ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ
Υλοποιητικό Πρότζεκτ
Χειμερινό Εξάμηνο 2020-2021

Παναγιώτης Χριστόπουλος 1054409
Χρήστος Στεμιτσιώτης 1054375

Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Εισαγωγή

Η εργασία υλοποιήθηκε στη γλώσσα Python με τη χρήση του Visual Studio Code. Χρησιμοποιήθηκαν οι παρακάτω βιβλιοθήκες:

- elasticsearch
- json
- csv

Ερώτημα 1

Στο script με όνομα elasticImport.py, δημιουργείται μια συνάρτηση η οποία φορτώνει αρχεία csv στην ElasticSearch στο index "movies". Στη συνέχεια, η συνάρτηση καλείται για να φορτωθεί το αρχείο movies.csv.

Αποτελέσματα εκτέλεσης στο τερματικό

http://127.0.0.1:9200/movies/

Στο script με όνομα bm25Similarity.py, ζητείται από το χρήστη να πληκτρολογήσει το όνομα της ταινίας που θέλει να ψάξει.
Το ζητούμενο query φορτώνεται σύμφωνα με τη βιβλιοθήκη ElasticSearch της Python.
Εμφανίζεται ο αριθμός των αποτελεσμάτων και το όνομα των αντίστοιχων ταινιών ακολουθούμενο από το δείκτη ομοιότητας

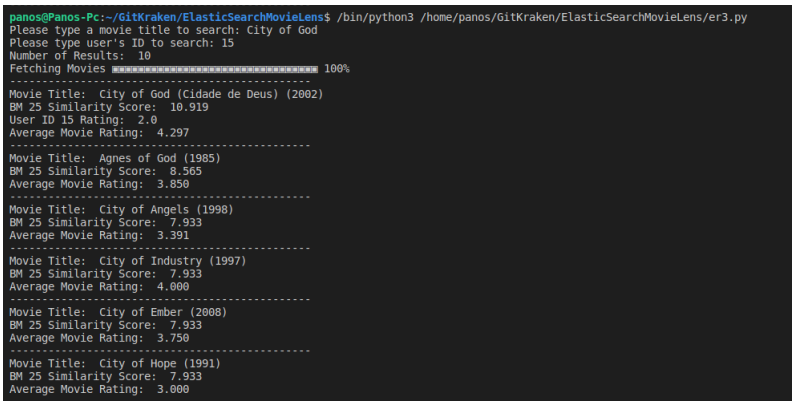
```
panos@Panos-Pc:~/Desktop/Ανακτηση$ /bin/python3 /home/panos/Desktop/Ανακτηση/bm25Similarity.py
Please type a movie title to search: City of God
Number of Results:  10
Movie Title: City of God (Cidade de Deus) (2002) Similarity Score:  11.001702
Movie Title: Agnes of God (1985) Similarity Score:  8.665753
Movie Title: City of Industry (1997) Similarity Score:  7.9139977
Movie Title: City of Angels (1998) Similarity Score:  7.9139977
Movie Title: City of Hope (1991) Similarity Score:  7.9139977
Movie Title: City of Ember (2008) Similarity Score:  7.9139977
Movie Title: Children of a Lesser God (1986) Similarity Score:  7.2449474
Movie Title: Prince of the City (1981) Similarity Score:  7.2072883
Movie Title: Dear God (1996) Similarity Score:  7.011532
Movie Title: Playing God (1997) Similarity Score:  7.011532
```

Εκτέλεση script bm25Similarity.py

Ερώτημα 2

Αρχικά, διαβάζουμε τα αρχεία ratings.csv και movies.csv και μέσω της βιβλιοθήκης pandas τα μετρατρέπουμε σε data frames.

