ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΑΣ, ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ
Τμήμα Πληροφορικής και Τηλεπικοινωνιών

Μηχανή Αναζήτησης FALSE

Εργασία μαθήματος ΑΝΑΚΤΗΣΗ ΚΑΙ ΕΞΟΡΥΞΗ ΠΛΗΡΟΦΟΡΙΩΝ

Συγγραφείς:
ΓΙΑΝΝΟΠΟΥΛΟΣ ΓΕΩΡΓΙΟΣ, 2022202000039, dit20039@go.uop.gr
ΓΙΑΝΝΟΠΟΥΛΟΣ ΙΩΑΝΝΗΣ, 2022201900032, dit19032@go.uop.gr

Δεκέμβριος 2023

Table of Contents

1 Εισαγωγή

Αυτή η αναφορά συνιστά την αξιολόγηση της πρώτης εξαμηνιαίας εργασίας μας στο μάθημα της Ανάκτησης και Εξόρυξης Πληροφορίας. Κατά τη διάρκεια αυτού του εξαμήνου, είχαμε ως στόχο την ανάπτυξη μιας μηχανής αναζήτησης τραγουδιών με τη χρήση του Lucene framework. Προκειμένου να επιτευχθεί αυτός ο στόχος με μεγαλύτερη ευελιξία, αποφασίσαμε να χρησιμοποιήσουμε το Pylucene, έναν wrapper του Lucene στη γλώσσα προγραμματισμού Python.

Σε αυτήν την αναφορά, αναλύουμε τη διαδικασία εγκατάστασης του Pylucene στον υπολογιστή μας καθώς και την ανάπτυξη της μηχανής αναζήτησης τραγουδιών, παρέχοντας επίσης μια εκτενή αξιολόγηση των αποτελεσμάτων και των προκλήσεων που παρουσιάστηκαν κατά την υλοποίηση του έργου.

Η ανάπτυξη της μηχανής αναζήτησης τραγουδιών αποτέλεσε μια σημαντική ευκαιρία να εφαρμόσουμε τις γνώσεις που αποκτήσαμε κατά τη διάρκεια του μαθήματος, πειραματιζόμενοι με την ανάπτυξη λύσεων για την ανάκτηση πληροφοριών σε πραγματικά σενάρια.

2 ΕΓΚΑΤΑΣΤΑΣΗ PYLUCENE

Για την εγκατάσταση του Pylucene στον υπολογιστή μας ακολουθήσαμε βήματα από την επίσημη ιστοσελίδα Apache Pylucene . Η εγκατάσταση του έγινε σε λειτουργικό σύστημα Linux και διανομή Ubuntu 22.04 και για την έκδοση Pylucene-8.11.0 .Το αρχείο αυτό θα βρίσκετε στον φάκελο του project άλλα δεν θα ακολουθήσουμε αυτήν την λύση καθώς σε προσπάθεια μας να το εγκαταστήσουμε σε παλιότερη ή νεότερη διανομή παρουσιάστηκαν κάποια προβλήματα που είναι επιλύσημα άλλα απαιτούν παραπάνω χρόνο για την επίλυση τους. Οπότε αφού ο προηγούμενος οδηγός μπορεί να εμφανίσει θέματα από υπολογιστή σε υπολογιστή ακολουθήσαμε την λύση της χρήσης docker βασιζόμενοι στο image του coady/pylucene:8 που έχει την εγκατάσταση του Pylucene-8.11.0. Παρακάτω ακολουθούν τα βήματα για την εγκατάσταση του docker στο terminal του Ubuntu, μπορεί να παραληφθεί εάν είναι εγκαταστημένο, την εγκατάσταση του image που δημιουργήσαμε και περιέχει πέρα από το Pylucene όλα τα απαραίτητα python libraries που χρησιμοποιήθηκαν καθώς και την δημιουργία του container μέσα στο image.

2.1 ΕΓΚΑΤΑΣΤΑΣΗ DOCKER

- Αρχικά ελέγχουμε ότι το docker δεν είναι εγκατεστημένο
  - `docker --version`

```
test@test-virtual-machine:~$ docker --version
Command 'docker' not found, but can be installed with:
sudo snap install docker          # version 20.10.24, or
sudo apt  install docker.io       # version 24.0.5-0ubuntu1~22.04.1
sudo apt  install podman-docker   # version 3.4.4+ds1-1ubuntu1.22.04.2
See 'snap info docker' for additional versions.
test@test-virtual-machine:~$
```

- Update the package lists
  - `sudo apt update`

- Install necessary dependencies to add a repository over HTTPS:
  - `sudo apt install apt-transport-https ca-certificates curl software-properties-common`

- Add Docker's official GPG key:

    - `curl -fsS https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add -`
- Add the Docker repository to your system:
    - `sudo add-apt-repository "deb [arch=amd64] https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable"`

