Homework #1 – MongoDB

Introduzione - L'esercizio richiede di caricare in MongoDB un dataset di film, applicare una validazione dei dati e rispondere a 5 domande analitiche utilizzando Python o la shell di MongoDB (ad esempio Studio3T). Il dataset è quello disponibile a questo indirizzo: https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

Descrizione del compito - Scaricare e salvare il datase dentro un database MongoDB sotto forma di documenti JSON. Durante il salvataggio, organizzare in automatico i film in diverse collezioni a seconda del certificato acquisito da ciascun film (U="Universal", ovvero adatto a tutti; UA="Parental Guidance", ovvero da vedere con la presenza di una persona adulta; A="Adult"). I film che non rientrano in nessuna delle categorie precedenti vanno salvati in una collezione a parte ("Trash"), così come tutti film usciti prima del 1980 e con un voto complessivo inferiore a 8.

Eseguire le seguenti interrogazioni sul database appena creato, escludendo dalla ricerca la collezione "Trash".

1. Trovare tutti i film pubblicati dopo il 2015 raggruppati per genere (ad esempio "fantascienza" o "giallo"); visualizzare in un istogramma il numero di film per ogni genere
2. Trovare i primi 5 film per ogni genere, sulla base del voto.
3. Trovare il film più vecchio e più recente per ogni genere.
4. Trovare il film più vecchio con la valutazione più alta.
5. Trovare i primi 5 film con la maggiore durata.
6. Trovare l'ammontare degli incassi per ogni genere e visualizzarlo in un istogramma.

Ripetere le interrogazioni precedenti, però questa volta distinguendo per certificato acquisito da ciascun film. Fare delle considerazioni in merito alle tempistiche richieste da MongoDB in un caso (ovvero senza distinguere per certificato) e nell'altro. Come si potrebbe trarre il massimo vantaggio, in termini di tempi di risposta, volendo mantenere tutti e due i casi?

Si supponga ora che il docente di Sistemi Informativi Evoluti e Big Data abbia una predilezione per il genere Sci-Fi e Adventure, ma non ami particolarmente i film troppo lunghi. Proporre un algoritmo di raccomandazione basato su queste preferenze e che gli suggerisca 3 film da guardare. Come si classificherebbe tale algoritmo?
Verificare se esistono nel database dei film classificati sia come Sci-Fi e Adventure, che non durano più di 90 minuti. Se sì, visualizzare i primi 3 (in base al rating) e confrontare il risultato di questa query rispetto all'output dell'algoritmo di raccomandazione. Se non esistono film con questi requisiti nel database, provare a ipotizzare quale potrebbe essere un rating di un film classificato sia come Sci-Fi e Adventure, che non duri più di 90 minuti. Confrontare anche in quest'ultimo caso la previsione con quanto suggerito dall'algoritmo di raccomandazione.

Dettagli sulla consegna – Predisporre un notebook Python (o un file .py) in cui è riportato lo svolgimento di tutti gli esercizi e un file PDF in cui sono riportati i commenti sui risultati laddove esplicitamente richiesto. Includere tutti i file in uno zip e caricare l'archivio su Moodle.