Relazione Interazione Uomo Macchina

Componenti del gruppo:

* [Raffaele Melluso](mailto:raffaelemelluso01@gmail.com), matr. 1045495.

[Repository github solo per la parte di IUM](https://github.com/RaffaeleMelluso/IUM-Progetto-Melluso)

**Introduzione**

Da sempre per fattori storici e culturali, gli Stati Uniti sono stati il paese di riferimento per la produzione di film e serie televisive. Questo ha portato ad una grande quantità di contenuti prodotti, ma non sempre di qualità. Forse questo è dovuto da un circolo che si è venuto a creare di produzione di contenuti per soddisfare la domanda, che a sua volta è cresciuta grazie alla grande fama di Hollywood e simili. Perché si ha la sensazione di una forbice tra la qualità dei film più famosi e quelli dozzinali? Questo affossa la valutazione media dei film prodotti negli Stati Uniti? Come è cambiata la valutazione dei film prodotti negli Stati Uniti nel tempo? I migliori studios di produzione sono quelli che producono film migliori?

Nel seguente progetto si intende rispondere a queste domande verificando la veridicità di esse e cercando di rispondervi con l’analisi dei dataset forniti.

**Technical Tasks**

**Parte 1: pulizia dei dataset**

* L’approccio scelto per la pulizia dei dati è stato quello di pensare ad un futuro utilizzo dei dataset all’interno di un database, quindi adottando una pulizia generale per creare un database “ready to use” e successivamente una pulizia specifica per permettere la corretta analisi dei dati.

Inoltre non c’è stato bisogno di fare casting dei tipi dato che i valori numerici erano già correttamente convertiti in automatico e i tipi “stringa” anche dopo la conversione restano “object”.

* Pulizia di Actors:
  + rimozione attori con nome nullo;
  + rimozione attori duplicati.
* Pulizia di Countries:
  + normalizzazione nomi dei paesi per l’uso di geopandas.
* Pulizia di Crew:
  + rimozione membri della crew con nome nullo;
  + rimozione membri della crew duplicati.
* Pulizia di Genres:
  + già pulito.
* Pulizia di Languages:
  + già pulito.
* Pulizia di Movies:
  + conversione di ‘date’ e ‘minute’ ad int64;
  + rimozione film con nome nullo.
* Pulizia di Posters:
  + già pulito.
* Pulizia di Releases:
  + aggiunta ai records di movies con ‘date’ nulla la loro release.
* Pulizia di Studios:
  + rinominati alcuni studios con nome nullo in ‘Unknown’;
  + rimozione dei duplicati.
* Pulizia di Themes:
  + già pulito.
* Pulizia di Rotten\_tomatoes\_reviews:
  + standardizzazione delle valutazioni con punteggio in percentuale;
  + rimozione dei duplicati.
* Pulizia di The\_oscar\_awards:
  + rimozione oscar duplicati.

**Problemi:**

Gli unici problemi riscontrati sono stati nella verifica e nella correzione dei tipi di essi, perché effettuando la conversione dei tipi object a stringa restavano object.

Inoltre un problema enorme è stato la conversione dei countries per renderli compatibili con la mappa usata su geopandas.

**Requisiti**

Alla fine della pulizia i dataset vengono salvati come .csv in una cartella chiamata Output, essa se non esiste viene creata dal sistema. Inoltre vengono conservati tutti i dataset originali nella cartella Data, per confrontarli con quelli puliti.

**Limitazioni**

Alcune colonne vengono lasciate nulle per permettere la ricerca e l’inserimento di dati come ad esempio in poster, in cui si fa la ricerca per id del film e viene caricata.

Alcuni campi composti da liste di stringhe non vengono convertiti perché in nessuna delle due parti del progetto c’è il bisogno di utilizzarli in modo separato.

**Parte 2: analisi dei dataset**

Per l’analisi dei dati si è deciso di caricare i files creati dopo la pulizia direttamente da csv come suggerito dal professore. Quindi diamo già per scontato che i singoli dataset non abbiano bisogno di controlli aggiuntivi se non per le merge. Cominciamo con il creare il dataset più utilizzato sin dall’inizio, quello che unisce movies, countries e studios. Per rispondere alla domanda iniziale del progetto questo dataset sarà di grande importanza. Inoltre per rispondere alle domande poste nel documento si è deciso prima di valutare la veridicità di esse, e successivamente di cercare di rispondervi. Il lavoro viene suddiviso in varie sezioni, quali:

* Visualizzazione dei dati sulla produzione dei film mondiali, utilizzando grafici a barre, a torta e una mappa del mondo;
* Confronto della valutazione media dei film tra USA e resto del mondo tramite grafico a violino e a barre;
* Analisi degli studios che hanno prodotto più film (e serie tv o anime) bubble chart, grafici combinati, regplot e scatterplot;
* Confronto valutazione media con numero di film prodotti per anno durante gli anni tra il mercato USA e quello francese con grafici combinati;
* Confronto per anno di studios attivi negli stati uniti e film prodotti per anno con grafici di correlazione;
* Analisi della valutazione media durante gli anni per genere dei film statunitensi, usando un grafico a linee interattivo;
* Analisi della valutazione media dei generi USA vs resto del mondo tra il 2000 ed il 2020, usando grafici a linee comparativi.

**Problemi**

I problemi riscontrati sono stati: l’utilizzo di una mappa personalizzata per geopandas, l’uso di grafici animati con la loro personalizzazione e riuscire a trovare dei dati che rispondessero alle domande poste all’inizio del progetto.

**Requisiti**

Bisogna installare la mappa al link [www.naturalearthdata.com/downloads/10m-cultural-vectors/](http://www.naturalearthdata.com/downloads/10m-cultural-vectors/) per utilizzare il grafico con geopandas. Inoltre si è deciso di utilizzare come tecniche e documentazioni quelle fornite durante le lezioni più l’utilizzo di intelligenza artificiale per consultazione sulle personalizzazioni ulteriori dei grafici.

**Conclusioni e divisione del lavoro**

Nonostante il dataset non permettesse una risposta definitiva e univoca, è servito per avvicinarsi di qualche passo alla risposta delle domande poste all’inizio del progetto. Nel documento viene correlato un articolo molto interessante che fa una overview sul mercato cinematografico statunitense molto correlato alle domande poste in precedenza.

Il progetto è stato svolto in singolo quindi non c’è stata una divisione del lavoro.

**Note**

Ho provato a includere ricerche e analisi che avvalorassero le mie tesi, però prevedevano un contributo pecuniario abbastanza alto anche solo per leggerle.

**Bibliografia**

<https://chatgpt.com/>

<https://www.w3schools.com>

<https://informatica.i-learn.unito.it/course/view.php?id=2950>

<https://www.goldenwaymediafilms.com/movies-dropped-in-quality/>