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

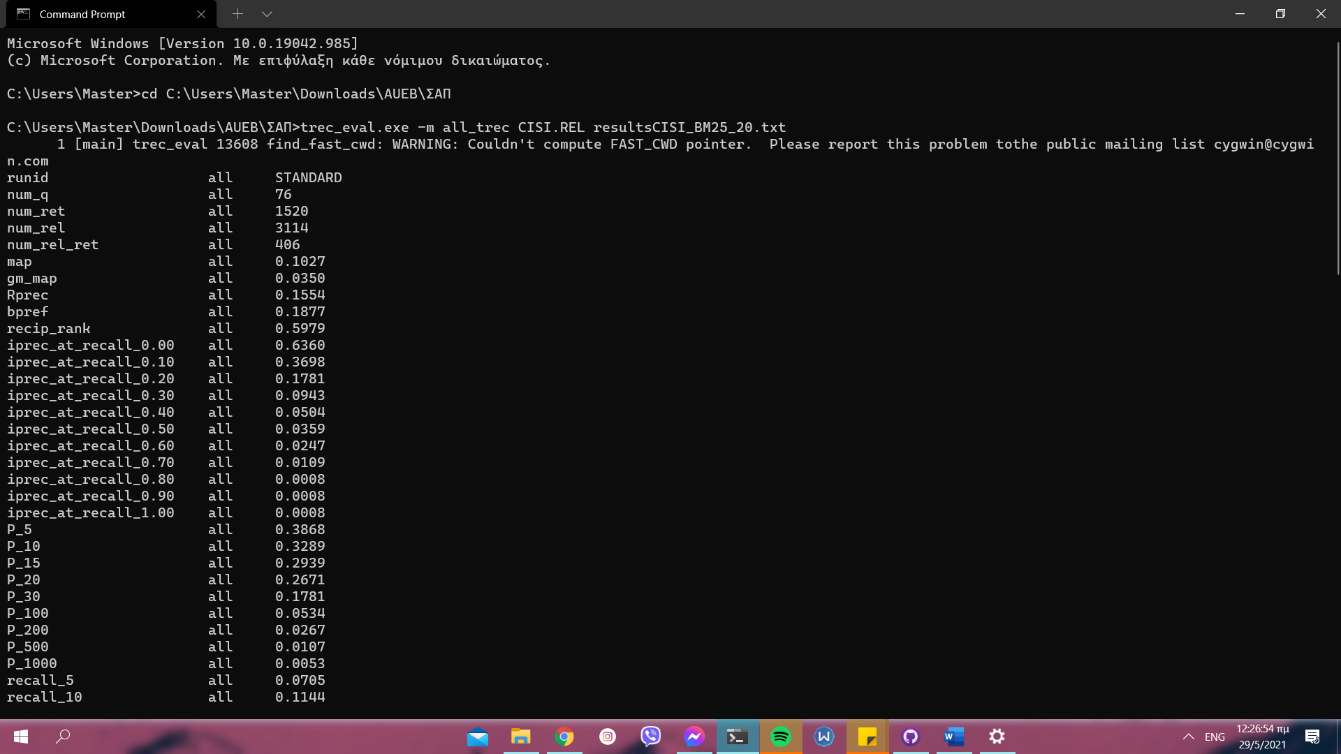
Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, ηλεκτρονικές συσκευές, υπολογιστής

