# ΜΥΕ003: Ανάκτηση Πληροφορίας
# Τελική Αναφορά Εργασίας

Καραπατής Κωνσταντίνος Α.Μ. : 2719
Παπαδόπουλος Γιώργος Α.Μ. : 2904

Ακολουθεί link με demo της εργασίας σε πραγματικό χρόνο (phase1 και phase2)
Το βίντεο έχει μικρό voice over σχολιασμό
https://drive.google.com/file/d/1rWXZmnXh2ZnSDRFVdqbiMV0V6189PsyX/view?usp=sharing

## Phase 1 :

Για τη συλλογή του συστήματος επιλέχθηκαν 622 άρθρα που αφορούν τον Covid-19.

### Προ-επεξεργασία δεδομένων

Τα άρθρα χωρίζονται σε τίτλους και το περιεχόμενο τους και αποτελούν ξεχωριστά αρχεία csv.
Για τη δημιουργία της συλλογής χρησιμοποιήσαμε ένα script σε python (scrape.py). Το script χρησιμοποιεί την βιβλιοθήκη selenium η οποία χρησιμοποιώντας τον κατάλληλο driver (στην περίπτωση μας τον chromedriver) μπορεί να κάνει scrape από σελίδες. Αντλούμε τα άρθρα μας από 2 ξεχωριστές σελίδες ώστε να έχουμε τον επαρκή αριθμό (500+), οι οποίες είναι οι εξής.
https://www.thelancet.com/coronavirus/collection?pageSize=100&startPage=&ContentItemCategory=Editorial
https://www.the-scientist.com/tag/covid-19

## Phase 2:

### Δημιουργία ευρετηρίου

Η δημιουργία του ευρετηρίου υλοποιείται στο αρχείο "DatasetIndexer.java". Το αρχείο αυτό περιέχει τις παρακάτω συναρτήσεις.
*createIndex()*: όταν γίνει κλήση της, αρχικά δημιουργείται ένα Directory της Lucene το οποίο θα αποθηκευτεί στο δίσκο(FSDirectory.open()) και ένας IndexWritter ο οποίος χρησιμοποιεί έναν StandarAnalyzer. Στη συνέχεια διαβάζονται επαναληπτικά τα αρχεία της συλλογής, καλώντας σε κάθε βήμα της επανάληψης τη συνάρτηση indexFile() για να προστεθεί το αρχείο στο ευρετήριο.
*indexFile()*: αρχικά διαβάζονται οι απαιτούμενες πληροφορίες από το αρχείο, χρησιμοποιώντας τον χαρακτήρα ( '/-/' ) για να γίνει το split ανάμεσα σε τίτλο και περιεχόμενο και στη συνέχεια προστίθενται ως πεδία(fields) σε μία μονάδα εγγράφου(document), η οποία και προστίθεται στο ευρετήριο. Κάθε μονάδα εγγράφου περιέχει τις πληροφορίες ενός άρθρου. Τα πεδία κάθε εγγράφου είναι:

- *articleName*: περιέχει τον τίτλο του άρθρου και δημιουργείται ευρετήριο (TextField.TYPE_STORED) για αυτό καθώς χρησιμοποιείται για αναζήτηση
- *articleContents*: περιέχει τα περιεχόμενα του άρθρου και θα δημιουργηθεί ευρετήριο (TextField.TYPE_STORED) για αυτό καθώς χρησιμοποιείται για αναζήτηση
-

### Εκτέλεση αναζήτησης

Η λειτουργία της αναζήτησης υλοποιείται στο αρχείο "SearchEngine.java". Το αρχείο αυτό περιέχει τις παρακάτω συναρτήσεις.
*createIndex()*: δημιουργεί, αν δεν υπάρχει, ήδη το ευρετήριο καλώντας τη συνάρτηση DatasetIndexer.createIndex(), ανοίγει το ευρετήριο(FSDirectory.open()) και αρχικοποιεί ένα IndexSearcher για αναζήτηση σε αυτό.

