ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ – PROJECT 2021

ΠΡΩΤΗ ΦΑΣΗ

Το παρακάτω κείμενο επιμελήθηκαν οι:

Θωμαή Τσιαούση

Νεφέλη Ελευθερία Σέξτου

1. ΣΥΝΟΠΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΣΥΛΛΟΓΗΣ

Η συλλογή (corpus) που επιλέχθηκε να χρησιμοποιηθεί για την λειτουργία της μηχανής, είναι ένα κομμάτι μιας έτοιμης συλλογής από την ιστοσελίδα kaggle [1], και αποτελείται από data που έχουν συλλεχθεί από online πηγές.

Η αρχική συλλογή (τύπου .csv) αποτελούνταν από περίπου 350.000 αρχεία, τα οποία είχαν δημοσιευθεί από τον Ιανουάριο του 2020 μέχρι και το Δεκέμβριο του 2020. Προς το παρόν, έχει κρατηθεί ένα δείγμα 200 αρχείων, που καλύπτει και τους 12 μήνες του προηγούμενου έτους, και έχει μετατραπεί σε μορφή .txt με εφαρμογή του παρακάτω κώδικα σε Python.

```
import csv

fileName = "covid19_data"
rowNum = 200
rowCount = 0
with open('COVID_DATA.csv', encoding="utf8") as csv_file:
    csv_reader = csv.reader(csv_file)
    for row in csv_reader:
        if rowCount == 0:
            print(f'Columns in CSV file are {", ".join(row)}')
            rowCount += 1
        else:
            text_file = open(''.join([fileName, str(rowCount), '.txt']), 'w', encoding="utf8")
            text_file.write("Author: "+row[0]+'\n'+"Topic: "+row[6]+'\n'+"Date: "+row[1]+'\n'+"Title: "
                            +row[3]+'\n\n'+"Body:\n"+row[5])
            text_file.flush()
            text_file.close()
            rowCount += 1

        if rowCount > rowNum:
            print('txt files created!')
            break
    print(f'Processed {rowNum} rows.')
```

[1] Covid-19 Public Media Dataset by Anacode | Kaggle

2. ΣΥΝΟΠΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΣΧΕΔΙΑΣΜΟΥ ΜΗΧΑΝΗΣ ΑΝΑΖΗΤΗΣΗΣ

2.1 ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ ΜΗΧΑΝΗΣ ΑΝΑΖΗΤΗΣΗΣ

Στόχος είναι η δημιουργία μιας μηχανής αναζήτησης η οποία θα επιστρέφει αποτελέσματα από τη συλλογή που έχει συγκεντρωθεί, τα οποία θα πρέπει να ικανοποιούν την ανάγκη πληροφόρησης του χρήστη σχετικά με τον COVID-19.

Επίσης, η μηχανή θα υποστηρίζει διάφορα είδη ερωτημάτων του χρήστη με βάση λέξεις κλειδιά και θα επιστρέφει αποτελέσματα παρουσιασμένα με βάση τη συνάφεια τους με το ερώτημα και με συγκεκριμένο format (επιστροφή 10 αρχείων ανά σελίδα, ανάλογα με τη συχνότητα εμφάνισης του ερωτήματος του χρήστη στο κάθε αρχείο).

2.2 ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΟΥ -ΚΑΤΑΣΚΕΥΗ ΕΥΡΕΤΗΡΙΟΥ

Οι βασικές έννοιες στη Lucene είναι το Index, Documents, Fields, και Terms. Κάθε Index εμπεριέχει μια ακολουθία από Documents, κάθε Document μια ακολουθία από Fields και κάθε Field είναι μια ακολουθία από Terms. Σε κάθε πεδίο θα αντιστοιχεί ένα inverted index το οποίο θα εμπεριέχει τους αντίστοιχους όρους.

Σε αυτό το στάδιο έχουν επιλεχθεί τα πεδία fileName και fileContent. Αργότερα, από το πεδίο fileContent μπορούν να εξαχθούν τα πεδία Title, Author, Date και Body.

Για τα Body και Title αρχικά θα γίνει ανάλυση με τον StandardAnalyzer (αποτελεί τον πιο πλήρες από τους ενσωματωμένους, καθώς αναγνωρίζει συγκεκριμένα είδη tokens , μετατρέπει τη γραμματοσειρά των όρων σε lowercase, αφαιρεί stop words). Τα πεδία Author και Date δεν θα αναλυθούν καθώς θα κρατηθούν ολόκληρα ως tokens για την αναζήτηση.

Το περιεχόμενο όλων των παραπάνω πεδίων θα αποθηκευτεί καθώς επιδιώκεται η δυνατότητα παρουσίασής του στο Document που θα επιστραφεί στο χρήστη.

2.3 ΑΝΑΖΗΤΗΣΗ

Ο χρήστης θα εισάγει το ερώτημά του στο User Interface.

Το ερώτημα αυτό θα αναλυθεί αντίστοιχα με τα πεδία με βάση τα οποία υλοποιείται η αναζήτηση και έτσι προκύπτουν οι λέξεις κλειδιά, βάσει των οποίων θα γίνεται η αναζήτηση στα Documents.

Επιπλέον, στόχος είναι το σύστημα να υποστηρίζει τις παρακάτω μορφές ερωτημάτων:

- αναζήτηση με λέξη κλειδί σε όλα τα πεδία (Author, Date, Title και Body που ανήκουν στο πεδίο Content του file)
- αναζήτηση με λέξη κλειδί σε συγκεκριμένο πεδίο (πχ. Στον τίτλο ή κάποια συγκεκριμένη ημερομηνία)
- αναζήτηση Boolean

Επίσης επιδιώκεται να διατηρείται το ιστορικό των αναζητήσεων του χρήστη και να προτείνονται εναλλακτικά ερωτήματα παρόμοια με αυτά που έχει ψάξει ή της μορφής «did you mean …».

Τέλος, θα χρησιμοποιηθούν embeddings ώστε η αναζήτηση να γίνεται με τον βέλτιστο δυνατό τρόπο που μπορεί να επιτευχθεί.

2.4 ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Σκοπός είναι η εφαρμογή να επικοινωνεί με το χρήστη μέσω ενός απλού GUI το οποίο:

- θα έχει μια μπάρα αναζήτησης όπου ο χρήστης θα μπορεί να εισάγει το ερώτημά του το οποίο θα καταχωρείται στο ΣΑΠ με το πάτημα ενός «κουμπιού»
- θα εμφανίζει τα αποτελέσματα ανά 10 (top 10 results από την δομή TopDocs) και θα παρέχεται η δυνατότητα να προχωρήσει στα επόμενα
- θα εμφανίζει στα αποτελέσματα τις λέξεις κλειδιά τονισμένες (highlighted)
- θα παρέχεται η δυνατότητα ομαδοποίησης των αποτελεσμάτων βάσει των πεδίων, όπως το Date και Author

Σημείωση: Παραπάνω παρουσιάζεται μια αρχική εκτίμηση του σχεδιασμού της μηχανής αναζήτησης που πρέπει να κατασκευαστεί. Ωστόσο, κατά τη υλοποίηση του, μπορεί να υποστεί αλλαγές, οι οποίες θα αποσκοπούν στην καλύτερη και ορθότερη λειτουργία της μηχανής.