Περιγραφή που δημιουργήθηκε αυτόματα

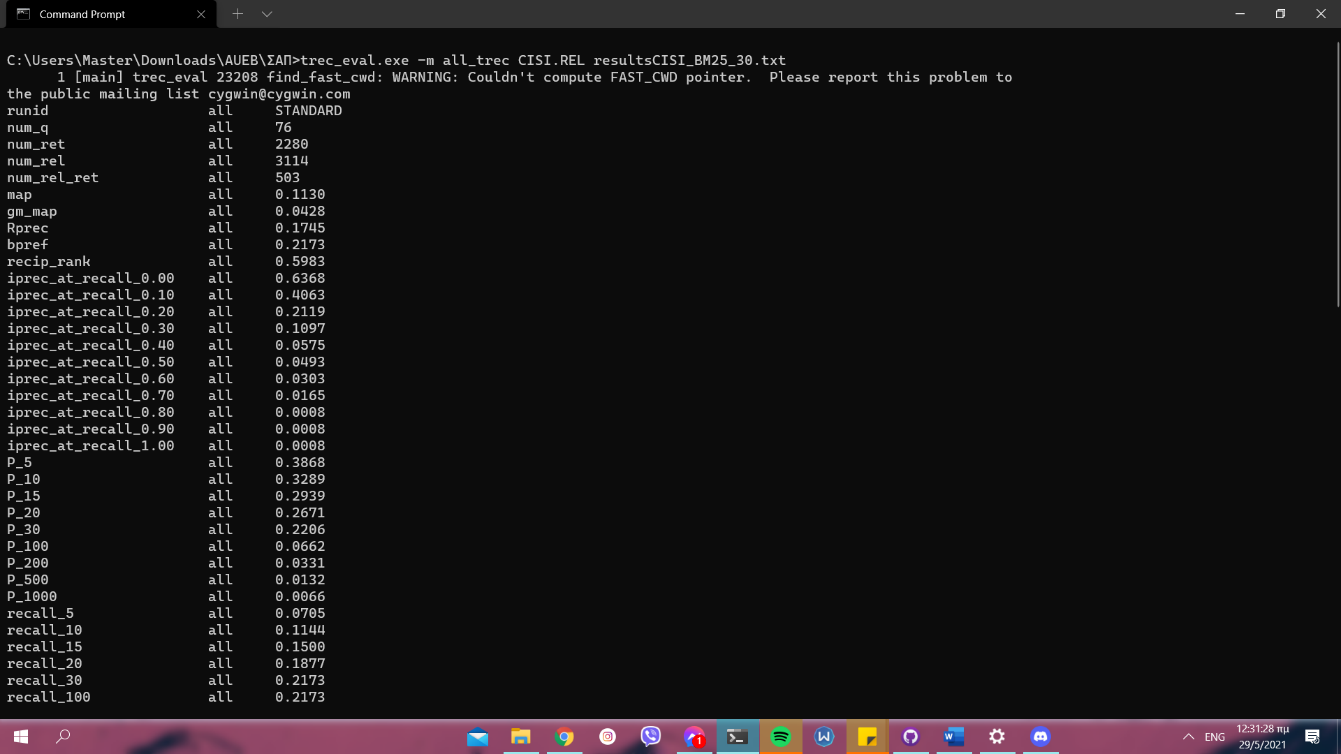
Εικόνα : Embeddings.java

Στη συνέχεια περάσαμε τα αποτελέσματα από το Trec\_eval:

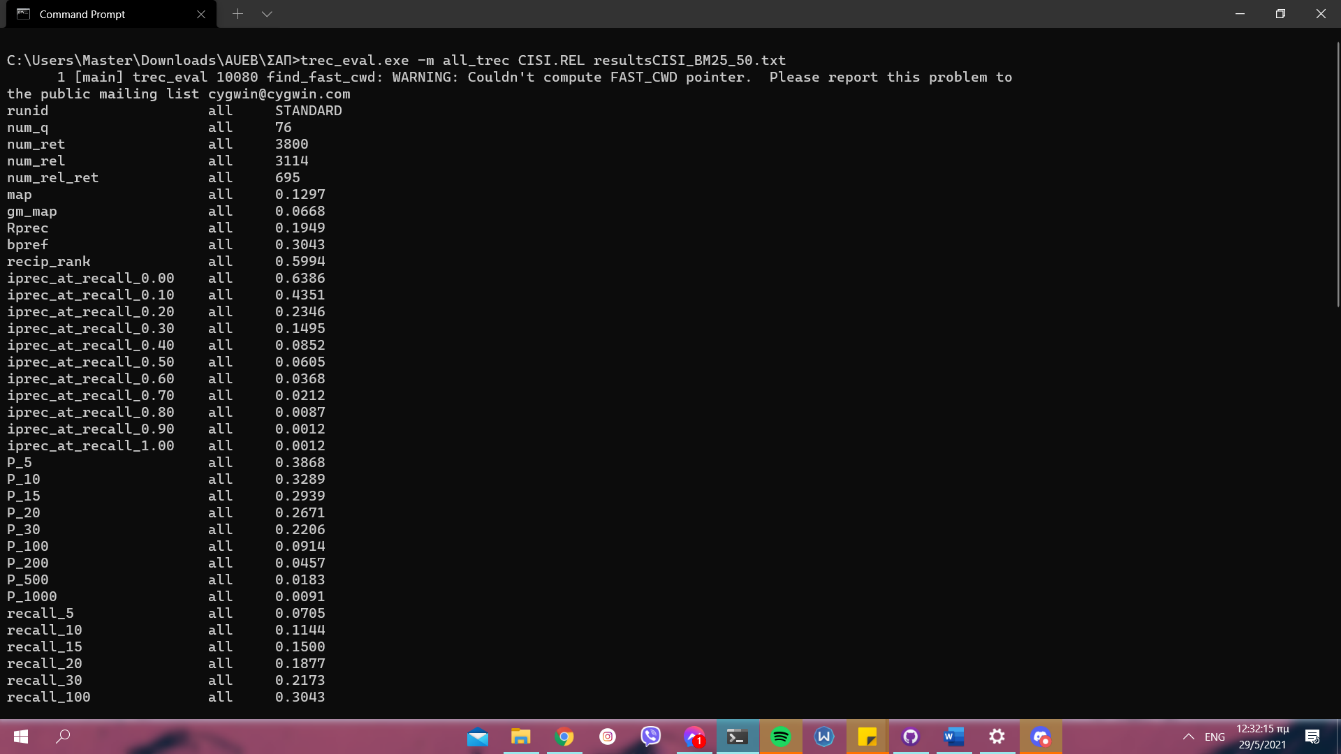
Εκτελούμε trec\_eval.exe -m all\_trec CISI.REL resultsCISI\_BM25\_20.txt στα 20 πιο σχετικά κείμενα



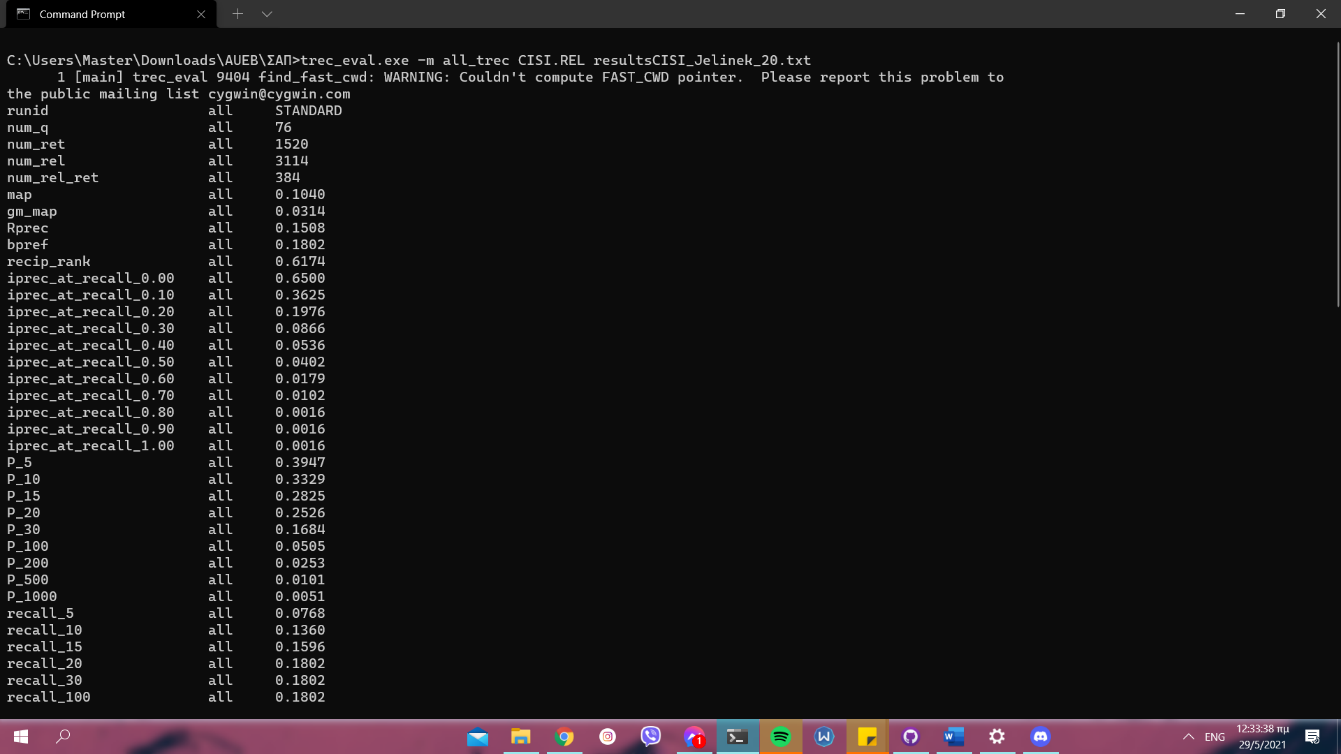
Εκτελούμε trec\_eval.exe -m all\_trec CISI.REL resultsCISI\_BM25\_30.txt στα 30 πιο σχετικά κείμενα



