Ανάκτηση Πληροφοριών και Εξόρυξη Δεδομένων
1η Εργασία
Ανάκτηση Κειμένων, Φράσεις και Συνώνυμα

Βογιατζή Αμαλία, 56542
Γκαντίδης Χρήστος, 56483

GitHub Project Link: https://github.com/christosg88/IR

Μέρος Α

	MAP	*R-precision*	*Precision @ 10*
`titles`	`0.2217`	`0.2691`	`0.4273`
`titles-desc`	`0.2295`	`0.2767`	`0.4347`
`titles-desc-narr`	`0.2241`	`0.2706`	`0.4360`

Βλέπουμε ότι όταν χρησιμοποιούμε το ***`titles-desc`***, το ***MAP*** είναι το μέγιστο, κάτι που δηλώνει ότι τα σχετικά topics εμφανίζονται σχετικά στην αρχή της λίστας των αποτελεσμάτων.

Το ***`R-precision`*** είναι πάλι μέγιστο άρα εμφανίζουμε και το μέγιστο αριθμό από σχετικά topics στα πρώτα k topics, όπου k ο αριθμός των σχετικών topics.

Το ***`Precision @ 10`*** είναι λίγο χειρότερο από ότι στο ***`titles-desc-narr`*** αλλά δεν μας πειράζει ιδιαίτερα γιατί υποθέτουμε ότι ο χρήστης είναι διατεθειμένος να δει μερικά άσχετα topics στην πρώτη σελίδα αλλά σε αντάλλαγμα να βρει περισσότερα σχετικά topics συνολικά.

Μέρος Β

	MAP	*R-precision*	*Precision @ 10*
titles	0.2217	0.2691	0.4273
titles-desc	0.2295	0.2767	0.4347
titles-desc-narr	0.2241	0.2706	0.4360
titles.syn	0.2143	0.2627	0.4080
titles-desc.syn	0.2249	0.2703	0.4260
titles-desc-narr.syn	0.2187	0.2655	0.4220
titles.lem	0.1715	0.2186	0.3153
titles-desc.lem	0.1499	0.1998	0.3007
titles-desc-narr.lem	0.1180	0.1709	0.2600
titles.lem.cat	0.1742	0.2229	0.3233
titles-desc.lem.cat	0.1767	0.2279	0.3453
titles-desc-narr.lem.cat	0.1550	0.2100	0.3313
titles.sim.syn	0.2050	0.2475	0.3785
titles-desc.sim.syn	0.2171	0.2581	0.4133
titles-desc-narr.sim.syn	0.2061	0.2543	0.3920

Legend:

- ***titles***: Μόνο οι τίτλοι των topics χρησιμοποιήθηκαν για το query.
- ***titles-desc***: Χρησιμοποιήθηκαν οι τίτλοι και οι περιγραφές για το query.
- ***titles-desc-narr***: Χρησιμοποιήθηκαν οι τίτλοι οι περιγραφές και οι αφηγήσεις για το query.
- ***.syn***: Χρησιμοποιήθηκαν όλα τα συνώνυμα της κάθε λέξης, προσέχοντας την κατηγορία της λέξης (ρήμα/ουσιαστικό/επίθετο κτλ) (synonyms).
- ***.lem***: Χρησιμοποιήθηκαν τα δύο πρώτα λήμματα της κάθε λέξης (lemma).
- ***.lem.cat***: Χρησιμοποιήθηκαν τα δύο πρώτα λήμματα κάθε λέξης, προσέχοντας την κατηγορία της λέξης (ρήμα/ουσιαστικό/επίθετο κτλ) (lemma category).
- ***.sim.syn***: Χρησιμοποιήθηκε το συνώνυμο της κάθε λέξης, που είχε την μεγαλύτερη σημασιολογική ομοιότητα με την αρχική λέξη (similar synonyms).

Στις περιπτώσεις που χρησιμοποιήθηκαν συνώνυμα που παρήγαγε το wordent, τα τοποθετούμε μαζί με την αρχική λέξη μέσα σε αγκύλες, κάτι που στο Indri Query Language δηλώνει πως οι λέξεις είναι συνώνυμες.

Για να βρούμε το συνώνυμο με την μεγαλύτερη σημασιολογική ομοιότητα με την αρχική λέξη, χρησιμοποιούμε την μέθοδο ***`path similarity`*** του wordnet.

Στις περιπτώσεις που παίρνουμε όλα τα συνώνυμα και χρησιμοποιούμε μαζί τα ***`titles-desc`*** ή το ***`titles-desc-narr`***, οι μετρικές χαλάνε γιατί το μεγάλο πλήθος λέξεων έχει ως αποτέλεσμα να τραβάμε πολλά topics που δεν είναι σχετικά.

Αντίθετα, στις περιπτώσεις που διαλέγουμε το πιο κοντινό στην αρχική λέξη συνώνυμο, το ***`title-desc`*** πετυχαίνει το καλύτερο αποτέλεσμα.

Συμπεράσματα και Επιπλέον Βελτιώσεις

Αν και η χρήση του wordnet για την εύρεση συνωνύμων δεν οδήγησε σε καλύτερες ανακτήσεις, μια επιπλέον βελτίωση θα μπορούσε να είναι η εύρεση συνωνύμων για λέξεις που δεν εμφανίζονται σε «μεγάλο» πλήθος topics. Η εύρεση του πλήθους των topics που χαρακτηρίζεται «μεγάλο» αποτελεί αντικείμενο έρευνας και πειραματισμού.

Αναφορές

Indri
https://www.lemurproject.org/indri/

Indri Query Language
https://www.lemurproject.org/lemur/IndriQueryLanguage.php

Wordnet
https://wordnet.princeton.edu/

NLTK
http://www.nltk.org/howto/wordnet.html