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.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

1. 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 2η Φάση (ο τρόπος που χειριστήκαμε τον SVD…).

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα 1: Indexer.java

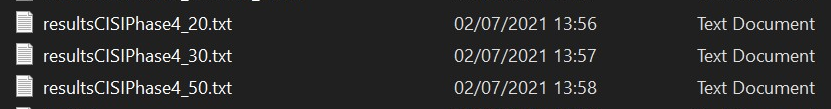
Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, ηλεκτρονικές συσκευές, υπολογιστής

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα 2: Embeddings.java

Εικόνα που περιέχει κείμενο, οθόνη, στιγμιότυπο οθόνης, ηλεκτρονικές συσκευές

Περιγραφή που δημιουργήθηκε αυτόματα Αποθηκεύουμε τα αποτελέσματα σε αρχεία .txt

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Στη συνέχεια περάσαμε τα αποτελέσματα από το Trec\_eval:

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

1. trec\_eval.exe -m all\_trec CISI.REL resultsCISIPhase4\_20.txt

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

2. trec\_eval.exe -m all\_trec CISI.REL resultsCISIPhase4\_30.txt

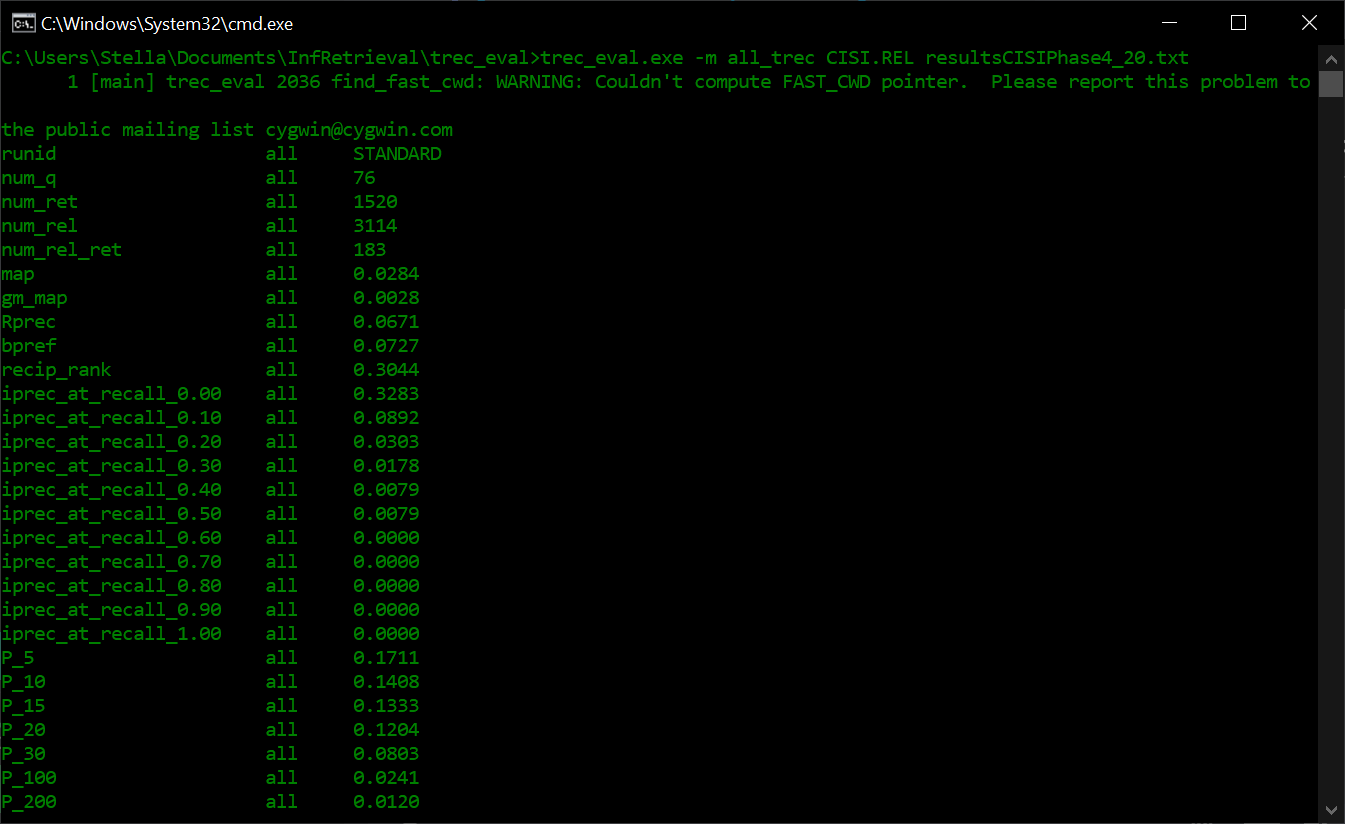
Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