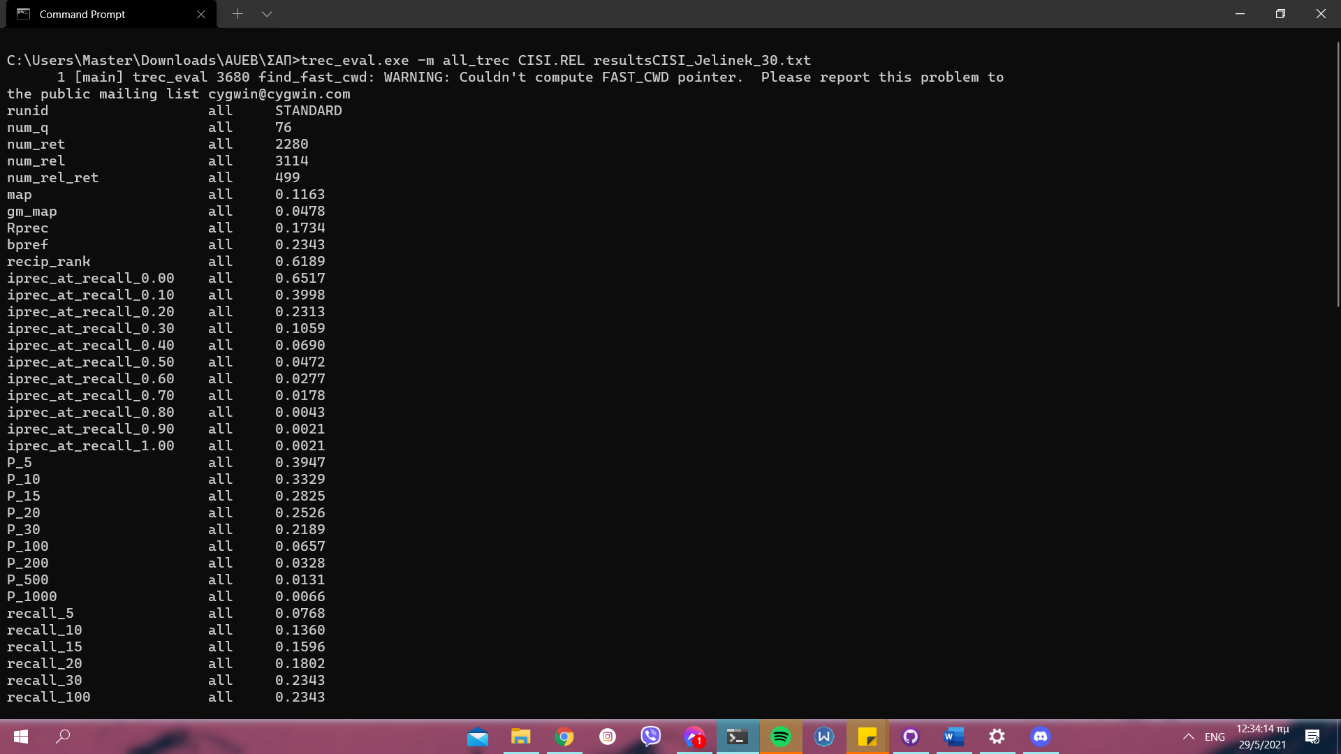
Εκτελούμε trec\_eval.exe -m all\_trec CISI.REL resultsCISI\_BM25\_50.txt στα 50 πιο σχετικά κείμενα