- Update package lists again to include the Docker repository:
    - `sudo apt update`

- Install Docker:
    - `sudo apt install docker-ce`

- Start the Docker service:
    - `sudo systemctl start docker`

- Verify Docker installation by checking its version:
    - `docker --version`

```
test@test-virtual-machine:~$ docker --version
Docker version 24.0.7, build afdd53b
```

```
test@test-virtual-machine:~$ sudo systemctl start docker
test@test-virtual-machine:~$ sudo systemctl status docker
● docker.service - Docker Application Container Engine
     Loaded: loaded (/lib/systemd/system/docker.service; enabled; vendor preset: enabled)
     Active: active (running) since Fri 2023-12-29 14:22:45 EET; 1min 1s ago
TriggeredBy: ● docker.socket
       Docs: https://docs.docker.com
   Main PID: 5131 (dockerd)
      Tasks: 9
     Memory: 26.1M
        CPU: 252ms
     CGroup: /system.slice/docker.service
             └─5131 /usr/bin/dockerd -H fd:// --containerd=/run/containerd/containerd.sock
```

Μπορεί να χρειαστεί να κάνετε restart τον υπολογιστή σας, ώστε να λειτουργήσει.

2.2 PULL IMAGE

Αφού έχει εγκατασταθεί το docker στον υπολογιστή μας, μπορούμε να προχωρήσουμε στο pull του image μας.

- Προβολή των images στον υπολογιστή μας
  - `docker images`

```
test@test-virtual-machine:~$ docker images
REPOSITORY    TAG       IMAGE ID    CREATED    SIZE
```

- Pull image
  - `docker pull giorgosgian/irm_project:version1`

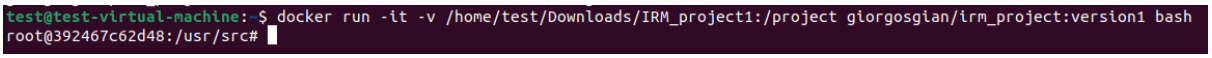
```
test@test-virtual-machine:~$ docker pull giorgosgian/irm_project:version1
version1: Pulling from giorgosgian/irm_project
6aefca2dc61d: Pull complete
967757d56527: Pull complete
c357e2c68cb3: Pull complete
c766e27afb21: Pull complete
32a180f5cf85: Pull complete
1535e3c1181a: Pull complete
ca398dbb0a27: Pull complete
fc3fb1727276: Pull complete
13ca01dc6e0b: Pull complete
92c7dc3149c2: Pull complete
4c0a669b4785: Pull complete
67df5c2ab1ba: Pull complete
020ac928aa8c: Pull complete
49e2aa17ae4f: Pull complete
51ccaa61c7ea: Pull complete
0580de134223: Pull complete
4f4fb700ef54: Pull complete
7cf067ae2dda: Pull complete
7c50f4b32d62: Pull complete
53fbd79851a1: Pull complete
Digest: sha256:2a1415908e937a094b04373cce7fabb2ed3fa4de1fd809e0ed09bd64b170d2e7
Status: Downloaded newer image for giorgosgian/irm_project:version1
docker.io/giorgosgian/irm_project:version1
```

- Προβολή ότι εγκαταστάθηκε
  - `docker images`

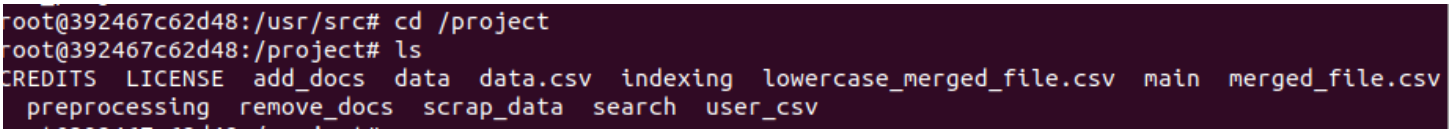
```
test@test-virtual-machine:~$ docker images
REPOSITORY                TAG       IMAGE ID       CREATED         SIZE
giorgosgian/irm_project   version1  3eb85ff5041a   45 hours ago    2.98GB
```

2.3 ΔΗΜΙΟΥΡΓΙΑ CONTAINER ΓΙΑ ΤΟ PROJECT

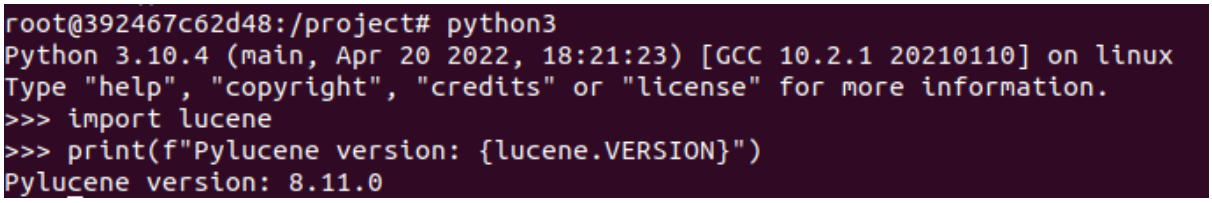
- Ανάλογα την διάδρομή που βρίσκετε ο φάκελος του project IRM_project1
  - `docker run -it -v {directory}:/project giorgosgian/irm_project:version1 bash`



