
gm_bpref       all      0.0005

```

2. trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_30.txt

```

C:\Users\Master\Downloads\AUEB\ΣΑΠ>trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_50.txt
1 [main] trec_eval 14076 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          107
map            all          0.0039
gm_map         all          0.0003
Rprec          all          0.0211
bpref          all          0.0340
recip_rank     all          0.0786
iprec_at_recall_0.00 all    0.0848
iprec_at_recall_0.10 all    0.0094
iprec_at_recall_0.20 all    0.0009
iprec_at_recall_0.30 all    0.0009
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          0.0184
P_10           all          0.0250
P_15           all          0.0246
P_20           all          0.0283
P_30           all          0.0263
P_100          all          0.0141
P_200          all          0.0070
P_500          all          0.0028
P_1000         all          0.0014
recall_5       all          0.0027
recall_10      all          0.0071
recall_15      all          0.0131
recall_20      all          0.0175
recall_30      all          0.0224
recall_100     all          0.0340
recall_200     all          0.0340
recall_500     all          0.0340
recall_1000    all          0.0340
infAP          all          0.0039
gm_bpref       all          0.0016

```

3. *trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_50.txt*

3. Χρήση Προεκπαιδευμένου μοντέλου

Για τις ανάγκες αυτού του σταδίου της εργασίας, χρησιμοποιήσαμε μόνο την κλάση Searcher και σε αυτήν φορτώσαμε το προεκπαιδευμένο μοντέλο από την Wikipedia αντί του δικού μας.

Τα αποτελέσματα του trec_eval ήταν διαφορετικά:

```

C:\Windows\System32\cmd.exe
C:\Users\Stella\Documents\InfRetrieval\trec_eval>trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_20.txt
1 [main] trec_eval 2036 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          183
map            all          0.0284
gm_map         all          0.0028
Rprec          all          0.0671
bpref          all          0.0727
recip_rank     all          0.3044
iprec_at_recall_0.00 all    0.3283
iprec_at_recall_0.10 all    0.0892
iprec_at_recall_0.20 all    0.0303
iprec_at_recall_0.30 all    0.0178
iprec_at_recall_0.40 all    0.0079
iprec_at_recall_0.50 all    0.0079
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          0.1711
P_10           all          0.1408
P_15           all          0.1333
P_20           all          0.1204
P_30           all          0.0803
P_100          all          0.0241
P_200          all          0.0120

```

4. trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_20.txt

```

C:\Windows\System32\cmd.exe

C:\Users\Stella\Documents\InfRetrieval\trec_eval>trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_20.txt
1 [main] trec_eval 7320 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      246
map            all      0.0321
gm_map         all      0.0043
Rprec          all      0.0799
bpref          all      0.0918
recip_rank     all      0.3066
iprec_at_recall_0.00 all 0.3338
iprec_at_recall_0.10 all 0.0995
iprec_at_recall_0.20 all 0.0432
iprec_at_recall_0.30 all 0.0196
iprec_at_recall_0.40 all 0.0079
iprec_at_recall_0.50 all 0.0079
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      0.1711
P_10           all      0.1408
P_15           all      0.1333
P_20           all      0.1204
P_30           all      0.1079
P_100          all      0.0324

```

5. trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_30.txt

