ΜΥΕ003-Ανάκτηση Πληροφορίας

Αρχικός Σχεδιασμός

Γιαννικόπουλος Χαράλαμπος | Α.Μ. : 2417
Νικόλαος-Ορέστης Ντάλλας | Α.Μ. : 2507

Θα παρουσιαστεί ένα σύντομο σχέδιο του συστήματος αναζήτησης σε άρθρα της Wikipedia.

Συλλογή:
Για τη δημιουργία του corpus.json(που αποτελεί τη συλλογή μας) φτιάξαμε ένα πρόγραμμα web scraping σε python(scraping.py). Εκεί υλοποιείται επαναληπτικά ένα http get αίτημα στο *https://en.wikipedia.org/wiki/Special:Random*, το οποίο επιστρέφει τυχαίο άρθρο από το σύνολο του αρχείου της Wikipedia σε μορφή hmtl. Για κάθε άρθρο, χρησιμοποιούμε τη βιβλιοθήκη bs4 και συγκεκριμένα το BeautifulSoup για την εξαγωγή του ωφέλιμου κειμένου μέσα από τα html elements. Δημιουργούμε αντικείμενο τύπου dictionary με πεδία “link, title, text” που αντιστοιχούν στο link, τον τίτλο και το κυρίως κείμενο του άρθρου, το οποίο προστίθεται σε μια λίστα. Τέλος, αφού συλλέξουμε όλα τα άρθρα στην παραπάνω μορφή, αποθηκεύουμε τη λίστα στο αρχείο *corpus.json.*

Η συλλογή περιέχει 5000 διαφορετικά άρθρα. Επίσης κατά την δημιουργία της, απορρίψαμε άρθρα της μορφής '*Category*' π.χ: https://en.wikipedia.org/wiki/Category:Places τα οποία δεν έχουν κείμενο χρήσιμο για τη συλλογή μας.

Governorates of Yemen	Main Street After Dark
John Mullen (baseball executive)	Julia London
Brzeziny County	Hemelum
Competency-based performance management	The Time of the Hero
Hemiphyllodactylus dushanensis	Billy Adams (rockabilly musician)
Vibe Tribe	Karate at the 2006 Asian Games 2013 Men's kumite +80 kg
Battle of Suez	Dares Phrygius
Credlin (surname)	Wei Commandery
Vampire ground finch	Antetonitrus
Tri-State Peak	Medulla oblongata

Στον παραπάνω πίνακα φαίνονται 20 τυχαία άρθρα (μόνο ο τίτλος του) από την συλλογή μας.

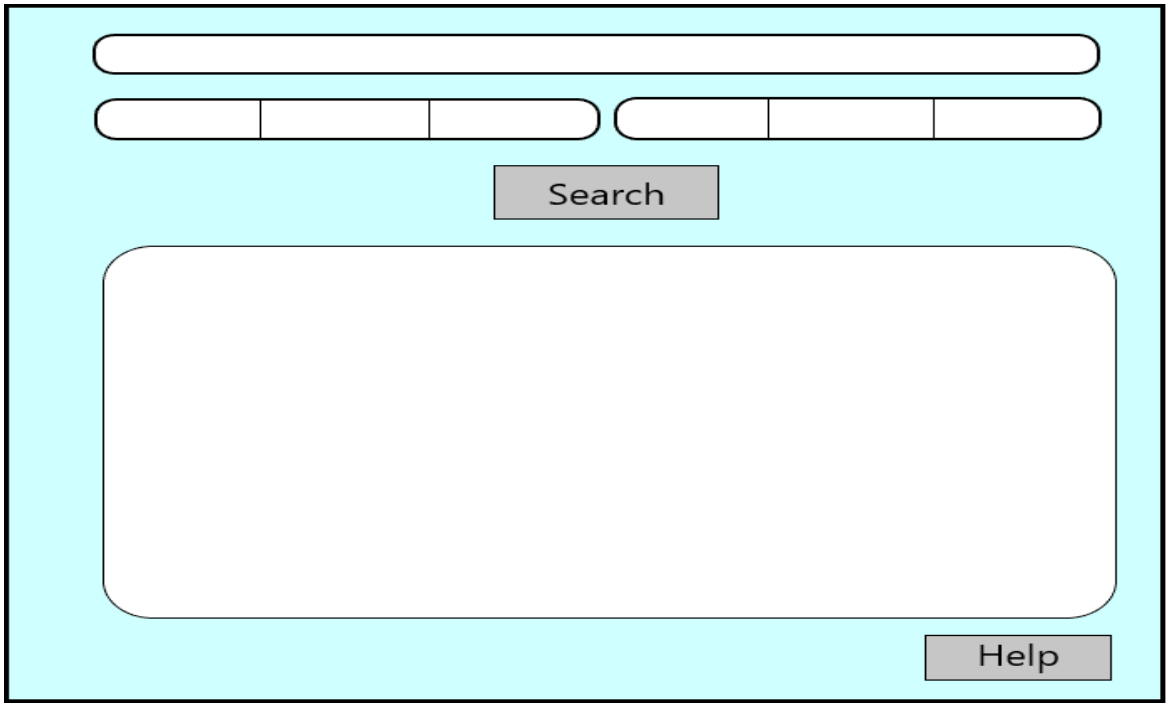
Ανάλυση κειμένου και κατασκευή ευρετηρίου:
Όσον αφορά την προεπεξεργασία, τα δεδομένα του αρχείου *corpus.json* θα τα τροφοδοτήσουμε στον analyzer που παρέχει η Lucene προκειμένου να γίνουν διαδικασίες της μορφής stemming, απαλοιφή stop words, επέκταση συνωνύμων, κλπ. Μετά την ανάλυση θα δημιουργήσουμε τα documents με βάση τη πληροφορία του analyzer και του αρχείου *corpus.json,* τα οποία θα τα βάλουμε στα ευρετήρια. Θα έχουμε δύο ευρετήρια, όπου το ένα θα αφορά το πεδίο *title* και το άλλο το πεδίο *text.* Κάθε document που θα υπάρχει στα ευρετήρια, θα αφορά ένα συγκεκριμένο άρθρο. Το document θα περιέχει επίσης και τα εξής πεδία: *link, title, text.*

