Ανάκτηση Πληροφοριών και Εξόρυξη Δεδομένων
Εργασία 1

Ομάδα 3
Γιαλαμής Αναγνώστης (58006) – Γουγούσης Παναγιώτης (58198)

Α’ μέρος

Για το πρώτο μέρος δημιουργείται ένα ευρετήριο (index) από τα κείμενα που βρίσκονται στους τέσσερις φακέλους fbis, fr94, ft, latimes. Στην συνέχεια, εκτελούνται τα 150 ερωτήματα, χωρισμένα σε πενηντάδες, στα αρχεία μορφής .trecX που δόθηκαν, όπου X είναι ο αριθμός 6, 7 ή 8. Η αναζήτηση γίνεται με τρεις διαφορετικούς τρόπους: πρώτα βασιζόμενη μόνο στον τίτλο του εγγράφου, στην συνέχεια στον τίτλο και το αφήγημα και τέλος στον τίτλο, το αφήγημα και την περιγραφή του.

```
runid   all     DPH
num_q   all     50
num_ret all     50000
num_rel all     4611
num_rel_ret     all     2214
map     all     0.2150
gm_map  all     0.1238
Rprec   all     0.2561
bpref   all     0.2345
recip_rank      all     0.7033
iprec_at_recall_0.00    all     0.7536
iprec_at_recall_0.10    all     0.4841
iprec_at_recall_0.20    all     0.3851
iprec_at_recall_0.30    all     0.2869
iprec_at_recall_0.40    all     0.2431
iprec_at_recall_0.50    all     0.1904
iprec_at_recall_0.60    all     0.1439
iprec_at_recall_0.70    all     0.0896
iprec_at_recall_0.80    all     0.0364
iprec_at_recall_0.90    all     0.0218
iprec_at_recall_1.00    all     0.0192
P_5     all     0.5080
P_10    all     0.4060
P_15    all     0.3640
P_20    all     0.3370
P_30    all     0.2947
P_100   all     0.1722
P_200   all     0.1191
P_500   all     0.0703
P_1000  all     0.0443
```

Ακολούθως, αξιολογήθηκαν τα αποτελέσματα με βάση τις ακόλουθες μετρικές που εξήγαγε το Terrier μέσω του evaluation. Η μετρική Precision@10 δηλώνει, από τα πρώτα 10 έγγραφα που ανακτήθηκαν, πόσα από αυτά είναι σχετικά με το ερώτημα. Η μετρική R-precisionrecision για ένα ερώτημα δείχνει το μέρος των σχετικών με αυτό εγγράφων, το οποίο ανακτάται από την αναζήτηση. Τέλος, η Average Precision δείχνει την μέση πιθανότητα να ανακτηθεί σχετικό έγγραφο με ένα ερώτημα.

Στο evaluation, χρησιμοποιούνται τα .qrelX αρχεία, τα οποία αντιστοιχούν σε κάποιο αρχείο .trecX, και παράγονται αρχεία μορφής .eval που περιέχουν τις μετρικές για την αξιολόγηση της αποτελεσματικότητας της ανάκτησης. Υπολογίζεται κάθε μετρική ξεχωριστά ανά ερωτήματα και, στο τέλος, ανά αρχείο υπολογίζεται η μέση τιμή καθεμιάς από αυτές. Σύμφωνα με την μορφή του αρχείου .eval, όπως φαίνεται στην διπλανή εικόνα, η τιμή P_10 είναι η Precision@10, η Rprec είναι η R-precision και η map είναι η Average Precision.

Τα αποτελέσματα που ελήφθησαν από την διαδικασία του evaluation συνοψίζονται στον παρακάτω πίνακα:

Titles				
	Q 301-350	Q 351-400	Q 401-450	Αριθμητικός Μέσος
Precision@10	0.404	0.444	0.47	0.439333
R-precision	0.2641	0.2323	0.299	0.265133
Average precision	0.2223	0.1842	0.2482	0.218233
Titles+ Description				
	Q 301-350	Q 351-400	Q 401-450	Αριθμητικός Μέσος
Precision@10	0.406	0.476	0.496	0.459333
R-precision	0.2561	0.262	0.3117	0.2766
Average precision	0.215	0.2096	0.2604	0.228333
Titles + Narrative + Description				
	Q 301-350	Q 351-400	Q 401-450	Αριθμητικός Μέσος
Precision@10	0.434	0.452	0.462	0.449333
R-precision	0.2463	0.2578	0.2875	0.263867
Average precision	0.2059	0.2143	0.2435	0.221233

Από τα παραπάνω αποτελέσματα παρατηρείται ένα πολύ καλό precision@10, δηλαδή περίπου 4 με 5 στα 10 πρώτα έγγραφα που ανακτώνται είναι σχετικά με το ερώτημα. Όμως, στην κάθε αναζήτηση εμφανίζεται περίπου το 26-28% των σχετικών εγγράφων για κάθε περίπτωση, σύμφωνα με το R-precision, και υπάρχει πιθανότητα περίπου 22% να βρεθεί σχετικό λήμμα, σύμφωνα με το Average Precision.

Παρατηρείται, επίσης, ότι κατά μέσο όρο τα καλύτερα αποτελέσματα τα δίνει η αναζήτηση με βάση μόνο τον τίτλο και το περιγραφή. Αν γίνει η αναζήτηση αποκλειστικά με βάση τον τίτλο, επειδή περιέχει περιορισμένο λεξιλόγιο, δεν μπορεί να παραγάγει το καλύτερο δυνατό αποτέλεσμα για την αναζήτηση. Για την αναζήτηση βασιζόμενη σε τίτλο, αφήγημα και περιγραφή, οι όροι για τους οποίους γίνεται η αναζήτηση είναι πολλοί, με αποτέλεσμα να εμφανίζονται πολλά άσχετα με το ερώτημα αποτελέσματα. Συνεπώς, τις καλύτερες επιδόσεις έχει η αναζήτηση με βάση τον τίτλο και το αφήγημα, καθώς περιέχει λεξιλόγιο πλούσιο αρκετά για την ανάκτηση των περισσότερων δυνατών σχετικών με το ερώτημα αποτελεσμάτων, χωρίς να γίνεται υπερτροφοδότηση του συστήματος και δεύτερη καλύτερη μέθοδος αναζήτησης είναι η αναζήτηση που βασίζεται σε όλα τα στοιχεία του ερωτήματος.

Μέρος Β

Προκειμένου να βελτιωθούν τα παραπάνω ποσοστά, στο επόμενο μέρος θα ακολουθηθεί η επέκταση ερωτήματος με συνώνυμα μέσω του WordNet. Υπάρχει βιβλιοθήκη για το WordNet μέσα στο πακέτο nltk της Python. Η διαδικασία χωρίζεται σε δύο στάδια, την εύρεση των συνωνύμων και το φιλτράρισμα μέσω του TSN (Term Semantic Network).

Για το πρώτο στάδιο, το αρχικό ερώτημα χωρίζεται στις επιμέρους λέξεις του. Για κάθε λέξη, ακολουθείται η παρακάτω διαδικασία. Πρώτα ελέγχεται αν ανήκει στις stopwords, οι οποίες ανακτώνται από το αρχείο τους, και, μετά, για όσες δεν ανήκουν σε αυτήν την κατηγορία, βρίσκονται τα συνώνυμά τους και προστίθενται στο τελικό ερώτημα.

Στο επόμενο στάδιο, φιλτράρεται το εκτεταμένο ερώτημα (expanded query), μέσω του Term Semantic Network. Από τα συνώνυμα δημιουργείται ένας γράφος με βάρη. Για κάθε δυάδα όρων από το κείμενο του αιτήματος υπολογίζεται η εννοιολογική εγγύτητά τους (similarity) μέσω του

συντομότερου μονοπατιού, ανηγμένο σε μία κλίμακα από το 0 ως το 1. Αν είναι μεγαλύτερη από το 0, δημιουργείται ένας κλάδος με βάρος ίδιο με το similarity, ο οποίος συνδέει τους δύο κόμβους (τα δύο συνώνυμα). Στην συνέχεια, τους ταξινομεί με βάση τον αριθμό των γειτόνων και επιστρέφει τους πέντε με τους περισσότερους γείτονες.

Τέλος, οι φιλτραρισμένοι όροι που προκύπτουν προστίθενται στο αρχικό ερώτημα και αφαιρούνται οι stopwords. Έτσι, προκύπτει το τελικό, εμπλουτισμένο ερώτημα. Τα αποτελέσματα είναι τα παρακάτω:

Titles				
	Q 301-350	Q 351-400	Q 401-450	Αριθμητικός Μέσος
Precision@10	0.314	0.316	0.3551	0.328367
R-precision	0.2139	0.1858	0.2281	0.209267
Average precision	0.1785	0.1361	0.1799	0.164833
Titles+ Description				
	Q 301-350	Q 351-400	Q 401-450	Αριθμητικός Μέσος
Precision@10	0.334	0.376	0.3939	0.367967
R-precision	0.2003	0.2136	0.2459	0.219933
Average precision	0.1633	0.1613	0.198	0.1742
Titles + Narrative + Description				
	Q 301-350	Q 351-400	Q 401-450	Αριθμητικός Μέσος
Precision@10	0.34	0.402	0.4163	0.3861
R-precision	0.1971	0.216	0.2346	0.2159
Average precision	0.1623	0.1616	0.1799	0.167933

Αυτό που παρατηρείται είναι ότι ο εμπλουτισμός των ερωτημάτων με συνώνυμα, έχει χειρότερη απόδοση σε σχέση με τα αρχικά. Τα συνώνυμα που βρίσκονται και προστίθενται στο ερώτημα, προσθέτουν αποτελέσματα μη σχετικά με αυτό με τα χειρότερα αποτελέσματα να παρουσιάζονται στην περίπτωση αναζήτησης με βάση τον τίτλο, λόγω του περιορισμένου λεξιλογίου. Στις άλλες δύο περιπτώσεις, δίνονται περισσότερα σχετικά αποτελέσματα με καλύτερη την αναζήτηση που βασίζεται σε όλα τα στοιχεία του ερωτήματος.

Πιο συγκεκριμένα, στην περίπτωση που η αναζήτηση γίνεται με βάση όλα τα στοιχεία του ερωτήματος είναι πιο πιθανό να βρεθούν σχετικά λήμματα στα πρώτα 10 έγγραφα, ενώ στην περίπτωση που γίνεται η αναζήτηση με βάση τον τίτλο και την περιγραφή, εμφανίζεται μεγαλύτερο μέρος των σχετικών εγγράφων με το ερώτημα και είναι πιο πιθανό να βρεθεί σχετικό έγγραφο. Αυτό το φαινόμενο παρουσιάζεται εξαιτίας της υπερτροφοδότησης που συμβαίνει στην πρώτη περίπτωση, η οποία ενώ είναι πιο πιθανό να εμφανίσει σχετικό αποτέλεσμα στα πρώτα 10 έγγραφα, περιορίζει τον αριθμό των σχετικών εγγράφων που θα ανακτηθούν.