**ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ**

**ΕΡΓΑΣΙΑ 2023**

**ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ ΤΡΑΓΟΥΔΙΩΝ**

Μάιος 2023

ΑΝΑΦΟΡΑ ΓΙΑ ΤΗΝ 2Η ΦΑΣΗ ΤΗΣ ΕΡΓΑΣΙΑΣ

ΟΜΑΔΑ

ΠΕΤΤΗΣ ΠΑΡΑΣΚΕΥΑΣ-ΧΡΗΣΤΟΣ 3136

ΒΑΦΕΙΑΣ ΔΗΜΗΤΡΙΟΣ 3150

Corpus

Συλλογή Εγγράφων

(2000 εγγραφές/αρχεία)

Για την συλλογή εγγράφων μας χρησιμοποιήσαμε το αρχείο από Kaggle με τραγούδια από το Spotify. Το αρχείο το πήραμε σε CSV(Comma Separated Values) μορφή με περίπου 40.000 διαφορετικές εγγραφές. Με τη βοήθεια της γλώσσας python και της βιβλιοθήκης pandas μετατρέψαμε τις πρώτες 2000 εγγραφές από το αρχείο, σε ξεχωριστά αρχεία, 1 για κάθε εγγραφή/τραγούδι. Ο λόγος που το κάναμε αυτό είναι επειδή δεν μπορέσαμε να πάρουμε τα δεδομένα κατευθείαν από το CSV αρχείο γιατί μας τα επέστρεφε όλα σε μια θέση της λίστας όταν φτιάχναμε το Index. Το πρόγραμμα με το οποίο φτιάξαμε τα αρχεία μπορεί να βρεθεί στον φάκελο του project με το όνομα <file\_scrapper.py>.

Αφού λοιπόν μαζέψαμε τα αρχεία για το ευρετήριο μας, φτιάξαμε τον λεγόμενο Indexer, ο οποίος παίρνει τα αρχεία, τα διαβάζει γραμμή-γραμμή και αναλογεί το περιεχόμενο κάθε γραμμής στο αρχείο, σε ένα πεδίο του ευρετηρίου, που θα χρησιμοποιηθεί αργότερα για την αναζήτηση.

Index

Ευρετήριο

Το ευρετήριο της εφαρμογής μας φτιάχνετε στην κλάση Indexer, του πηγαίου κώδικα, και είναι αναγκαίο να φτιαχτεί μόνο ΜΙΑ φορά. Ακολούθως αρκεί να υπάρχει ένα έγκυρο segments αρχείο στον root φάκελο, για να μπορεί να γίνει η αναζήτηση μέσα στο ευρετήριο.

Για πρώτη εγκατάσταση του ευρετηρίου, σιγουρευτείτε ότι έχετε τα αρχεία από τις εγγραφές/τραγούδια στον φάκελο song\_files του root directory, και αφού φορτώσετε την εφαρμογή επιλέξτε Index από το menu bar, και ακολούθως την επιλογή Create Index.

Αρχικά, τα διαφορετικά αρχεία στον δίσκο πρέπει να μετατραπούν σε αντικείμενα Document, τα οποία υποστηρίζει η βιβλιοθήκη και θα χρειαστούν ακολούθως για την πρόσθεση των δεδομένων στο ευρετήριο. Το ευρετήριο φτιάχνετε με τη βοήθεια της βιβλιοθήκης Lucene για την Java και έτσι κάνουμε χρήση γνωστών συναρτήσεων και αντικειμένων που προσφέρει η βιβλιοθήκη(IndexWriter, IndexWriterConfig, κ.α.), για να «γράψουμε» τα δεδομένα/αρχεία στο ευρετήριο μας.