- Το project μας βρίσκετε τώρα στο /project
  - `cd /project`
  - `ls`



- Προαιρετικά μπορούμε να ελέγξουμε την έκδοση του Pylucene
  - `python3`
  - `import lucene`
  - `print(f"Pylucene version: {lucene.VERSION}")`



- Για το τρέξιμο του προγράμματος αρκεί να τρέξουμε την main/main.py
  - `python3 main/main.py`

```
root@e8fafb46bb51:/project# python3 main/main.py
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
_________________Welcome to the FALSE search engine_____________________
_________Menu_____________________
Press 1 to preprocess the data
Press 2 to index the data
Press 3 to add a document
Press 4 to delete a document
Press 5 to search a document
Press 6 to scrap data directly from AZlyrics
Press 0 to exit from the FALSE search engine
```

Η λειτουργικότητα του project θα επεξηγηθεί παρακάτω.

3 ΛΕΠΤΟΜΕΡΕΙΕΣ ΥΛΟΠΟΙΗΣΗΣ

Η δομή του φακέλου του project είναι η εξής:

/IRM_project1

/main

/main.py[1]

/preprocessing

/preprocess.py

/index

/index.py

/add_docs

/add.py

/add_album.py

/remove_docs

/remove.py

/remove_album.py

/search

/search.py

/VSM_search.py

/Phrase_search.py

/scrap_data

/scrap.py

/data

/index

/original_data

/preprocessed_data

[1] Για προβολή του κώδικα μέσα από το docker αρκεί `nano όνομα_αρχείου.py`

3.1 MAIN

Μέσα στην main γίνεται import το lucene και η εκκίνηση του JAVA VM ώστε να μπορούμε να τρέξουμε το lucene μέσα από την python. Ακόμα κάνουμε import[2] τις συναρτήσεις που θα χρησιμοποιήσουμε από τους άλλους φακέλους. Με το τρέξιμο της main έχουμε και την εμφάνιση του menu όπου εκεί ο χρήστης εισάγει τις εντολές του.

3.2 PREPROCESSING

Η προ επεξεργασία των δεδομένων λαμβάνει χώρα σε αυτό το αρχείο. Για τα αρχεία songs.csv και lyrics.csv τα ενώνουμε σε ένα αρχείο csv βάση όνομα τραγουδιστή και όνομα τραγουδιού και διαγράφουμε τα διπλά κελιά. Εφαρμόζουμε πεζά σε πεδία 'singer_name', 'song_name', 'link','lyrics'. Ακόμα στο πεδίο lyrics εφαρμόζουμε stemming και αφαίρεση stopwords με την χρήση του SnowballStemmer και της βιβλιοθήκης NLTK. Δημιουργούμε ένα νέο πεδίο με όνομα stemmed_lyrics ώστε εκεί να πραγματοποιείτε η αναζήτηση έπειτα και να παρουσιάζουμε στον χρήστη τα lyrics ολόκληρα με το πεδίο lyrics.

Για το αρχείο albums.csv που δεν έχει κάποιο κοινό πεδίο με τα προηγούμενα απλά εφαρμόζουμε πεζά καθώς η αναζήτηση στα πεδία του δεν απαιτεί κάποια άλλη προεπεξεργασία.

3.3 INDEX

Σε αυτό το αρχείο βρίσκονται οι συναρτήσεις για την ευρετηρίαση των δεδομένων. Χρησιμοποιούμε τον MMapDirectory για την αποθήκευση του ευρετηρίου. Για τον ευρετηριασμό του combined csv ευρετηριάζουμε τα πεδία singer_name, song_name,song_link, lyrics, stemmed_lyrics. Για τα πεδία singer_name, song_name, stemmed_lyrics κρατάμε πληροφορία για τα έγγραφα, συχνότητα όρων και την θέση των όρων. Ενώ για τα άλλα δύο πεδία δεν απαιτείται καθώς τα θέλουμε μόνο για εμφάνιση και όχι για αναζήτηση. Η συνάρτηση που καλείται είναι η index() και η διαδρομή του είναι /data/index/main_index.