*search()*: παίρνει ως ορίσματα το ερώτημα του χρήστη και την επιλογή των πεδίων προς αναζήτηση. Αρχικά προσθέτει τους όρους της αναζήτησης σε μία λίστα προτεινόμενων όρων που χρησιμοποιείται για την επιπρόσθετη λειτουργία. Στη συνέχεια ελέγχει ποια πεδία είναι προς αναζήτηση, δημιουργεί το αντίστοιχο ερώτημα, με χρήση των κλάσεων MultiFieldQueryParser και Query της Lucene, και εκτελεί την αναζήτηση των εκατό πιο σχετικών εγγράφων στο ευρετήριο (IndexSearcher.search().scoreDocs) χρησιμοποιώντας ένα StandarAnalyzer. Δημιουργεί ένα Highlighter ώστε να γίνει επισήμανση των όρων αναζήτησης στα αποτελέσματα. Τέλος επιστρέφει τις πληροφορίες προς παρουσίαση στο χρήστη για κάθε ένα από τα δέκα πρώτα αποτελέσματα.
*getNext()*: επιστρέφει τις πληροφορίες προς παρουσίαση στο χρήστη για κάθε ένα από τα δέκα επόμενα αποτελέσματα.
*getResult()*: παράγει την πληροφορία προς εμφάνιση στο χρήστη για κάποιο αποτέλεσμα. Ελέγχει σε ποια πεδία έγινε η αναζήτηση και κρατάει για καθένα από αυτά τα "καλύτερα" τους σημεία με επισημασμένους τους όρους αναζήτησης. Η επιστρεφόμενη πληροφορία περιέχει τον τίτλο και το περιεχόμενο του άρθρου.
*getDocumentInfo()*: επιστρέφει όλες τις πληροφορίες για κάποιο άρθρο που εμφανίστηκε στο χρήστη ως αποτέλεσμα αναζήτησης, εκτελώντας επισήμανση στα περιεχόμενα των πεδίων στα οποία έγινε αναζήτηση.
*updateAutocomplete()*: επιστρέφει τη λίστα προτεινόμενων όρων που χρησιμοποιείται για την επιπρόσθετη λειτουργία.

Επιπρόσθετη λειτουργικότητα
Το σύστημα παρέχει στο χρήστη τη δυνατότητα αυτόματης συμπλήρωσης όρων αναζήτησης με βάση τους όρους που έχουν εισαχθεί από προηγούμενες αναζητήσεις. Για το σκοπό αυτό χρησιμοποιείται μία λίστα με τους όρους των ερωτημάτων που υποβάλλονται κάθε φόρα από το χρήστη.

Το αρχείο "LuceneConstants" περιέχει τις παρακάτω σταθερές που χρησιμοποιούνται από το υπόλοιπο σύστημα:

- *COLLECTION_PATH*: το μονοπάτι της συλλογής
- *INDEX_PATH*: το μονοπάτι του ευρετηρίου
- *TITLE*: το πεδίο για το όνομα ενός άρθρου
- *CONTENTS*: το πεδίο για τα περιεχόμενα ενός άρθρου
- *MAX_SEARCH*: συνολικό μέγιστο πλήθος αποτελεσμάτων της αναζήτησης
- *MAX_PROJECT*: μέγιστο πλήθος αποτελεσμάτων προς εμφάνιση στο χρήστη κάθε φορά

Η διεπαφή χρήστη αποτελείται από δύο παράθυρα. Το βασικό παράθυρο υλοποιείται στο αρχείο "MainSearchWindow" και το παράθυρο με τις αναλυτικές πληροφορίες ενός άρθρου στο αρχείο "DocumentInfoWindow".
Στο βασικό παράθυρο αρχικά υπάρχει ένα πεδίο εισαγωγής κειμένου(search box), στο οποίο ο χρήστης εισάγει τα ερωτήματα του, ένα κουμπί 'Search', το οποίο ο χρήστης πατάει για την υποβολή των ερωτημάτων του, μία επιλογή για να εκτελέσει ο χρήστης σύνθετη αναζήτηση, δηλαδή να επιλέξει με βάση ποια πεδία θα γίνει η αναζήτηση. Όταν ο χρήστης υποβάλει ένα ερώτημα, εμφανίζονται πληροφορίες τα δέκα πρώτα αποτελέσματα-έγγραφα καθώς και ένα κουμπί 'Next' το οποίο μπορεί να πατήσει ο χρήστης για να εμφανιστούν τα επόμενα 10 αποτελέσματα. Αν ο χρήστης πατήσει στην επιλογή 'Advanced Search', εμφανίζεται ένα μενού από το οποίο ο χρήστης μπορεί να επιλέξει σε ποια πεδία να γίνει η αναζήτηση. Αν δεν επιλεχθεί κανένα πεδίο η αναζήτηση γίνεται και στα δύο πεδία. Αυτό είναι και το default είδος αναζήτησης του συστήματος. Τέλος αν ο χρήστης επιλέξει κάποιο από τα εμφανιζόμενα αποτελέσματα, τότε εμφανίζεται ένα παράθυρο που περιέχει όλες τις πληροφορίες του επιλεγμένου αποτελέσματος-άρθρου.
Στο δευτερεύον παράθυρο εμφανίζονται οι αναλυτικές πληροφορίες ενός άρθρου, με επισημασμένους τους όρους αναζήτησης στα περιεχόμενα των πεδίων στα οποία έγινε η αναζήτηση.