A Deep Dive into Film Box-Office Data

Alberto Odierna (907341), Biagio Spiezia (920172), Luca Bazzetto (907034)

2024

Indice

1 Introduzione

Il cinema è da sempre una delle forme più popolari di intrattenimento, con milioni di persone che frequentano le sale cinematografiche per immergersi ed apprezzare le storie, che i registi traducono in esperienze visive narrative che suscitano fantasia e curiosità in tutto il mondo. Il botteghino, o gli incassi dal box office, giocano un ruolo cruciale nel finanziare l'industria cinematografica e nel determinare il successo commerciale di un film. Questi incassi rappresentano il denaro guadagnato dalla vendita dei biglietti per assistere alle proiezioni nelle sale cinematografiche. Il successo del botteghino dipende da diversi fattori come la qualità del film, il suo appeal per il pubblico, il periodo in cui esce, la simultanietà di uscita con altri film o l'efficacia delle strategie di marketing e pubblicità. I profitti derivanti dal botteghino possono influenzare il finanziamento di futuri progetti cinematografici, dando vita a prequel, sequel e franchise che continuano ad intrattenere gli spettatori nel corso del tempo.

Due delle principali piattaforme italiane specializzate nella critica cinematografica sono MyMovies e ComingSoon, apprezzate per la loro affidabilità, completezza e la vasta gamma di contenuti offerti che variano dalle recensioni alle diverse news e aggiornamenti forniti sul mondo dell'intrattenimento. MyMovies offre servizi agli utenti come la possibilità di creare una lista di film preferiti, di votare, recensire film e partecipare alle discussioni sulla cinematografia. ComingSoon, invece, è particolarmente noto per fornire aggiornamenti tempestivi sui film in uscita, anteprime esclusive ed interviste con attori e registi. Nel panorama internazionale, invece, una delle piattaforme più influenti e affidabili nel settore dell'intrattenimento è IMDb, la quale nel corso degli anni è diventata una risorsa indispensabile per professionisti del settore e appassionati di cinema e televisione.

Lo scopo di questo studio, dunque, è quello di creare un dataset in cui vengono raccolti e integrati dati da queste piattaforme al fine di confrontarli per ricercare differenze o analogie.

2 Acquisizione dati

L'acquisizione dei dati costituisce la fase iniziale del nostro progetto. Questo processo è fondamentale per ottenere tutte le informazioni necessarie. La raccolta dei dati viene effettuata tramite l'esecuzione dello script `scraper.py`, il quale è stato progettato per estrarre automaticamente le informazioni desiderate. In particolare permette di specificare due parametri, ovvero il paese di interesse e le stagioni cinematografiche. Un esempio di utilizzo è riportato di seguito:

```
scraper.py --countries usa --seasons 2001 2005 2010
```

Questo comando, in particolare, è configurato per estrarre i dati relativi ai film con i maggiori incassi nelle stagioni del 2001, 2005 e 2010 negli Stati Uniti. In questo modo, è possibile ottenere un insieme di dati dettagliato e specifico per le analisi successive.

2.1 MyMovies

La prima fase della raccolta dati è stata condotta tramite una tecnica di web scraping applicata alla sezione box office del sito MyMovies. Questo processo ha impiegato la libreria BeautifulSoup per estrarre in modo sistematico le informazioni presenti sulle pagine web.

Attraverso l'uso di BeautifulSoup, sono state ottenute le seguenti informazioni per ciascun film elencato:

- Titolo
- Titolo in lingua originale (se disponibile)
- Regista

- Anno di produzione
- Anno del box office
- Paese del box office
- Paesi di produzione
- Genere
- Posizione in classifica
- Incasso totale
- Link alla recensione del film

Per ogni anno, dal 2000 al 2023, sono stati selezionati e analizzati i 100 film con i maggiori incassi negli Stati Uniti.

Inoltre, utilizzando il link alla recensione del film ottenuto durante lo scraping iniziale, sono stati ulteriormente estratti i dati relativi al punteggio del film e al numero di votanti. Anche questa operazione è stata effettuata con BeautifulSoup, completando così il set di dati raccolti dal sito MyMovies.

