Εργασία: Μηχανή αναζήτησης άρθρων σχετικών με τον COVID-19

Ομάδα
Ιωάννα Κουγιά 2731
Ιωάννης Πρίφτη 3321

Φάση1: Αρχικός σχεδιασμός και συλλογή δεδομένων

1. Περιγραφή της συλλογής και του τρόπου συλλογής των δεδομένων

Ξεκινήσαμε επιλέγοντας μια έτοιμη συλλογή από το kaggle. Λέγεται “COVID19 Tweets” και περιέχει τα tweets με hashtag #covid19. Από αυτήν επιλέξαμε να χρησιμοποιήσουμε για την εργασία τις 3000 πρώτες εγγραφές.

2. Συνοπτική περιγραφή του σχεδιασμού του συστήματος

Η εργασία αφορά τον σχεδιασμό και την υλοποίηση μιας μηχανής αναζήτησης tweets σχετικών με τον covid-19, χρησιμοποιώντας την βιβλιοθήκη lucene.

2.1. Το πρώτο στάδιο, το οποίο έχει υλοποιηθεί, είναι η συλλογή των εγγράφων (documents). Τα έγγραφα μας βρίσκονται σε ένα αρχείο csv. Σε κάθε γραμμή βρίσκεται ένα διαφορετικό tweet, το οποίο αποτελεί την μονάδα εγγράφου, και σε κάθε στήλη ένα διαφορετικό πεδίο (πχ. Όνομα λογαριασμού, ημερομηνία, ακόλουθοι του λογαριασμού κ.α).

2.2. Το δεύτερο στάδιο, είναι η ανάλυση κειμένου και η κατασκευή ευρετηρίου. Σε αυτό το στάδιο θα χρησιμοποιήσουμε και τη lucene.

> **2.2.1.**Το πρώτο βήμα για την κατασκευή ευρετηρίου είναι η συλλογή δεδομένων η οποία έχει ήδη γίνει.
>
> **2.2.2.** Το δεύτερο βήμα είναι η δημιουργία των documents. Τα documents θα έχουν τα εξής πεδία: username (όνομα χρήστη), location (τοποθεσία χρήστη), followers (ακόλουθοι χρήστη), friends (φίλοι χρήστη), favorites, date (ημερομηνία), text (το κείμενο της

δημοσίευσης), link, και hashtags. Στη συνέχεια θα αποφασίσουμε ποια από αυτά τα πεδία θα είναι indexed ή stored ή και τα δύο.

2.2.3. Για την ανάλυση των εγγράφων σε tokens θα χρησιμοποιήσουμε τον StandardAnalyzer, που είναι ο πιο εξελιγμένος analyzer. Έπειτα για να δημιουργήσουμε το ευρετήριο θα κάνουμε χρήση της μεθόδου IndexWritter. Το ευρετήριο θα είναι πάνω στο πεδίο text.

2.3. Το επόμενο στάδιο είναι η αναζήτηση με λέξεις κλειδιά. Στον χρήστη θα παρέχεται μια σελίδα, με ένα textbox αναζήτησης στο οποίο θα πληκτρολογεί λέξεις κλειδιά. Κάτω από το textbox θα υπάρχει η δυνατότητα χρήσης κάποιων φίλτρων όπως πχ. αναζήτηση με βάση συγκεκριμένα πεδία(όνομα λογαριασμού, αγαπημένος χρήστης, τοποθεσία, ημερομηνία) ή κάποιου είδους ταξινόμηση/ομαδοποίηση πχ χρονική. Για την επεξεργασία του ερωτήματος του χρήστη θα χρησιμοποιήσουμε την queryParser που θα “σπάει”την ερώτηση σε όρους και θα λειτουργεί όπως ο analyzer που χρησιμοποιήσαμε. Η υλοποίηση της αναζήτησης θα γίνει μέσω του IndexSearcher πάνω στο ευρετήριο που δημιουργήσαμε. Επίσης, κάθε ερώτηση που κάνει ο χρήστης, θα διατηρείται σε κάποια δομή δεδομένων, ώστε να χρησιμοποιηθεί αργότερα για προτάσεις εναλλακτικών ερωτημάτων. Και τέλος θα χρησιμοποιήσουμε embeddings για βελτίωση της αναζήτησης.

2.4. Το τέταρτο στάδιο αφορά την παρουσίαση των αποτελεσμάτων.
Στον χρήστη θα εμφανίζονται τα 10 πιο συναφή αποτελέσματα αναζήτησης, στο γραφικό περιβάλλον κάτω από την ερώτηση που υπέβαλε. Θα δίνεται στον χρήστη η δυνατότητα να επιλέγει κάποιου είδους ταξινόμηση /ομαδοποίηση μέσω κατάλληλου πεδίου όπως πχ με το πεδίο date. Επιπλέον στο κάτω μέρος της σελίδας, θα υπάρχει ένα βελάκι προς την επόμενη σελίδα με τα επόμενα 10 λιγότερα συναφή αποτελέσματα. Όσον αφορά τα αποτελέσματα, θα εμφανίζεται ένα μέρος του tweet που περιέχει τις λέξεις κλειδιά της ερώτησης. Τέλος, αφού κάνει κλικ σε κάποιο από τα αποτελέσματα θα γίνεται ανακατεύθυνση στον φιλομετρητή όπου θα φαίνεται το αντίστοιχο post του twitter.