Εργασία: Μηχανή αναζήτησης άρθρων σχετικών με τον COVID-19

Φάση 1: Αρχικός σχεδιασμός και συλλογή δεδομένων

Καταληκτική Ημερομηνία Παράδοσης

Δευτέρα 19 Απριλίου 2021, 5μμ

1. Στόχοι αυτής της φάσης: (1) η δημιουργία της συλλογής δεδομένων και (2) η κατανόηση των βασικών βημάτων της εργασίας και ο αρχικός σχεδιασμός της.

2. Για την εργασία θα παραδώστε link σε ένα **GitHub repository**.

Δημιουργείστε το repository ως PRIVATE και δώστε μου προσπέλαση.

Στα Settings του repository επιλέξτε Manage Access και κάντε με collaborator.

GitHub username: pitoura

3. Περιεχόμενο του GitHub για την Φάση 1.

(α) **Readme** με τις παρακάτω πληροφορίες (500-1000 λέξεις):

- Σύντομη περιγραφή της συλλογής και του τρόπου που συλλέξατε τα δεδομένα.

- Συνοπτική περιγραφή του σχεδιασμό του συστήματός σας. Ο σχεδιασμός μπορεί να αλλάξει κατά την υλοποίηση. Παρακάτω, είναι μια προτεινόμενη δομή:

> *Γενική περιγραφή* της μηχανής αναζήτησης που θα κατασκευάσετε: Στόχος, λειτουργικότητα
>
> *Ανάλυση κειμένου και κατασκευή ευρετηρίου:* Ποια θα είναι η προεπεξεργασία των άρθρων για τη δημιουργία του εγγράφου, ποια είναι η μονάδα εγγράφου και τα αντίστοιχα πεδία (fields), τα ευρετήρια που σκοπεύετε να δημιουργήσετε (σε ποια πεδία, και τι είδους), κλπ, ώστε να υποστηρίζονται οι διάφοροι τρόποι αναζήτησης
>
> *Αναζήτηση:* Πως θα γίνετε η αναζήτηση των εγγράφων, τα είδη των ερωτημάτων
>
> *Παρουσίαση Αποτελεσμάτων:* Πως θα παρουσιάζονται τα αποτελέσματα.

Στις απαντήσεις στα παραπάνω όπου είναι δυνατόν αναφερθείτε και στα αντίστοιχα τμήματα της Lucene, όπως Build/Analyze/Index Document IndexSearcher/QueryParser, TopDocs, ScoreDocs, κοκ

(β) Τουλάχιστον **20 από τα έγγραφα της συλλογής** (σε όποιο format θέλετε, csv, json, κοκ). Εξηγείστε στο readme το format.

4. Τρόπος παράδοσης: Υποβάλλετε στο ecourse του μαθήματος ένα αρχείο με τα ονόματα και τους ΑΜ των μελών της ομάδας και το link στο GitHub.