2.2 ComingSoon

La seconda fase della raccolta dati è stata realizzata utilizzando una combinazione delle librerie Selenium e BeautifulSoup. Questa fase si concentra sulla ricerca del punteggio e del numero di voti per ciascun film presenti sul sito ComingSoon. I passaggi seguiti sono i seguenti:

- Se è disponibile il titolo in lingua originale, la ricerca viene effettuata utilizzando questo titolo.
- Se il titolo in lingua originale non è disponibile, la ricerca viene effettuata utilizzando il titolo in italiano.

Per eseguire queste ricerche, viene utilizzata la libreria requests per formulare una query direttamente nel motore di ricerca Google, includendo il titolo del film seguito dalla parola chiave "comingsoon". La pagina dei risultati di Google viene poi analizzata per individuare nella preview del risultato il punteggio e il numero di voti.

Qualora questa ricerca non produca risultati soddisfacenti, viene effettuata una ricerca direttamente all'interno dell'archivio del sito ComingSoon.it. In questo caso, viene utilizzata la libreria Selenium per gestire i cookie e il form di ricerca presenti sul sito. Selenium consente di navigare e interagire con la pagina web in modo dinamico, facilitando così l'estrazione dei dati necessari.

2.3 IMDb

La terza fase di estrazione dati è stata completata tramite l'utilizzo di API. In particolare, per ottenere il punteggio e il numero di voti su IMDb, è stata sfruttato il progetto open source OMDb API.

Il processo di raccolta avviene nel modo seguente: utilizzando la libreria requests, viene effettuata una chiamata all'endpoint dell'API. I parametri passati nella chiamata includono il titolo del film e l'anno di produzione, garantendo così l'esattezza e la specificità dei risultati ottenuti. Si estrapolano poi i dati necessari e si aggiungono alle informazioni del film.

Questa metodologia consente di accedere rapidamente e accuratamente a un vasto database di informazioni relative ai film, permettendo di integrare nel nostro progetto i dati riguardanti il punteggio e il numero di voti registrati su IMDb. La scelta di utilizzare OMDb API facilita l'automazione del processo di raccolta dati, riducendo significativamente il tempo e le risorse necessarie rispetto a metodi manuali o semi-automatici.

2.4 Problematiche riscontrate

Inizialmente, per la ricerca dei voti, veniva utilizzato esclusivamente Selenium. Questo strumento veniva impiegato per eseguire ricerche su Google, inserendo il titolo del film insieme al nome del regista e all'anno di produzione per ottenere risultati più precisi. Dai risultati di Google, era possibile estrarre il voto del film tramite la locandina visualizzata. Tuttavia, questo approccio presentava diverse problematiche:

- Incoerenza delle fonti di voti: i risultati provenivano da varie fonti (IBS, Feltrinelli, Movieplayer, ecc.), il che significava che non sempre le informazioni erano uniformi.
- Mancanza di risultati: in alcuni casi, la ricerca non restituiva alcun risultato utile.
- Efficienza: Selenium richiede un notevole consumo di tempo e risorse rispetto a BeautifulSoup.

Per queste ragioni, si è cercato di limitare l'utilizzo di Selenium, impiegandolo solo quando strettamente necessario.

Un altro problema significativo riguardava l'ottenimento dei dati da IMDb, uno dei più importanti siti di recensioni cinematografiche al mondo. La prima opzione considerata è stata l'uso delle API ufficiali di IMDb, che però comportavano costi aggiuntivi.

La seconda opzione esplorata è stata l'uso di un dataset disponibile su Kaggle. Questa soluzione è stata approfondita e, dopo aver trovato un dataset che includeva i film per gli anni di nostro interesse, sono emersi problemi durante la fase di integrazione. Questi problemi erano dovuti a una rielaborazione dei dati da parte del proprietario del dataset, come ad esempio la riscrittura dei

titoli dei film. Per risolvere questo problema, si è tentato di utilizzare librerie di RecordLinkage per associare i titoli dei film, ma le basse percentuali di associazione ottenute hanno portato a scartare questa opzione.

