

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ
Υλοποιητικό Πρότζεκτ
Χειμερινό Εξάμηνο 2020-2021

Παναγιώτης Χριστόπουλος 1054409
Χρήστος Στεμτσιώτης 1054375



Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

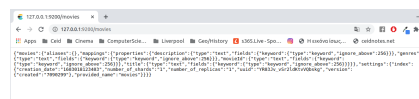
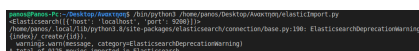
Εισαγωγή

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.



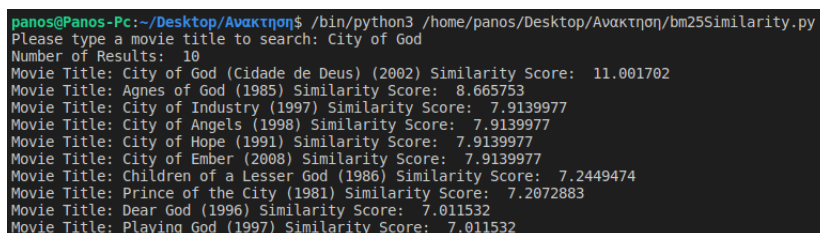
Αποτελέσματα εκτέλεσης στο τερματικό

<http://127.0.0.1:9200/movies/>

Στο script με όνομα `bm25Similarity.py`, ζητείται από το χρήστη να πληρολογήσει 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



Εκτέλεση script bm25Similarity.py

Ερώτημα 2

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

Εν συνεχεία, υλοποιήσαμε τρεις συναρτήσεις. Η `movieID` μας βοηθάει ώστε να βρούμε το ID της ταινίας έχοντας 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 `findMovieAverage` που υλοποιήθηκε βρίσκει το μέσο όρο των ψήφων κάθε ταινίας.

Τέλος, το σύστημα ζητάει από τον χρήστη να πληκτρολογήσει το όνομα κάποιας ταινίας και το ID του χρήστη και εμφανίζει τα αποτελέσματα. Όλα αυτά βρίσκονται στο script er2.py.

```
panos@Panos-PC:~/GitKraken/ElasticSearchMovieLens$ ./bin/python3 /home/panos/GitKraken/ElasticSearchMovieLens/er3.py
Please type a movie title to search: City of God
Please type user's ID to search: 15
Number of Results: 10
Fetching Movies ..... 100%

Movie Title: City of God (Cidade de Deus) (2002)
BM 25 Similarity Score: 10.919
User ID 15 Rating: 2.0
Average Movie Rating: 4.297
-----
Movie Title: Agnes of God (1985)
BM 25 Similarity Score: 8.565
Average Movie Rating: 3.850
-----
Movie Title: City of Angels (1998)
BM 25 Similarity Score: 7.933
Average Movie Rating: 3.391
-----
Movie Title: City of Industry (1997)
BM 25 Similarity Score: 7.933
Average Movie Rating: 4.600
-----
Movie Title: City of Ember (2008)
BM 25 Similarity Score: 7.933
Average Movie Rating: 3.750
-----
Movie Title: City of Hope (1991)
BM 25 Similarity Score: 7.933
Average Movie Rating: 3.000
```

Εκτέλεση script er2.py

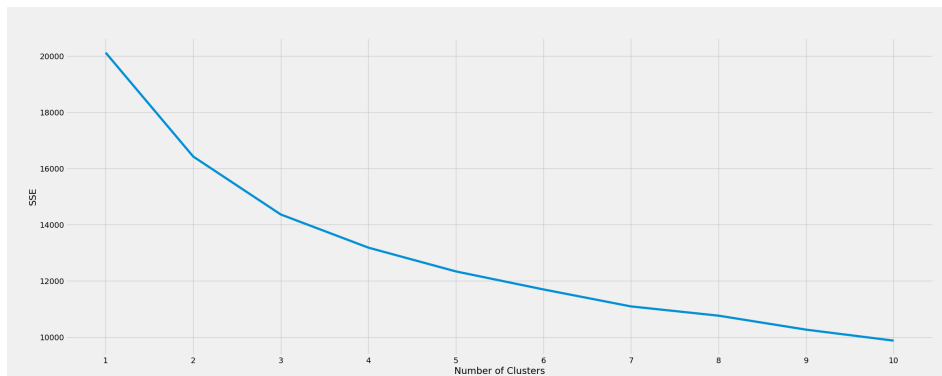
Η παραπάνω εκτύπωση του ερωτήματος 2 είναι για το query "City of God". Όπως βλέπουμε στο πρώτο αποτέλεσμα, προβάλλει το BM25 Similarity Score, το μέσο όρο βαθμολογίας της ταινίας αλλά και τη βαθμολογία του χρήστη αν υπάρχει.

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.

```
Please Type UserID: 16  
Processing ██████████ Users Calculated: 55/671 Estimated Time Remaining: 0:01:55
```

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
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...
incient 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.

userId	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...	15611	15667	156726	157407	158238	158314	158993	159462	159858	160440	161594	161839	161918	162376	162542	162672
movieId	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...	15611	15667	156726	157407	158238	158314	158993	159462	159858	160440	161594	161839	161918	162376	162542	162672
rating	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
time	1208286400	1208290000	1208293600	1208297200	1208300800	1208304400	1208308000	1208311600	1208315200	1208318800	1208322400	1208326000	1208329600	1208333200	1208336800	1208340400	...	1208344000	1208347600	1208351200	1208354800	1208358400	1208362000	1208365600	1208369200	1208372800	1208376400	1208380000	1208383600	1208387200	1208390800	1208394400	1208398000

Train και Validation Sets

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

Please type user ID: 15	class pandas.core.series.Series>	
>title	NaN	
>direct Performances> Cats (1998)	NaN	
>50 50 (2005)	NaN	
>Hellboy: The Seeds of Creation (2004)	NaN	
>Heath Arizona Sides (1934)	NaN	
>Round Midnight (1996)	NaN	
>	NaN	
>xxx (2002)	1.0	
>xxx: State of the Union (2005)	NaN	
>Three Amigos! (1989)	4.0	
>I'm on a Liberty (freedom for us) (1931)	NaN	
>Harrison Var (2014)	NaN	
>name: 35, length: 3664, dtype: float64		
>class pandas.core.series.Series>		
>021, 'Adventures in Babysitting' (1987), '12 Angry Men (1957)', '2001: A Space Odyssey (1968)', '22 Jump Street (2014)', '28 Days Later (2002)', '9 1/2 Weeks (Nine 1/2 Weeks) (1986)', 'A Boy and a Girl (2002)', 'Abyss, The (1989)', 'Act of Killing, The (2012)', 'Adaptation (2002)', 'Adventures in Babysitting' (1987), 'Airplane' (1980)', 'Alien (1979)', 'Alien (1986)', 'All About My Mother (Todo sobre mi madre) (1999)', 'All the President's Men (1976)', 'Altered States (1986)', 'Amadeus (1984)', 'Amateur (1994)', 'Amazing Spider-Man, The (2002)', 'American Beauty (1999)', 'American Beauty (Lower) (1995)', 'Anchorman: The Legend (2001)', 'Antonia's Line (Antonia) (1995)', 'Anvil! The Story of Anvil (2008)', 'Appointment (The) (1990)		

Ενδεικτικές Βαθμολογίες χρήστη 15 και παρουσίαση λίστας θετικών ψήφων