Εισαγωγή

Στις 9 Ιανουαρίου 2020 οι υγειονομικές αρχές της Κίνας ανακοινώνουν ένα νέο στέλεχος κορωνοϊού (2019-nCoV). Μέχρι σήμερα, έχουν καταγραφεί εκατόν σαράντα εκατομμύρια κρούσματα και τρία εκατομμύρια θάνατοι παγκοσμίως. Ο κορωνοϊός SARS-CoV-2 (COVID-19) έχει πλέον γίνει η καθημερινή μας έγνοια με χιλιάδες ειδήσεις, ρεπορτάζ και άρθρα. Ο όγκος της πληροφορίας που έχει δημιουργηθεί για τον ιό, είναι αρκετά μεγάλος. Είναι σημαντικό λοιπόν η ανάκτηση μιας τέτοιας πληροφορίας να γίνεται έμμεσα και άμεσα χωρίς να εμπλέκεται η κακοπροαίρετη διάθεση τρίτων με ψευδή στοιχεία και ανακοινώσεις. Για αυτό τον λόγο η εργασία του μαθήματος *Ανάκτηση Πληροφορίας* του Πανεπιστημίου Ιωαννίνων για το ακαδημαϊκό έτος 2020-2021, θα ασχοληθεί με τη δημιουργία μίας μηχανής αναζήτησης που θα ανακτά και θα παρουσιάζει άρθρα που αφορούν τον ιό. Στη παρούσα αναφορά θα αναφερθούμε στη συλλογή των δεδομένων μας αλλά και τον σχεδιασμό του συστήματος που θα απαρτίζει τη μηχανή αναζήτησης μας, με όνομα DogeDogeGo.

1 Συλλογή Δεδομένων

Στη συλλογή μας υπάρχουν πεντακόσια άρθρα που αφορούν την πανδημία της Covid-19 τα οποία εξάγαμε από το web. Αναλυτικότερα δημιουργήσαμε ένα scrapper ο οποίος συλλέγει συνδέσμους από ιστοσελίδες με άρθρα που αφορούν την Covid-19 χρησιμοποιώντας την αναζήτηση της σελίδας Wikipedia. Στη συνέχεια, ο scrapper μας εξάγει τον τίτλο κάθε άρθρου από το σύνδεσμο και χρησιμοποιεί το API της Wikipedia κάνοντας request για να μας επιστρέψει τον τίτλο, το κείμενο και το σύνδεσμο του άρθρου. Τέλος, δημιουργούμε ένα αρχείο τύπου JSON το οποίο αποθηκεύει κάθε άρθρο σε ένα αντικείμενο με δομή:

```
    {"url": url, "title": title, "text": text}
```

Το αρχείο που παράγει ο scrapper βρίσκεται στο repository της εργασίας μας στο Github με όνομα data.json. Το αρχείο αυτό είναι το corpus το οποίο θα επεξεργαστεί η Lucene για να δημιουργήσουμε την μηχανή αναζήτησης της εργασίας.

2 Σχεδιασμός Συστήματος

Στόχος της μηχανής αναζήτησης είναι η ανάκτηση άρθρων που αφορούν την Covid-19 με λέξεις κλειδιά καθώς επίσης και αναζήτηση σε πεδίο του *Document*. Το Document είναι η μονάδα ευρετηριοποίησης και αναζήτησης. Το Document είναι ένα σύνολο πεδίων. Κάθε πεδίο έχει ένα όνομα και μια τιμή κειμένου. Ένα πεδίο μπορεί να αποθηκευτεί με το Document και επιστρέφεται με αναζητήσεις στο έγγραφο. Συνεπώς, κάθε έγγραφο πρέπει τυπικά να περιέχει ένα ή περισσότερα αποθηκευμένα πεδία που ταυτίζονται με μοναδικό τρόπο.

2.1 Index

Στην ενότητα 1 περιγράψαμε τη δομή της συλλογής μας αποθηκεύοντας τα πεδία τα οποία θα εισαχθούν στο index της Lucene. Το ευρετήριο αποθηκεύει στατιστικά στοιχεία σχετικά με τους όρους προκειμένου να κάνει την αναζήτηση με βάση όρους πιο αποτελεσματική. Η Lucene δημιουργεί ένα inverted index. Αυτό συμβαίνει επειδή μπορεί να παραθέσει, για έναν όρο, τα έγγραφα που το περιέχουν. Το indexing της Lucene αφορά το *build* των Documents που αποτελείται από τα Fields και την εισαγωγή των Documents με την χρήση του IndexWriter. Στη συνέχεια ο IndexWriter χρησιμοποιεί έναν Analyzer ο οποίος μετατρέπει το κείμενο σε μικρότερες και ακριβείς ενότητες. Το κείμενο περνά από διάφορες λειτουργίες εξαγωγής λέξεων-κλειδιών, αφαίρεσης κοινών λέξεων και σημείων στίξης, αλλαγής λέξεων σε πεζά. Εμείς θα χρησιμοποιήσουμε τον StandardAnalyzer ο οποίος αναλύει με βάση την γραμματική και αφαιρεί stop words. Επίσης, μπορεί να αναγνωρίσει συνδέσμους και emails το οποίο μας είναι χρήσιμο καθώς στη συλλογή μας κρατάμε πληροφορία για το σύνδεσμο του άρθρου.

2.2 Αναζήτηση

Η αναζήτηση στη Lucene πραγματοποιείται μέσω της κλάσης QueryParser για να δημιουργηθεί ένα Query και στη συνέχεια, χρησιμοποιώντας την κλάση IndexSearcher εκτελεί το query στο ευρετήριο και λαμβάνουμε τα αποτελέσματα. Το σύστημα μας θα υποστηρίζει TermQuery και PhraseQuery στον αρχικό μας σχεδιασμό.

2.3 Παρουσίαση Αποτελεσμάτων

Στόχος μας είναι να δημιουργήσουμε μία web εφαρμογή χρησιμοποιώντας τη Lucene στο backend και στο frontend θα σχεδιάσουμε μία διεπαφή παρόμοια με κάποιες από τις γνωστές μηχανές αναζήτησης (για παράδειγμα DuckDuckGo). Πιο συγκεκριμένα, η αρχική σελίδα της εφαρμογής θα έχει ένα κουτί για το κείμενο της αναζήτησης και μια δεύτερη σελίδα όπου θα παρουσιάζονται τα αποτελέσματα ταξινομημένα ανάλογα με τη συνάφεια.