Για τον ευρετηριασμό του album.csv χρησιμοποιούμε πάλι τον MMapDirectory και το αποθηκευουμε στο /data/index/album_index .Ευρετηριάζουμε τα πεδία singer_name,name, type, year κρατώντας όλη την πληροφορία για τα έγγραφα, συχνότητα όρων και την θέση των όρων. Χρησιμοποιείτε η συνάρτηση album_index().

[2] Όλα τα μονοπάτια έχουν τροποποιηθεί ώστε να τρέχουν στο docker, για να τρέξει εκτός docker θα πρέπει να αφαιρεθεί χειροκίνητα το '/project/' από όλα τα μονοπάτια που υπάρχουν στα αρχεία. Με αυτόν τον τρόπο τα μονοπάτια θα είναι relative οπότε θα τρέχουν.

3.4 ADD DOCS

Μέσα σε αυτόν τον φάκελο εμπεριέχονται δύο αρχεία python. Το πρώτο για την προσθήκη νέων εγγραφών στο κοινό ευρετήριο τραγουδιών και στοίχων. Το δεύτερο για την προσθήκη νέων εγγραφών στο ευρετήριο των άλμπουμ.

3.4.1 ADD.py

Αρχείο για την προσθήκη νέων πεδίων στο main index. Δυνατότητα προσθήκης νέου τραγουδιού πληκτρολογώντας εσύ τα δεδομένα ή χρησιμοποιώντας ένα είδη έτοιμο αρχείο csv όπου πρέπει να έχει την σωστή δομή ώστε να ευρετηριαστεί και να ανανεωθεί με την νέα/νέες εγγραφή/ές το υπάρχων csv αρχείο.

Για την εισαγωγή από πληκτρολόγιο θα ζητηθεί από τον χρήστη να εισάγει όνομα τραγουδιστή, όνομα τραγουδιού, σύνδεσμος και οι στίχοι.Καλείτε η συνάρτηση add_document_to_index(singer_name, song_name, link, lyrics).

Για την εισαγωγή δεδομένων από έτοιμο αρχείο πρέπει να βρίσκεται μέσα στον φάκελο /user_csv/ . Ο χρήστης εισάγει το όνομα του αρχείου, δεν απαιτείται να βάλει την κατάληξη .csv η οποία θα μπει αυτόματα εάν δεν υπάρχει. Ελέγχονται τα κελιά να είναι σωστά και εφόσον είναι ξεκινάει η διαδικασία προ επεξεργασίας και ευρετηρίασης των νέων δεδομένων καθώς και την ανανέωση του csv. Για δοκιμή υπάρχει το αρχείο song_add.csv μέσα στον φάκελο user_csv. Η συνάρτηση που θα κληθεί είναι η add_csv_to_index(new_data,csv_file_path)

3.4.2 Add_album.py

Ίδια λειτουργία με το add.py άλλα για το album index. . Δυνατότητα προσθήκης νέου τραγουδιού πληκτρολογώντας εσύ τα δεδομένα ή χρησιμοποιώντας ένα είδη έτοιμο αρχείο csv όπου πρέπει να έχει την σωστή δομή ώστε να ευρετηριαστεί και να ανανεωθεί με την νέα/νέες εγγραφή/ές το υπάρχων csv αρχείο.

Για την εισαγωγή από πληκτρολόγιο θα ζητηθεί από τον χρήστη να εισάγει όνομα τραγουδιστή, όνομα αλμπουμ, τύπο άλμπουμ και χρονιά.Καλείτε η συνάρτηση add_album_document_to_index(singer_name, album_name, album_type, year).

Για την εισαγωγή δεδομένων από έτοιμο αρχείο πρέπει να βρίσκεται μέσα στον φάκελο /user_csv/ . Ο χρήστης εισάγει το όνομα του αρχείου, δεν απαιτείται να βάλει την κατάληξη .csv η οποία θα μπει αυτόματα εάν δεν υπάρχει. Ελέγχονται τα κελιά να είναι σωστά και εφόσον είναι ξεκινάει η διαδικασία προ επεξεργασίας και ευρετηρίασης των νέων δεδομένων καθώς και την ανανέωση του csv. Για δοκιμή υπάρχει το αρχείο album_add.csv μέσα στον φάκελο user_csv. Η συνάρτηση που θα κληθεί είναι η add_album_csv_to_index(new_data,csv_file_path).

3.5 REMOVE DOCS

3.5.1 Remove.py

Διαγραφή ενός τραγουδιού από το ευρετήριο και ανανέωση του αρχείου csv. Ζητείτε από τον χρήστη να εισάγει όνομα τραγουδιστή και τραγουδιού ή να εισάγει πολλούς

τραγουδιστές και τα τραγούδια τους χωρισμένα με κόμμα. Καλείτε η συνάρτηση remove_from_index(singer_names, song_names) συντάσσεται ένα combined Boolean Query για το όνομα του τραγουδιστή και όνομα τραγουδιού και γίνεται αναζήτηση του τραγουδιού. Εάν βρεθεί παρουσιάζονται τα στοιχεία του και εμφανίζεται ένα μενού όπου ο χρήστης πληκτρολογεί είτε y για να διαγραφτεί είτε n ώστε να μην διαγραφτεί. Έπειτα ακολουθείτε η ίδια διαδικασία για τα επόμενα ονόματα και τραγούδια.

