Συστήματα Ανάκτησης Πληροφοριών

Φάση 5 ‐ Ανάκτηση χρησιμοποιώντας πολλαπλές συναρτήσεις ομοιότητας

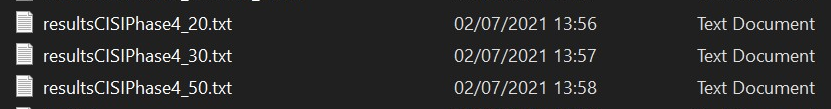
Στυλιανή Δούκα – p3170042

Βασίλειος Μπάλλας – p170115

**Περιγραφή Υλοποίησης**

Για τις ανάγκες αυτής της Φάσης της εργασίας επιστρέψαμε στο project της 3ης Φάσης και το αξιοποιήσαμε εφαρμόζοντας συνδυαστικά τις συναρτήσεις ομοιότητας που είχαμε χρησιμοποιήσει στις φάσεις 1 και 3. Παρακάτω αναλύουμε τους διάφορους συνδυασμούς που δοκιμάστηκαν.

Αποθηκεύουμε τα αποτελέσματα σε αρχεία .txt



Στη συνέχεια περάσαμε τα αποτελέσματα από το Trec\_eval:

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

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

3. 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 |
| Classic+BM25 |  |  |  |  |
| Classic+LMJelinekMercer |  |  |  |  |
| BM25+LMJelinekMercer |  |  |  |  |
| WV+Classic |  |  |  |  |
| WV+BM25 |  |  |  |  |
| WV+LMJelinekMercer |  |  |  |  |

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 |
| Classic+BM25 |  |
| Classic+LMJelinekMercer |  |
| BM25+LMJelinekMercer |  |
| WV+Classic |  |
| WV+BM25 |  |
| WV+LMJelinekMercer |  |

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