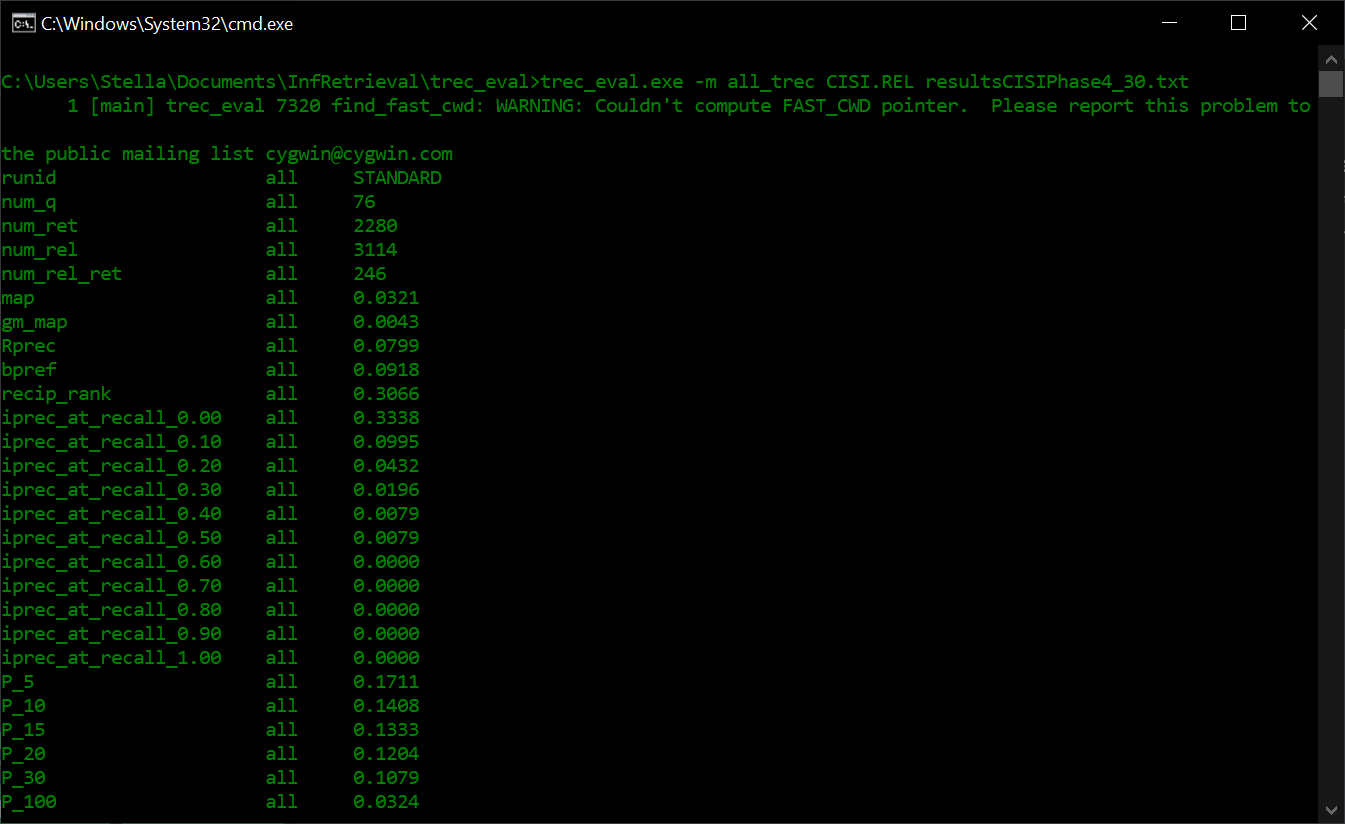
3. trec\_eval.exe -m all\_trec CISI.REL resultsCISIPhase4\_50.txt

1. 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.