3.5.2 Remove_album.py

Διαγραφή ενός άλμπουμ από το ευρετήριο και ανανέωση του αρχείου csv. Ζητείτε από τον χρήστη να εισάγει όνομα τραγουδιστή και άλμπουμ ή να εισάγει πολλούς τραγουδιστές και άλμπουμ χωρισμένα με κόμμα. Καλείτε η συνάρτηση
remove_album_from_index(singer_names, album_names)

Ακολουθείτε η ίδια διαδικασία με τον remove.py

3.6 SEARCH

3.6.1 Search.py

Σε αυτό το αρχείο είναι οι συναρτήσεις για την πραγματοποίηση Boolean ερωτημάτων. Εάν εμπεριέχεται στην αναζήτηση τα keywords AND, OR, NOT σε κάποιο πεδίο τότε πραγματοποιείτε αυτή η αναζήτηση. Ο χρήστης μπορεί να εισάγει πληροφορία σε ένα ή και παραπάνω πεδία και έπειτα του ζητείτε να εισάγει το πλήθος των απαντήσεων που θέλει να του επιστραφούν. Επιστρέφονται με σκορ σχετικότητας τα κείμενα που είναι πιο σχετικά με την αναζήτηση του χρήστη καθώς και όλη τους η πληροφορία.

3.6.2 VSM_search.py

Αρχείο με τις συναρτήσεις για vector space model ερωτήματα που είναι το default της μηχανής αναζήτησης καθώς εάν δεν εμπεριέχονται Boolean τελεστές ή ‘’ ‘’. Για τον υπολογισμό του σκορ χρησιμοποιείτε η BM25Similarity και το ερώτημα είναι ένα combined Boolean query.

3.6.3 Phrase_search.py

Αρχείο με τις συναρτήσεις για phrase queries χρησιμοποιώντας την PhraseQuery μέθοδο της lucene. Για την πραγματοποίηση αυτού του ερωτήματος πρέπει κάποιο από τα πεδία αναζήτησης να περιέχει ‘’ ‘’.

3.7 SCRAP DATA

Σε αυτό το αρχείο εμπεριέχονται οι συναρτήσεις για ανάκτηση της πληροφορίας των τραγουδιών από την AZlyrics ιστοσελίδα. Για τα τραγούδια και τους στίχους απαιτείτε να δοθεί το όνομα του τραγουδιστή και το όνομα τραγουδιού. Έπειτα συντάσσετε αυτόματα ο σύνδεσμος και με την χρήση των βιβλιοθηκών urllib3 και beautifulsoup4 ξεκινάει η ανάκτηση

της πληροφορίας εφόσον έχουν δοθεί σωστά τα δεδομένα. Τα δεδομένα αποθηκεύονται στην διαδρομή scrap_data/scraped_azlyrics.csv και έπειτα το αρχείο αυτό προστίθεται στο ευρετήριο και στο υπάρχων csv με τα υπόλοιπα δεδομένα.

Για την ανάκτηση άλμπουμ, απαιτείτε μόνο το όνομα του τραγουδιστή. Τα δεδομένα αποθηκεύονται /scrap_data/scraped_album_azlyrics.csv και έπειτα ευρετηριάζεται και ανανεώνεται το csv που περιέχει τα album.

4 ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ

Σε συνέχεια από το 2.3 κεφάλαιο έχουμε

```
georgios@giorgos:~$ docker run -it -v /home/georgios/Documents/IRM_project1:/project irm_project:version1 bash
root@d40111d706fc:/usr/src# cd /project
root@d40111d706fc:/project# python3 main/main.py
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
_________________Welcome to the FALSE search engine_____________________
_________Menu_________________
Press 1 to preprocess the data
Press 2 to index the data
Press 3 to add a document
Press 4 to delete a document
Press 5 to search a document
Press 6 to scrap data directly from AZlyrics
Press 0 to exit from the FALSE search engine
```

- Πληκτρολογούμε το 1 ώστε να αρχίσει η προ επεξεργασία των δεδομένων

```
1
Data preprocessing started
Stemming started for column lyrics....
Stemming done
Stemming took 51.46 seconds.
Data preprocessing completed
```

- Συνεχίζουμε πληκτρολογώντας 2 για ευρετηριασμό

```
2
All files inside the folder have been deleted.
Indexing songs and lyrics...
0 docs found in songs index
26459 docs found in index for merged_songs_lyrics_stemmed.csv
Album indexing...
All files inside the folder have been deleted.
0 docs found in songs index
1819 docs found in index for albums_normalized.csv
```