Αναζήτηση:
Στη μηχανή αναζήτησης θα παρέχεται εκ των προτέρων η δυνατότητα για αναζήτηση άρθρων με τη χρήση λέξεων κλειδιών.
Εν συνεχεία, θα υποστηρίζεται αναζήτηση στο πεδίο τίτλος του κάθε άρθρου.
Επιπλέον, θα μπορεί ο χρήστης να βλέπει ένα μικρό ιστορικό αναζητήσεων, το οποίο θα αφορά πρόσφατα ερωτήματα που τέθηκαν από τον χρήστη κατά την διάρκεια της δικιάς του συνεδρίας.
Τέλος, θα ενσωματώσουμε την χρήση embedding με σκοπό την βελτίωση της αναζήτησης.

Διάταξη και παρουσίαση των αποτελεσμάτων:
Από προεπιλογή, το πρόγραμμα θα εμφανίζει τα αποτελέσματα σύμφωνα με τη συνάφειά τους με το ερώτημα.
Επιπλέον, θα παρέχεται η δυνατότητα για αλφαβητική ταξινόμηση ως προς τον τίτλο.
Η διάταξη, τέλος, θα μπορεί να γίνεται και βάσει του μεγέθους του πεδίου *text* των συναφών εγγράφων σε αύξουσα σειρά.

Γραφικό περιβάλλον:
Για την κατασκευή του γραφικού περιβάλλοντος, με το οποίο θα επικοινωνεί ο χρήστης, θα χρησιμοποιηθεί η βιβλιοθήκη swing της Java.

Παρακάτω ακολουθεί ένα πρωτότυπο του περιβάλλοντος αυτού.



Στο πρώτο πεδίο ο χρήστης θα γράφει τις λέξεις κλειδιά που θέλει να αναζητήσει.
Η πρώτη τριπλέτα κουμπιών αφορά τους διαφορετικούς τρόπους αναζήτησης ενώ η δεύτερη αφορά τους τρόπους παρουσίασης των αποτελεσμάτων. Αρχικά, ο χρήστης θα επιλέγει τον τρόπο αναζήτησης που θέλει και στη συνέχεια το πως θα διατάσσονται τα πιθανά αποτελέσματα.
Κάτω από αυτές τις επιλογές, θα υπάρχει το κουμπί *"Search"*. Πατώντας το, ο χρήστης θα βλέπει τα αποτελέσματα στο μεγάλο λευκό πλαίσιο.
Κάτω δεξιά θα υπάρχει το κουμπί *"Help"*, μέσα από το οποίο θα ενημερώνεται για το πως μπορεί να αλληλεπιδράσει με τη μηχανή αναζήτησης.
Τέλος, εφόσον έχουν προηγηθεί άλλα ερωτήματα από τον ίδιο χρήστη, κάποια από αυτά θα εμφανίζονται προκειμένου να τα χρησιμοποιήσει σε επόμενο ερώτημα.