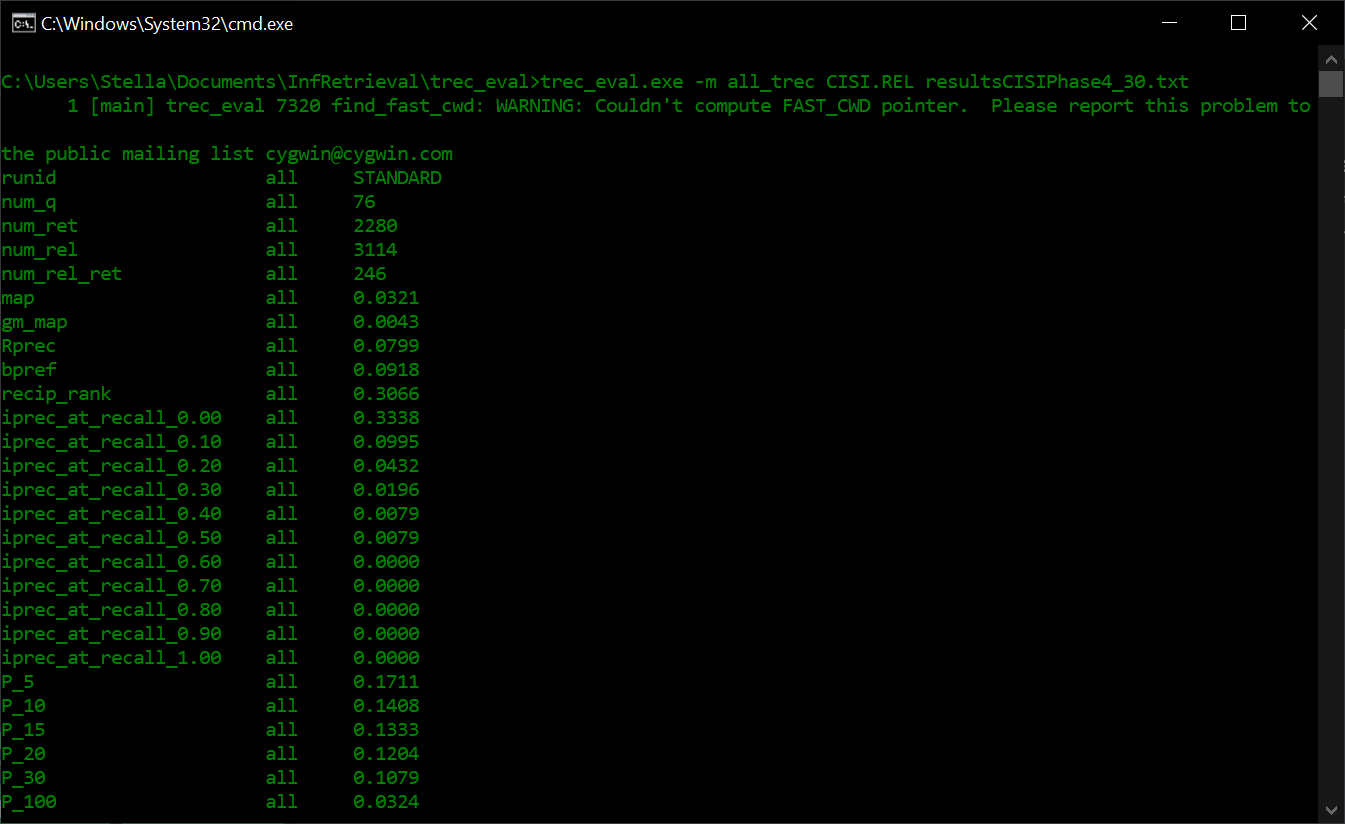
Τα αποτελέσματα του trec\_eval ήταν διαφορετικά:



4. trec\_eval.exe -m all\_trec CISI.REL resultsCISIPhase4\_20.txt



5. trec\_eval.exe -m all\_trec CISI.REL resultsCISIPhase4\_30.txt



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 Φαση 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