- Συνεχίζουμε πατώντας 3 για προσθήκη
    - Πατάμε 1 για προσθήκη τραγουδιου

```
3
Add document...
Press 1 to add a new doc in songs and lyrics
Press 2 to add a new doc in albums
Enter your input: 1
Type 1 to add your input
Type 2 to add an already csv file
Type exit in order to return to the main menu
Enter your input: 1
Enter singer name: French Montana
Enter song name: Unforgettable
Enter link: /frenchmontana/unforgettable.html
Enter lyrics: It's not good enough for me, since I been with you It's not gonna work for you
```

- Πατάμε 2 για να προσθέσουμε ένα csv file με δεδομένα τραγουδιού

```
Enter your input: 1
Type 1 to add your input
Type 2 to add an already csv file
Type exit in order to return to the main menu
Enter your input: 2
Enter the name of the CSV file: song_add
/project/user_csv/song_add.csv
Opening song_add.csv
Format is ok
Continuing with indexing and updating the csv file...
Data from /project/user_csv/song_add.csv added to the merged CSV file.
Indexing documents:
Data added to the CSV file successfully.
Document added to the existing index successfully.
```

- Πατάμε το 2 για την προσθήκη σε άλμπουμ και έπειτα 1 για να εισάγουμε εμείς

```
Press 1 to add a new doc in songs and lyrics
Press 2 to add a new doc in albums
Enter your input: 2
Type 1 to add your input
Type 2 to add an already csv file
Type exit in order to return to the main menu
Enter your input: 1
Enter singer name: Pop Smoke
Enter album name: Shoot For The Stars Aim For The Moon
Enter album type: album
Enter album year of release: 2020
Data added to the CSV file successfully.
```

- Πατάμε 2 για την εισαγωγή δεδομένων από αρχείο csv

```
Enter your input: 2
Type 1 to add your input
Type 2 to add an already csv file
Type exit in order to return to the main menu
Enter your input: 2
Enter the name of the CSV file: album_add
Opening album_add.csv
Format is ok
Continuing with indexing and updating the csv file...
Data from /project/user_csv/album_add.csv added to the merged CSV file.
Indexing documents:
```

- Πατάμε 4 για διαγραφή
  - Πατάμε 1 για διαγραφή τραγουδιού

```
4
Press 1 to remove a song
Press 2 to remove an album
Enter your input:1
Remove document...
Enter singer name : french montana
Enter song name : unforgettable
Singer query = singer_name:french singer_name:montana
Song name query = song_name:unforgettable
The term is found in the following documents:
Doc_id =26460
Number 1
Song name: Unforgettable
Singer name: French Montana
Lyrics: It's not good enough for me, since I been with you It's not gonna work for you, nobody can equal me
Link:/frenchmontana/unforgettable.html
----------------------------------------
Are you sure you want to delete this document? [y]/[n]y
Document with docID: 26460 and singer name: french montana and song name: unforgettable removed from the index.
```

  - Πατάμε 2 για διαγραφή άλμπουμ

```
4
Press 1 to remove a song
Press 2 to remove an album
Enter your input:2
Enter singer name : Pop smoke
Enter album name : Shoot For The Stars Aim For The Moon
Singer query = singer_name:pop singer_name:smoke
Album name query = name:shoot name:for name:the name:stars name:aim name:for name:the name:moon
The term is found in the following documents:
Doc_id =1819
Number 1
Singer name: Pop Smoke
Album name: Shoot For The Stars Aim For The Moon
Type: album
Year:2020
----------------------------------------
Are you sure you want to delete this document? [y]/[n]y
Document with docID: 1819 and singer name: Pop smoke and song name: Shoot For The Stars Aim For
The Moon removed from the index.
```

- Πληκτρολογούμε 5 για αναζήτηση
  - 1 για αναζήτηση τραγουδιού

```
5
Search document...
Where to search:
Press 1 to search a song based on artist, song name, lyrics
Press 2 to search an album based on artist, type of album, year
Enter your input: 1
Enter singer name:arctic AND monkeys
Enter song name:i wanna be yours
Enter lyrics:
Performing boolean query
How many results to output: 5
The term is found in the following documents:
Number 1
Doc_id =13990
Score: 12.470056533813477
Song name: i wanna be yours
Singer name: arctic monkeys lyrics
Link:../lyrics/arcticmonkeys/iwannabeyours.html
----------------------------------------
Number 2
Doc_id =13979
Score: 8.062274932861328
Song name: do i wanna know?
Singer name: arctic monkeys lyrics
```