Un'altra opzione esplorata è stata l'uso di un dataset disponibile su Kaggle. Questa soluzione è stata esaminata attentamente e, dopo aver trovato un dataset che includeva i film degli anni di nostro interesse, sono emersi problemi durante la fase di integrazione. Questi problemi erano dovuti alla rielaborazione dei dati da parte del proprietario del dataset, come la riscrittura dei titoli dei film. Per risolvere questo problema, si è tentato di utilizzare librerie di RecordLinkage per associare i titoli dei film, ma le basse percentuali di associazione ottenute hanno portato a scartare questa opzione.

L'opzione finale, descritta nel paragrafo precedente, consisteva nell'uso delle API di OMDb. Questa soluzione ha un'unica limitazione: un limite di utilizzo di 1000 richieste al giorno. È quindi necessario gestire attentamente le richieste per non superare questo limite giornaliero.

3 Archiviazione dei dati

L'archiviazione dei dati è stata progettata utilizzando una collection all'interno di MongoDB, in cui ogni film viene salvato come un documento distinto. La figura 1 mostra un esempio di documento (film "Cattivissimo me") presente all'interno della collection.

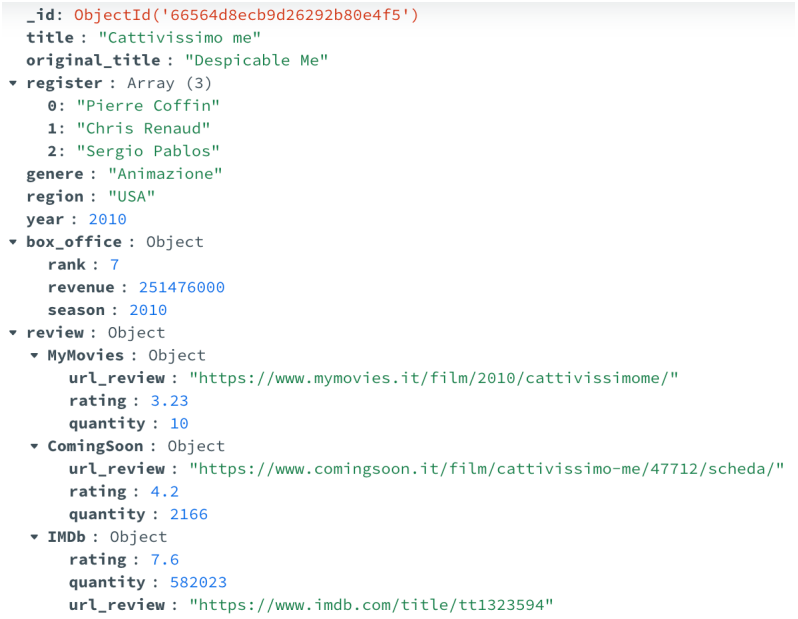

Figura 1: Esempio di documento presente all'interno della collection

Inizialmente, erano state create tre diverse collections: una contenente i dati generali del film, una per i dati del box office e una per le recensioni. Tuttavia, questa soluzione è stata abbandonata a favore di una struttura più integrata per le seguenti motivazioni:

- Il download dei dati non è frequente: i dati vengono scaricati una sola volta e includono tutte le informazioni necessarie. Durante l'anno alcune recensioni possono a causa di nuovi voti da parte degli utenti, un aggiornamento annuale dei dati è sufficiente.
- Semplificazione delle query: con la struttura attuale, è più semplice eseguire query di ricerca senza dover applicare diversi join per ottenere informazioni semplici. Questa integrazione facilita l'accesso e la gestione dei dati.
- Flessibilità del modello: questa tipologia di modello consente di includere o escludere attributi in base alle informazioni disponibili durante le ricerche. Ad esempio, se non si riesce a ottenere un certo dato per un film specifico, il modello non viene compromesso.