Εν συνεχεία, υλοποιήσαμε τρεις συναρτήσεις. Η movieID μας βοηθάει ώστε να βρούμε το ID της ταινίας έχωντας ως όρισμα το ονομά της. Η findUserVote χρησιμοποιείται ώστε να βρίσκουμε την ψήφο του χρήστη για μια συγκεκριμένη ταινία με ορίσματα το ID του χρήστη και το όνομα της ταινίας. Η findMovieAvarage που υλοποιήθηκε βρίσκει το μέσο όροτων ψήφων κάθε ταινίας.
Τέλος,το σύστημα ζητάει απο τον χρήστη να πληκτρολογήσει το όνομα κάποιας ταινίας και το ID του χρήστη και εμφανίζει τα αποτελέσματα. 'Ολα αυτα βρίσκονται στο script er2.py.

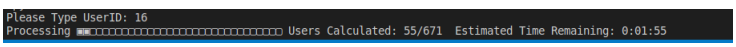


Εκτέλεση script er2.py

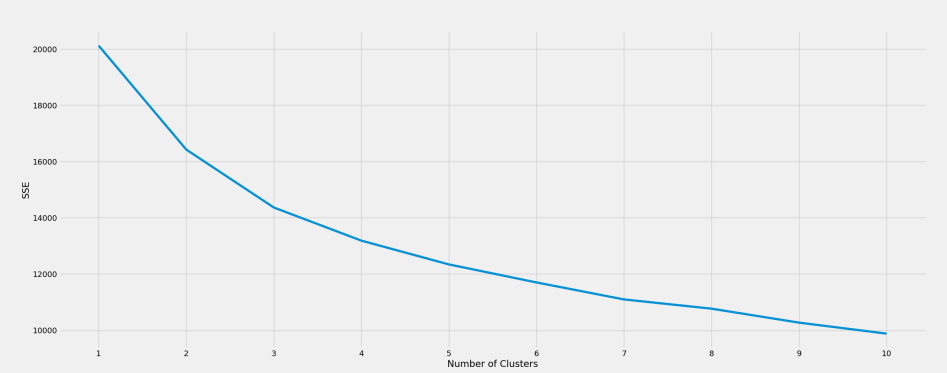
Η παραπάνω εκτύπωση του ερωτήματος 2 είναι για το query "City of God". Όπως βλέπουμε στο πρώτο αποτέλεσμα, προβάλλει το BM25 Similarity Score, το μέσο όρο βαθμολογίας της ταινίας αλλά και τη βαθμολογία του χρήστη αν υπάρχει.

Ερώτημα 3

Για το ερώτημα 3, δημιουργήθηκαν οι βοηθητικές συναρτήσεις findUserVote, findUserRatedMoviesList, findMoviesByGenre και userDataAcc, η οποία επιστρέφει μία λίστα με τα averages του κάθε χρήστη. Αφού δοθεί ως input το id του επιθυμητού χρήστη, βρίσκονται οι μέσοι όροι ανα genre τόσο του ίδιου όσο και των υπόλοιπων χρηστών.



Στη συνέχεια, με τη βοήθεια του KL Elbow, βρίσκεται ο κατάλληλος αριθμός clusters.



Αξίζει να σημειωθεί ότι οι ήδη βαθμολογημένες ταινίες δεν προτείνονται στο χρήστη.

```
Suggested movies by users in the same cluster:
[28, 'Star Wars: Episode IV - A New Hope (1977)', 5.0]
[27, 'Jurassic Park (1993)', 5.0]
[25, 'Terminator 2: Judgment Day (1991)', 5.0]
[24, 'Godfather, The (1972)', 5.0]
[33, 'Pulp Fiction (1994)', 4.256]
[22, 'Fight Club (1999)', 4.178]
[26, 'Star Wars: Episode VI - Return of the Jedi (1983)', 4.06]
[21, 'Sixth Sense, The (1999)', 4.018]
[21, 'Braveheart (1995)', 4.0]
[30, 'Forrest Gump (1994)', 4.0]
[29, 'Silence of the Lambs, The (1991)', 4.0]
[37, 'Fargo (1996)', 4.0]
[20, 'Star Wars: Episode V - The Empire Strikes Back (1980)', 4.0]
[23, 'Back to the Future (1985)', 4.0]
[21, 'Independence Day (a.k.a. ID4) (1996)', 3.484]
[21, 'Twister (1996)', 3.25]
[24, 'Matrix, The (1999)', 3.0]
```

