Report IUM

Sommario

1. Introduzione 1
2. Soluzione 1
2.1 Data cleaning 1
2.1.1 Problematiche 2
2.1.2 Requisiti 2
2.1.3 Limitazioni 2
2.2 Data analysis 2
2.2.1 Problematiche 3
2.2.2 Requisiti 3
2.2.3 Limitazioni 3
3. Conclusioni 3
4. Informazioni aggiuntive 3
5. Bibliografia 3

1. Introduzione

Il progetto, nella sua parte di IUM, si è rivolto alla gestione di grandi moli di dati, che andavano puliti per essere resi utilizzabili al fine della creazione del sito valutato nella parte di Tweb del corso IUM-Tweb (12 crediti). Per svolgerlo siamo partiti dalla pulizia dati, fino ad arrivare la parte di analisi che ha prodotto due risultati diversi, con risultati raccolti nei file “cartoon.ypnb” e “nolan.ypnb”.

2. Soluzione

2.1 Data cleaning

Sono stati creati diversi file Jupyter Notebook per fare la pulizia dei dataset fornitoci, uno per ognuno dei file che abbiamo. Abbiamo importato la libreria Pandas per eseguire le pulizie. Il lavoro ha avuto diverse fasi, che possono essere ricondotte a queste:

- Importare il dataset, visualizzarlo e capirne il contenuto.
- Salvataggio dei dati in locale, per modifiche non distruttive

- Rimuovere o gestire i valori NA
- Rimuovere o gestire i duplicati
- Conversione di eventuali types
- Esportare il file csv, caricandolo nei databases corretti

2.1.1 Problematiche

La pulizia dei dati è risultata un lavoro che nel complesso è scorso velocemente, nonostante le difficoltà riscontrate in due file in particolare:

- "**Rotten_tomatoes_review": ** La principale difficoltà affrontata con questo dataset è stata la conversione dei valori della colonna *review_score* in un'unica unità di misura universale, poiché le valutazioni erano espresse in formati diversi. Per uniformarle, abbiamo adottato una scala da 0 a 10. (Abbiamo scelto di mantenere il punteggio della critica in 0/5 volontariamente).
- "**Movies**": La difficoltà qui è stata quella di controllare la consistenza dei dati forniti, data appunto la presenza di film ancora non usciti o con minutaggio spropositato, che dopo alcuni controlli abbiamo eliminato.

2.1.2 Requisiti

- Sviluppo di uno o più Jupyter Notebook che permettano l'analisi dei dati. (Soddisfatto, i file si trovano nella cartella "cleaning_scripts")
- Possibilità di interrogare il database per estrarre un sottoinsieme di dati (ad esempio, statistiche su un film specifico o su un attore specifico) (Soddisfatto, i database sono pronti per lavorarci sopra in maniera soddisfacente).

2.1.3 Limitazioni

Data la mole dei dati, garantiamo dei buoni risultati per quanto riguarda la pulizia, ma è possibile ci siano elementi che non siamo riusciti a filtrare nella maniera migliore possibile.

2.2 Data analysis

Durante il processo di analisi dati abbiamo fatto l'analisi di diversi aspetti dei dataset, con l'obiettivo di ottenere dati significativi sui film, gli attori, i premi Oscar e le recensioni dei critici. Abbiamo importato le librerie **Pandas, Matplotlib, Seaborn, Numpy, e Geopandas** per poter applicare tecniche di manipolazione, aggregazione e visualizzazione dei dati per rispondere a domande chiave sulle performance dei film e degli attori.

2.2.1 Problematiche

La maggior parte delle problematiche riscontrate sono state relative alla creazione di grafici che fossero all'altezza delle nostra aspettative. Alla fine del lavoro, siamo riusciti ad ottenere molti grafici diversi e che riescono però a spaziare bene e dare l'idea dei dati che abbiamo voluto mostrare nelle storie.

2.2.2 Requisiti

- Ampia varietà di visualizzazioni, evitando l'uso di un solo tipo di grafico (ad esempio, non solo grafici a barre). È richiesto anche l'uso di visualizzazioni geografiche. (Soddisfatto, proponiamo circa 20 grafici a file, per la maggior parte diversi l'uno dall'altro)
- Consegna in formato già eseguito, mostrando tutti i grafici e le tabelle senza la necessità di eseguire nuovamente il codice durante la valutazione (Soddisfatto, il codice è consegnato con ogni diagramma eseguito)

2.2.3 Limitazioni

Ovviamente, i grafici possono in alcuni casi essere limitati a spazi di tempo/quantità di dati ridotte, ma è tutto puramente legato agli argomenti scelti per le due storie, che appunto spaziano di una trentina di anni nel loro spazio temporale.

3. Conclusioni

Reputiamo il lavoro di analisi dati soddisfacente e in linea con la consegna dataci. Riteniamo inoltre di aver appreso in maniera soddisfacente gli strumenti basi per analizzare dati e creare grafici soddisfacenti. Come per il lavoro di Tweb, la divisione dei compiti è stata equa, con un file di narrazione a ciascuno e la pulizia dei dati divisa esattamente a metà e portata avanti in parallelo. Dunque, la divisione risulterà essere tale:

- Agnese Andriani 50%
- Davide Giordano 50%

4. Informazioni aggiuntive

Nessuna informazione aggiuntiva.

5. Bibliografia

Abbiamo utilizzato le slide su moodle e chatGPT per la generazione automatica di codice. In seguito alla ricezione del codice da parte dell'intelligenza artificiale è stato eseguito il seguente procedimento: lettura e comprensione del codice, valutazione

della correttezza e infine adattamento al progetto, al fine di un utilizzo il più didattico possibile.