4 Arricchimento dei dati

Nel processo di ricerca e raccolta dei dati relativi a un film, ogni sito ha il proprio modo di memorizzare i titoli, il che può causare discrepanze significative. Queste differenze possono includere variazioni nei caratteri speciali, nella formattazione e nei titoli tradotti. Di seguito sono riportati alcuni esempi di queste varianti:

MyMovies	**ComingSoon**	**IMDb**
Maleficent - Signora del male	Maleficent 2: Signora del Male	Maleficent: Mistress of Evil
Gli Incredibili - Una "normale" famiglia di supereroi	Gli Incredibili - Una normale famiglia di supereroi	The Incredibles

Tabella 1: Esempi di titoli nei diversi siti

Per risolvere questi problemi e assicurare una maggiore coerenza nei dati, sono state adottate diverse strategie di arricchimento dei dati:

- Ricerche multiple per ogni film: Ogni film è stato ricercato utilizzando sia il titolo originale che quello italiano, per garantire una copertura completa delle varianti di titolo.
- Rimozione dei caratteri speciali: I caratteri speciali presenti nei titoli sono stati rimossi per uniformare i dati e facilitare le operazioni di confronto.

- Aggiunta di informazioni contestuali: Per migliorare l'accuratezza delle ricerche, sono state aggiunte informazioni supplementari come l'anno di produzione e il nome del regista. Questi dettagli aiutano a distinguere tra film con titoli simili o identici.

Queste misure hanno permesso di migliorare significativamente la qualità e l'affidabilità dei dati raccolti, assicurando che le informazioni sui film siano coerenti e complete, indipendentemente dalla fonte.

5 Queries

Una volta acquisite, integrate e memorizzate tutte le informazioni necessarie, è iniziata la fase di interrogazione del database. In questa fase, l'attenzione è stata rivolta a rispondere alle domande più rilevanti e interessanti per il progetto. Sono state quindi formulate una serie di query mirate per analizzare diversi aspetti dei dati raccolti, permettendo di avere una visione più chiara e approfondita delle dinamiche in gioco.

Il primo punto approfondito è stato il confronto tra la recensione media dei top 10 film e quella dei bottom 10. Per raggiungere questo obiettivo, sono stati selezionati, per ogni anno, i primi 10 film e successivamente è stata calcolata la valutazione media per ciascun sito. Lo stesso procedimento è stato applicato anche agli ultimi 10 film. Infine, è stata calcolata la media totale delle recensioni, sempre categorizzata per sito.

Il primo punto approfondito è stato il confronto tra la recensione media dei film con i 10 maggiori incassi e quelli con i 10 minori. Per raggiungere questo obiettivo, sono stati selezionati, per ogni anno, i primi 10 film e successivamente è stata calcolata la valutazione media per ciascun sito. Lo stesso procedimento è stato applicato anche ai 10 film con il rank più basso. Infine, è stata calcolata la media totale delle recensioni, sempre categorizzata per sito.

I risultati ottenuti sono i seguenti:

Categoria	**MyMovies**	**ComingSoon**	**IMDb**
Top 10	2.74	3.85	6.43
Bottom 10	2.74	3.85	6.43
Totale	2.83	3.84	6.52

Tabella 2: Media delle recensioni per ciascun anno

Cosa si può dedurre da questi risultati? La cosa più evidente è che la media delle recensioni dei film nella top 10 è simile a quella dei film nella bottom 10. Questo suggerisce che non c'è una differenza significativa nella qualità dei film. Pertanto, il ranking basato sugli incassi non è un indicatore affidabile della qualità dei film secondo le recensioni.

Le medie totali delle valutazioni mostrano valori leggermente superiori rispetto alle medie dei top 10 e bottom 10 per ciascun anno. Questo potrebbe

indicare che ci sono film che non rientrano né nei top 10 né nei bottom 10, ma che ottengono recensioni leggermente migliori.

