Capitolo 1

Metodologia

1.1 Data Acquisition

In uno studio di Data Analysis, lo step di Data Acquisition è il primo della pipeline da seguire. In questa fase abbiamo raccolto i dati necessari dal dataset [**?**, MovieLens]. I file sono stati raccolti in una cartella denominata `ml-25m` contente un README.

1.2 Dataset

Movielens è un Recomendation System per contenuti multimediali, quali film, serie tv, documentari ecc. Movielens mette a disposizione degli sviluppatori un dataset opensource, generator dal database TMDB (The Movie Database). Questo datasedt contiene recensioni con voti e tag di oltre 60.000 film, raccolte da oltre 150.000 utenti durante il periodo che va dal 1995 fino al 2019.
Il dataset è composto dai seguenti file:

- *genome-scores.csv*: contiene il *relevance score* di ogni tag per tutti i film (ovvero, quanto un tag è importante per il dato film).
- *genome-tags.csv*: contiene tutti i tag presenti all'interno del dataset.
- *links.csv*: contiene gli ID di ogni film per i due database TMBD e IMDB.
- *movies.csv*: contiene i titoli dei film, con i rispettivi generi.
- *ratings.csv*: contiene più di 25.000.000 votazioni provenienti dalle recensioni degli utenti.

- *tags.csv*: contiene i tag che sono stati assegnati ai film dagli utenti nelle rispettivi recensioni.

Per il nostro caso di studio, non abbiamo fatto uso di tutti i file disponibili, ma esclusivamente quelli che abbiamo considerato adatti per lo scopo. Di seguito viene mostrata la struttura dei tali.
Il dataset ***ratings*** eèstato selezionato per poter far uso dei voti assegnati ai film dagli utenti, per poter quindi ottenere il voto mediio di tali film. La struttura è la seguente

userId	**movieId**	**rating**	**timestamp**
1	296	5.0	1147880044
1	306	3.5	1147868817
1	307	5.0	1147868828
1	665	5.0	1147878820
1	899	3.5	1147868510
...	...	...	...

Tabella 1.1: ratings.csv

Il dataset ***genome-scores*** è stato selezionato per poter identificare le relazioni che ci sono tra un dato voto ed i valori dei tag assegnati

movieId	**tagId**	**relevance**
1	1	0.028749999999999998
1	2	0.023749999999999993
1	3	0.0625
1	4	0.07574999999999998
1	5	0.14075
...	...	...

Tabella 1.2: genome-scores.csv

Infine, si è fatto uso del dataset ***movies*** per ottenere le informazioni inerenti al genere di ogni film.

movieId	title	genres
1	Toy Story (1995)	Adventur|Animation|...
2	Jumanji (1995)	Adventure|Children|...
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama|...
...	...	...

Tabella 1.3: movies.csv

1.3 Data Visualization

Nella fase di *data visualization* andremo a visualizzare alcune proprietà e peculiarità del dataset.

Distribuzione dei ratings

Nella seguente figura, è mostrata la distribuzione dei voti sul dataset.

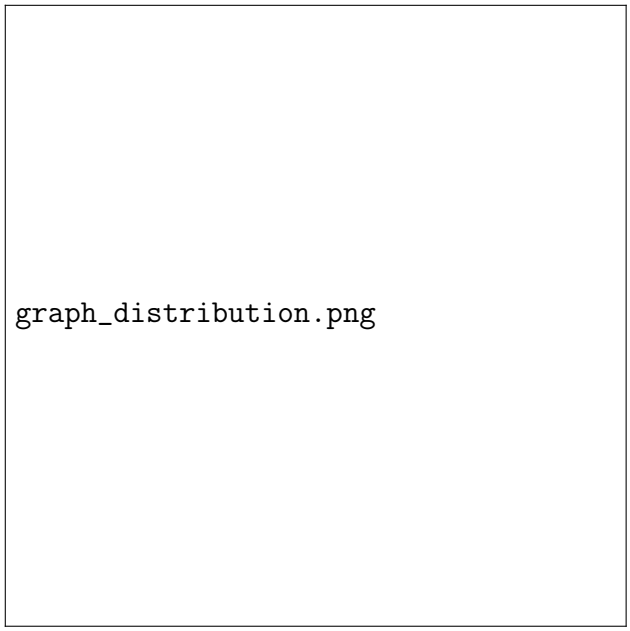

Figura 1.1: Distribution of Ratings.

Il gafico mostra la distribuzine dei vbti sul dataset, evidenziando un chiaro sbilanciamento nella distribuzione. In particolare, si può osservare una concentrzione significativamente maggiore di voti con valore di 3 o superiore rispetto ai voti inferiori. Questo sbilanciamento è ulteriormente evidenziato

dal fatto che i voti con valore di 3 o superiore sono 6 o più volte maggiori rispetto ai voti inferiori. Tale distribuzione suggerisce una maggiore prevalenza di valutazioni positive rispetto a quelle negative o neutre nel dataset. Questo risultato potrebbe essare utile per comprendere meglio le caratteristihe del dataset, ad esempio potrebbe suggerire la presenza di una tendenza positiva nelle valutazioni, o la necessità di bilancuiare meglio el categorie di voto.

Relazione tra numero di voti e voto medio

Figura 1.2: Relationship between Number of Ratings and Average Rating.

Voto medio per Genere

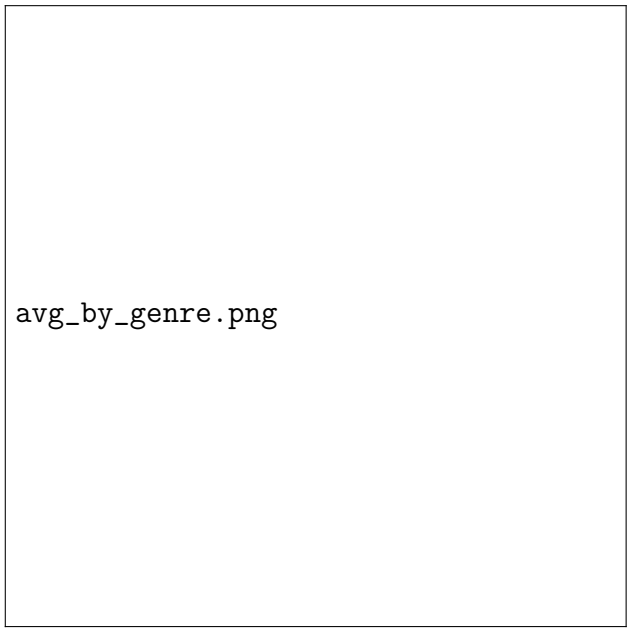

Figura 1.3: Average Ratings by Genre.

1.4 Tabular Data