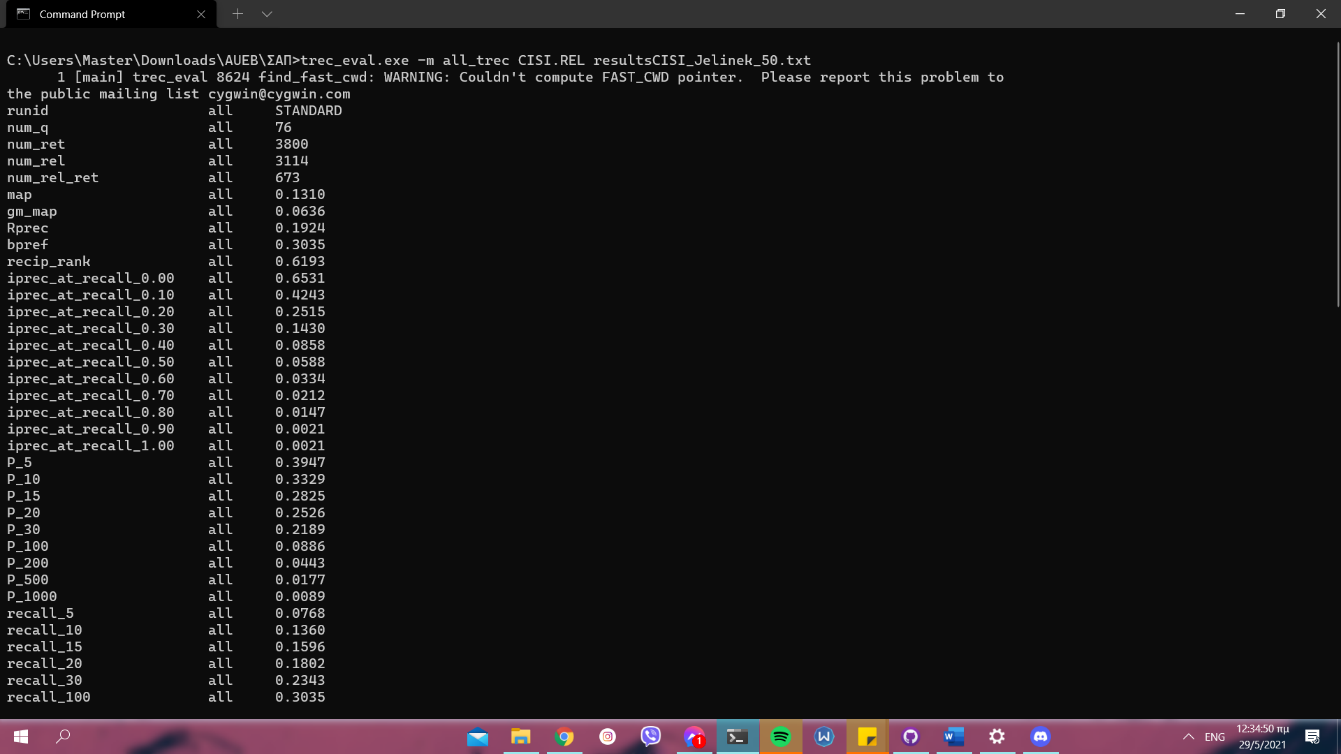
Εκτελούμε trec\_eval.exe -m all\_trec CISI.REL resultsCISI\_Jelinek\_20.txt στα 20 πιο σχετικά κείμενα



Εκτελούμε trec\_eval.exe -m all\_trec CISI.REL resultsCISI\_Jelinek\_30.txt στα 30 πιο σχετικά κείμενα



Εκτελούμε trec\_eval.exe -m all\_trec CISI.REL resultsCISI\_Jelinek\_50.txt στα 50 πιο σχετικά κείμενα



**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 Φαση 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: