Ανάκτηση Πληροφοριών και Εξόρυξη Δεδομένων

1η Εργασία

Ανάκτηση Κειμένων, Φράσεις και Συνώνυμα

Βογιατζή Αμαλία, 56542

Γκαντίδης Χρήστος, 56483

Μέρος Α

	MAP	*R-precision*	*Precision @ 10*
Titles	0.2217	0.2691	0.4273
Titles-Description	0.2295	0.2767	0.4347
Titles-Description-Narrative	0.2241	0.2706	0.4360

Βλέπουμε ότι όταν χρησιμοποιούμε το title μαζί με το description, 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, description, narrative 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-Description	0.2295	0.2767	0.4347
Titles-Description-Narrative	0.2241	0.2706	0.4360
Titles with Synonyms	0.2143	0.2627	0.4080
Titles-Description with Synonyms	0.1874	0.2370	0.3893
Titles-Description-Narrative with Synonyms	0.1808	0.2344	0.3847
Titles with Similar Synonyms	0.2050	0.2475	0.3785
Titles-Description with Similar Synonyms	0.2171	0.2581	0.4133
Titles-Description-Narrative with Similar Synonyms	0.2061	0.2543	0.3920

Στις περιπτώσεις with Synonyms, παίρνουμε όλα τα συνώνυμα που μας επιστρέφει το wordnet για κάθε λέξη στο query και τα τοποθετούμε μαζί με την αρχική λέξη μέσα σε αγκύλες, κάτι που στο Indri Query Language δηλώνει πως οι λέξεις είναι συνώνυμες. Στις περιπτώσεις with Similar Synonyms, παίρνουμε από τα συνώνυμα που μας επιστρέφει το wordnet αυτό που έχει το μεγαλύτερο path similarity με την αρχική λέξη.

Στις περιπτώσεις που παίρνουμε όλα τα συνώνυμα και χρησιμοποιούμε μαζί με τα titles, το description ή το description και το narrative, 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 μαζί με το description πετυχαίνουν 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