```

C:\Windows\System32\cmd.exe

C:\Users\Stella\Documents\InfRetrieval\trec_eval>trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_30.txt
1 [main] trec_eval 7320 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      246
map            all      0.0321
gm_map         all      0.0043
Rprec          all      0.0799
bpref          all      0.0918
recip_rank     all      0.3066
iprec_at_recall_0.00 all 0.3338
iprec_at_recall_0.10 all 0.0995
iprec_at_recall_0.20 all 0.0432
iprec_at_recall_0.30 all 0.0196
iprec_at_recall_0.40 all 0.0079
iprec_at_recall_0.50 all 0.0079
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      0.1711
P_10           all      0.1408
P_15           all      0.1333
P_20           all      0.1204
P_30           all      0.1079
P_100          all      0.0324

```

6. trec_eval.exe -m all_trec CISI.REL resultsCISIPhase4_50.txt

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
WordEmbeddingsSimilarity	0,0184	0,0250	0,0246	0,0283
PretrainedEmbeddingsWiki	0,1711	0,1408	0,1333	0,1204



Παρατηρούμε ότι με τα WordEmbeddings προκύπτουν πολύ χειρότερα αποτελέσματα από τους υπόλοιπους τρόπους ομοιότητας. Οι κλάσεις που χρησιμοποιήσαμε στη Φάση 4 έχουν παρόμοια αποτελέσματα με τον SVD που χρησιμοποιήσαμε στην Φάση 2.

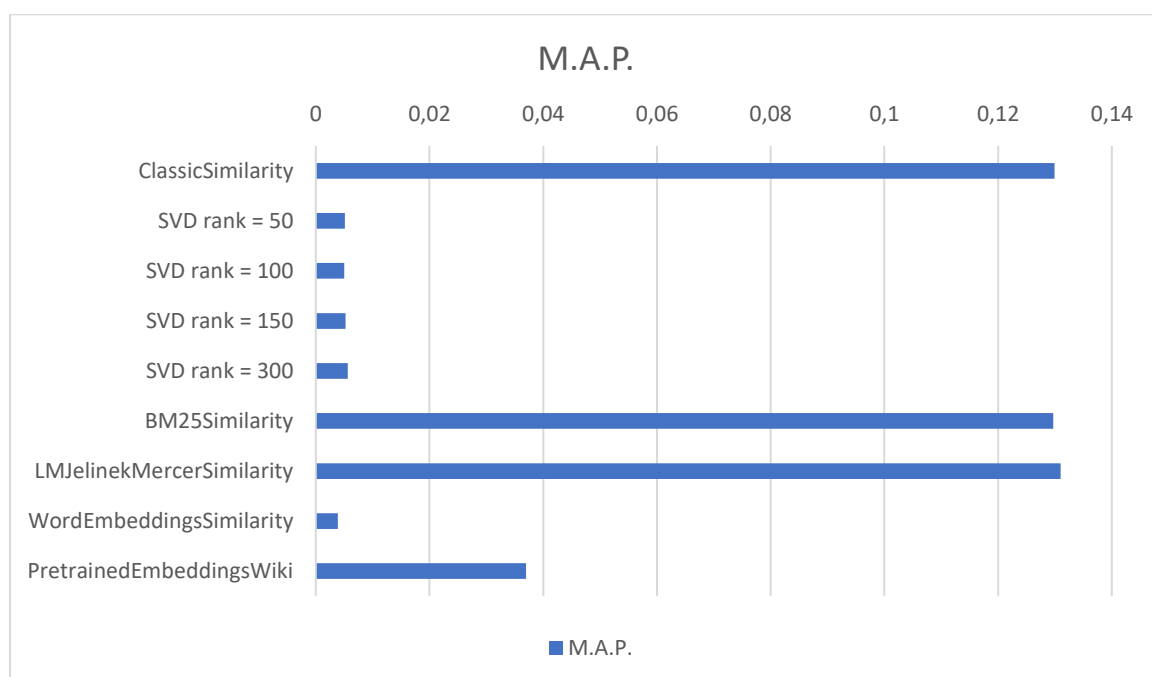
Πιστεύουμε ότι αυτό οφείλεται στην έλλειψη αρκετών κειμένων για την καλή εκπαίδευση του μοντέλου Word2Vec. Αυτό επιβεβαιώνεται και από την απόδοση του προεκπαιδευμένου μοντέλου το οποίο έχει εκπαιδευτεί σε πολύ περισσότερα δεδομένα

και επιτυγχάνει καλύτερο σκορ. Ούτε η δική του επίδοση είναι πλήρως ικανοποιητική καθώς έχει εκπαιδευτεί σε κείμενα πολύ γενικότερης φύσεως από αυτά που χρησιμοποιούμε. Ίσως θα ήταν πιο αποδοτικό να εκπαιδευτεί περαιτέρω το μοντέλο στα δικά μας κείμενα με την τεχνική του ***fine-tuning***.

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
WordEmbeddingsSimilarity	0,0039
PretrainedEmbeddingsWiki	0,0370



Πηγές:

<https://deeplearning4j.konduit.ai/language-processing/word2vec>

<https://stackoverflow.com/questions/2638200/how-to-get-a-token-from-a-lucene-tokenstream>

και λοιπές πηγές από την περιγραφή της άσκησης στο e-class