  - VSM ερώτημα σε τραγούδι

```
Enter your input: 1
Enter singer name:arctic monkeys
Enter song name:i wanna be yours
Enter lyrics:
Performing vsm query
How many results to output: 1
The term is found in the following documents:
Number 1
Doc_id =13990
Score: 12.470056533813477
Song name: i wanna be yours
Singer name: arctic monkeys lyrics
Link:../lyrics/arcticmonkeys/iwannabeyours.html
```

  - Phrase ερώτημα

```
5
Search document...
Where to search:
Press 1 to search a song based on artist, song name, lyrics
Press 2 to search an album based on artist, type of album, year
Enter your input: 1
Enter singer name:
Enter song name:
Enter lyrics:"i wanna be your vacuum cleaner"
Performing phrase query
How many results to output: 1
The term is found in the following documents:
Number 1
Doc_id =13990
Score: 10.993555068969727
Song name: i wanna be yours
Singer name: arctic monkeys lyrics
Lyrics:

i wanna be your vacuum cleaner
```

- Αναζήτηση σε άλμπουμ
    - Boolean ερώτημα

```
Enter your input: 2
Enter singer name:linkin AND park
Enter album name:
Enter the type of the album :
Enter year of release of the album:
Performing boolean query in albums docs...
How many results to output: 3
The term is found in the following documents:
Number 1
Score: 3.8544726371765137
Singer name: linkin park lyrics
Album name: xero
Type: demo
Year:1997
---------------------------------------
Number 2
Score: 3.8544726371765137
Singer name: linkin park lyrics
Album name: hybrid theory
Type: EP
Year:1999
```

- VSM ερώτημα σε άλμπουμ

```
Enter your input: 2
Enter singer name:
Enter album name:
Enter the type of the album :demo
Enter year of release of the album:
Performing vsm query
How many results to output: 2
The term is found in the following documents:
Number 1
Score: 5.801832675933838
Singer name: linkin park lyrics
Album name: xero
Type: demo
Year:1997
----------------------------------------
Number 2
Score: 5.801832675933838
Singer name: linkin park lyrics
Album name: hybrid theory 8-track demo
Type: demo
Year:1999
----------------------------------------
Search took 0.54 milliseconds
```

- Phrase ερώτημα σε άλμπουμ

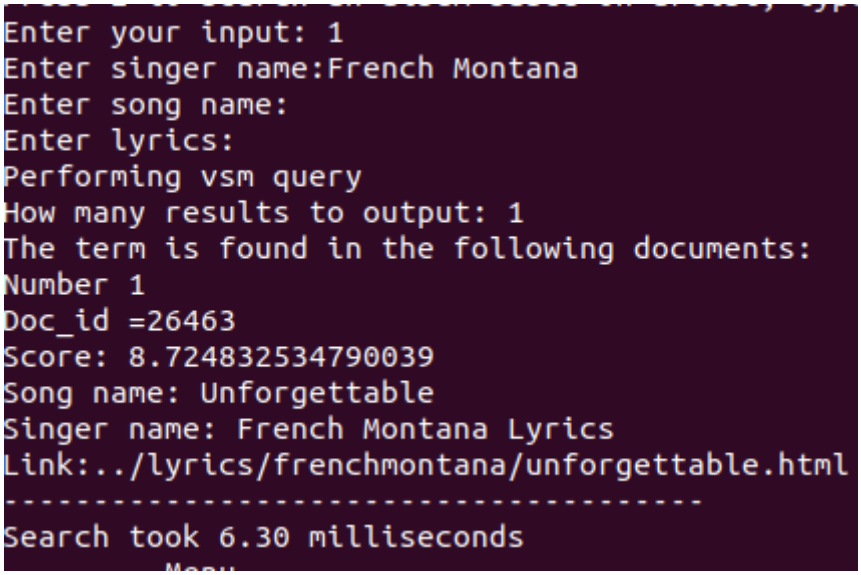
```
Enter your input: 2
Enter singer name:"coldplay"
Enter album name:
Enter the type of the album :album
Enter year of release of the album:
Performing phrase query...
How many results to output: 3
The term is found in the following documents:
Number 1
Score: 2.7635884284973145
Singer name: coldplay lyrics
Album name: parachutes
Type: album
Year:2000
---------------------------------------
Number 2
Score: 2.7635884284973145
Singer name: coldplay lyrics
Album name: a rush of blood to the head
Type: album
Year:2002
---------------------------------------
Number 3
Score: 2.7635884284973145
```

- Πληκτρολογούμε το 6 για να κάνουμε ανάκτηση πληροφορίας από την ιστοσελίδα AZlyrics

