Εργασία: Μηχανή αναζήτησης ταινιών

Φάση 1: Αρχικός σχεδιασμός και συλλογή δεδομένων

Καταληκτική Ημερομηνία Παράδοσης: Παρασκευή 15 Απριλίου 2021, 11μμ

Στόχοι αυτής της φάσης είναι: (1) η δημιουργία της συλλογής δεδομένων, (2) η κατανόηση των βασικών βημάτων της εργασίας και (3) ένας αρχικός σχεδιασμός της.

Γενικά για την εργασία θα παραδώστε 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

Μπορείτε επίσης να έχετε το repository public.

Περιεχόμενο του GitHub repository για την Φάση 1:

(1) Readme με σύντομη περιγραφή (500-1000 λέξεις) με πληροφορίες για τα παρακάτω

(α) Ποια θα είναι η συλλογή των εγγράφων που θα χρησιμοποιείστε στην εργασία: Περιγράψτε το format των δεδομένων, την πληροφορία που θα έχουν, την πηγή από την οποία θα τα κατεβάσατε ή/και πως θα τα συλλέξετε.

(β) Σύντομη περιγραφή του σχεδιασμού του συστήματος σας (ο σχεδιασμός μπορεί να αλλάξει κατά την υλοποίηση). Ενδεικτική δομή για αυτή την περιγραφή:

Εισαγωγή: Ποιος είναι ο στόχος και η λειτουργικότητας του συστήματος.

Ανάλυση κειμένου και κατασκευή ευρετηρίου: Ενδεικτικά, 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), ώστε να υποστηρίζονται οι διάφοροι τρόποι αναζήτησης, κλπ

Αναζήτηση: Πως θα γίνετε η αναζήτηση των ταινιών, 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

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

Τρόπος παράδοσης: Υποβολή ενός αρχείο pdf μέσω του ecourse του μαθήματος. Το αρχείο θα περιέχει μόνο

1. Τα ονόματα και ΑΜ των μελών της ομάδας.
2. Το link στο GitHub repository της εργασίας σας.
3. Το Readme που θα υπάρχει και στο GitHub repository