

## Εργασία: Μηχανή αναζήτησης Wikipedia άρθρων

Σύντομη περιγραφή σχεδιασμού και συλλογής δεδομένων

Καταληκτική Ημερομηνία Παράδοσης  
Τετάρτη 29 Απριλίου 2020, 5μμ

**Τρόπος παράδοσης:** Η εργασία θα υποβληθεί στο ecourse του μαθήματος. Μια υποβολή ανά ομάδα.

**Στόχος** αυτής της φάσης είναι να σας βοηθήσει να κατανοήσετε τα βήματα που πρέπει να ακολουθήσετε.

Περιγράψτε συνοπτικά το σχεδιασμό του συστήματός σας. Ο σχεδιασμός μπορεί να αλλάξει κατά την υλοποίηση.

Η περιγραφή θα πρέπει να είναι από 600-1000 λέξεις.

### Περιεχόμενο

Συμπεριλάβετε πληροφορίες για τα παρακάτω:

Συλλογή εγγράφων (corpus). Τον τρόπο που θα συλλέξετε τα άρθρα. Θα πρέπει να έχετε ήδη συλλέξει τουλάχιστον 10 άρθρα. Πείτε ποια και πως τα συλλέξατε.

Ανάλυση κειμένου και κατασκευή ευρετηρίου. την προεπεξεργασία των άρθρων για τη δημιουργία του εγγράφου, ποια θα 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. Αν σκοπεύετε να υλοποιήσετε κάτι πέρα της ελάχιστης απαίτησης, εξηγήστε τι και πως.

Παρουσίαση Αποτελεσμάτων. Αν σκοπεύετε να υλοποιήσετε κάτι πέρα της ελάχιστης απαίτησης, εξηγήστε τι και πως.

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