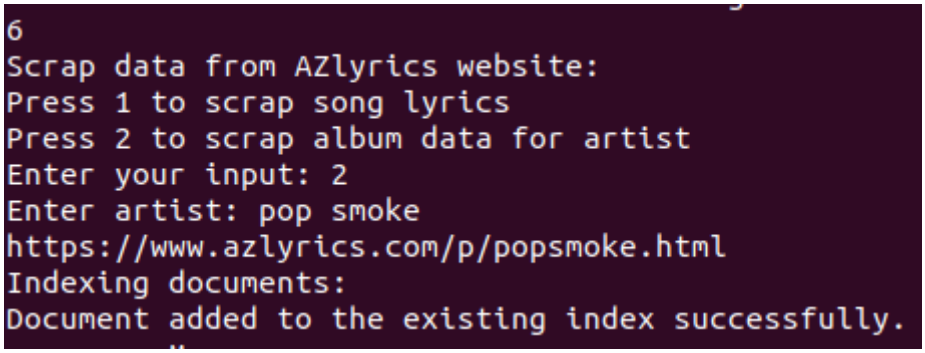
    - 1 για να εισάγουμε τραγούδι

```
6
Scrap data from AZlyrics website:
Press 1 to scrap song lyrics
Press 2 to scrap album data for artist
Enter your input: 1
Enter artist: French Montana
Enter a song by the same artist: Unforgettable
Lyrics successfully written to file for: French Montana LyricsUnforgettable
Indexing documents:
Data added to the CSV file successfully.
Document added to the existing index successfully.
```

Ας κάνουμε και μία αναζήτηση για να δούμε ότι υπάρχει



- Πληκτρολογούμε 2 για να εισάγουμε άλμπουμ



Αναζητούμε για να δούμε εάν υπάρχει πλέον στο ευρετήριο μας

```
Performing vsm query
How many results to output: 3
The term is found in the following documents:
Number 1
Score: 7.760578155517578
Singer name: Pop Smoke Lyrics
Album name: Meet The Woo
Type: mixtape
Year:2019
----------------------------------------
Number 2
Score: 7.760578155517578
Singer name: Pop Smoke Lyrics
Album name: Meet The Woo, Vol. 2
Type: mixtape
Year:2020
----------------------------------------
Number 3
Score: 7.760578155517578
Singer name: Pop Smoke Lyrics
Album name: Shoot For The Stars Aim For The Moon
Type: album
Year:2020
----------------------------------------
Search took 0.91 milliseconds
```

5 Επίλογος

Εν κατακλείδι, η εξαμηνιαία εργασία στην Ανάκτηση Πληροφοριών, η οποία επικεντρώθηκε στην ανάπτυξη μιας μηχανής αναζήτησης τραγουδιών με την χρήση του Lucene ήταν ένα σημαντικό ταξίδι έρευνας και μάθησης. Η δημιουργία μιας λειτουργικής μηχανής αναζήτησης τραγουδιών επέτρεψε την πρακτική εφαρμογή των θεωρητικών γνώσεων που αποκτήθηκαν κατά τη διάρκεια αυτού του μαθήματος.

Αυτό το έργο απαιτούσε προσεκτική προσοχή στη λεπτομέρεια, που περιλάμβανε την οργάνωση δεδομένων, την επιλογή αλγορίθμων και τις στρατηγικές υλοποίησης. Παρείχε μια πολύτιμη πρακτική εμπειρία στην κατανόηση της πολυπλοκότητας και των λεπτομερειών των συστημάτων ανάκτησης πληροφοριών.

Επιπλέον, πέρα από τις τεχνικές πτυχές, το εγχείρημα αυτό ανέδειξε την κρίσιμη διασταύρωση της θεωρίας και της πράξης σε πραγματικές εφαρμογές. Υπογράμμισε τη σημασία της σύνδεσης της ακαδημαϊκής γνώσης με την πρακτική εφαρμογή για αποτελεσματικές και αποδοτικές λύσεις.

Καθώς το έργο αυτό φτάνει στο τέλος του, σηματοδοτεί όχι μόνο την κορύφωση των προσπαθειών αυτού του εξαμήνου, αλλά χρησιμεύει και ως βήμα προς περαιτέρω εξερεύνησης στον τομέα της Ανάκτησης Πληροφοριών.

Αυτή η εμπειρία έχει προκαλέσει μια βαθύτερη εκτίμηση για τις ιδιαιτερότητες της ανάπτυξης μηχανών αναζήτησης και τη σημασία της στα σύγχρονα συστήματα πληροφοριών. Περιείχε πολύτιμες γνώσεις για τον τρόπο δημιουργίας και λειτουργίας των σύγχρονων μηχανών αναζήτησης.

6 ΠΗΓΕΣ

- Apache Pylucene, https://lucene.apache.org/pylucene/install.html
- coady/pylucene:8, https://hub.docker.com/layers/coady/pylucene/8/images
- Apache Lucene Documentation, https://lucene.apache.org/core/8_9_0/index.html
- Pylucene Examples, https://notebook.community/paulovn/ml-vm-notebook/vmfiles/IPNB/Examples/
- CREDITS file μέσα στον φάκελο του project