Προτεινόμενες ταινίες από χρήστες του ίδιου cluster

Υλοποιήθηκαν οι κατάλληλες συναρτήσεις έτσι ώστε μέσα από τη βιβλιοθήκη imdbpy να γίνεται η κατάλληλη παρουσίαση των ταινιών.

```
--------------------------------------------------
Title: Pulp Fiction        Average clustered user vote average: 4.256 from a total of 33 votes
Year: 1994
Countries: United States
Director: Quentin Tarantino
Actors: Tim Roth, Amanda Plummer, Laura Lovelace
IMDB Plot:
Faced with life's cruel irony, the unpredictable stories of a well-dressed pair of low-level hitmen; a gangster's statuesque moll, and a double-crossing prizefighter become inextricably intertwined, as the s
age to hold up their favourite L.A. diner. Entrusted with retrieving a glow-emitting leather suitcase which belongs to their boss--the powerful crime kingpin, Marsellus--instead, the dark-suited gunmen, Vinc
unately--with men like Mr Wolf always around to handle a crisis--there's time to cool off in a long twist contest, while at the same time, the proud champion boxer, Butch, makes the decision of a lifetime. S
incent find themselves in the perfect dead-end situation, exactly where it all began: an all-too-familiar cafeteria. Is truth stranger than fiction?::Nick Riganas
--------------------------------------------------
```

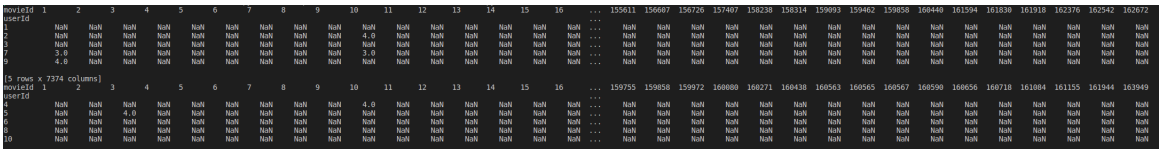
Όπως φαίνεται στην παραπάνω εικόνα, στην παρουσίαση εμφανίζονται τα εξής:

- Τίτλος
- Μέσος όρος βαθμολογίας χρηστών του ίδιου cluster
- Χρονιά παραγωγής
- Χώρες παραγωγής
- Σκηνοθέτης
- Ηθοποιοί
- Πλοκή

Εμφανίζονται οι 30 καταλληλότερες ταινίες

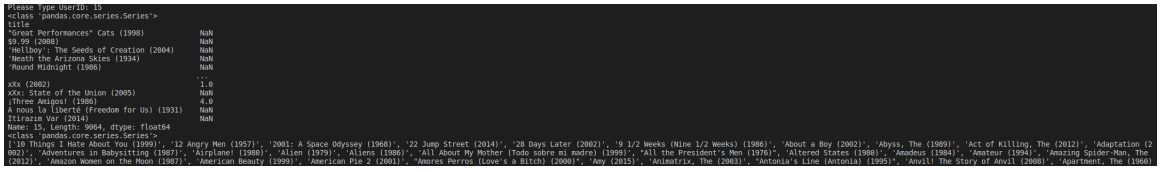
Ερώτημα 4

Αρχικά, δημιουργήθηκε μία συνάρτηση έτσι ώστε να χωρίσει τους χρήστες σε train και validation set με αναλογία 60-40.



Train και Validation Sets

Στη συνέχεια φιλτράρονται οι θετικές ψήφοι έτσι ώστε να χρησιμοποιηθούν στη συνέχεια οι τίτλοι των ταινιών ως word embeddings.



Ενδεικτικές Βαθμολογίες χρήστη 15 και παρουσίαση λίστας θετικών ψήφων