Analizzando le piattaforme, IMDb ha un valore medio di recensione (6.52 in totale) su una scala in decimi, mentre MyMovies e ComingSoon hanno medie rispettivamente di 2.83 e 3.84 su una scala in quinti. Effettuando un confronto orizzontale e tenendo conto delle diverse scale di valutazione, si può dedurre che in media le recensioni su ComingSoon sono più alte, mentre su MyMovies sono più basse. Questo potrebbe indicare una tendenza dei recensori di ComingSoon ad essere più indulgenti, o potrebbe riflettere una diversa scala di valutazione.

Successivamente, è stata effettuata un'analisi per identificare i generi cinematografici più comuni, concentrandosi su quelli che appaiono più di 100 volte. Per ciascuno di questi generi, è stato calcolato l'incasso medio. Vedi Tabella 3

Genere	**Incasso Medio**	**Frequenza**
Animazione	130,138,989.4	214
Azione	127,643,513.0	374
Commedia	59,821,164.8	419
Horror	53,307,530.4	172
Thriller	50,782,694.9	119
Drammatico	43,768,377.5	308

Tabella 3: Incassi medi e frequenza per genere cinematografico

Osservando i dati, possiamo dedurre diverse conclusioni interessanti. I generi con la frequenza più alta sono azione, commedia e drammatico. Questo indica che questi generi sono frequentemente presenti in top 100. I generi di animazione, azione e commedia hanno gli incassi medi più alti. In particolare, i film di animazione e azione si distinguono per i loro incassi significativamente superiori rispetto agli altri generi.

Un'altra domanda affrontata è stata quella di identificare i dieci registi che, dal 2000 al 2023, hanno avuto il maggior numero di film in top 100. Inoltre, è stato calcolato l'incasso medio dei loro film. I risultati sono stati riportati in Tabella 4.

Regista	Numero di Film	Incasso Medio
Steven Spielberg	16	104,800,670.5
Clint Eastwood	14	78,903,756.4
Ridley Scott	12	80,358,971.2
M. Night Shyamalan	10	80,591,666.1
Steven Soderbergh	10	57,732,247.3
Shawn Levy	10	116,387,148.7
Peter Jackson	9	235,784,008.0
Tim Burton	9	121,847,492.6
Michael Bay	9	178,607,831.7
Ron Howard	8	94,493,569.4

Tabella 4: Numero di film e incasso medio per regista (2000-2023)

La tabella riporta i nomi di alcuni dei più noti e grandi registi internazionali. La conclusione più intuitiva è che Steven Spielberg è il regista che, nei 23 anni esaminati, ha prodotto 16 film in grado di rientrare nei top 100 per incasso. Uno spunto interessante per future analisi sarebbe calcolare il rapporto tra i film in top 100 e il totale dei film prodotti da ciascun regista.

Un'altra osservazione meno intuitiva, poiché la tabella non è ordinata per incassi medi, è che Peter Jackson si distingue come il regista con l'incasso medio più alto, con una differenza significativa rispetto agli altri registi in classifica.

Infine, è stata effettuata una selezione dei 10 film con i maggiori incassi di tutti i tempi, ordinandoli in base ai guadagni. Per ciascuno di questi film, è stata anche annotata la stagione d'incasso. Vedi Tabella 5

Titolo	Incasso	Stagione
Star Wars: Episodio VII	935,518,389	2015
Avengers: Endgame	858,365,685	2019
Spider-Man: No Way Home	804,395,488	2021
Avatar	749,446,000	2022
Top Gun: Maverick	718,519,000	2022
Black Panther	700,004,026	2018
Avatar: La via dell'acqua	684,060,555	2022
Avengers: Infinity War	678,807,703	2018
Titanic	659,328,801	2023
Jurassic World	652,175,130	2015

Tabella 5: I 10 film con i maggiori incassi di tutti i tempi e le loro stagioni di incasso.

Tra questi, "Star Wars: Episodio VII - Il risveglio della forza" si distingue come il film con l'incasso più alto di sempre, con un guadagno complessivo di oltre 935 milioni di dollari. Questo risultato supera di ben 80 milioni il secondo classificato, "Avengers: Endgame".

6 Data Quality

In seguito al caricamento dei dati nel database, si è passati alla valutazione della qualità dei dati. Per tale valutazione, è stato scelto di utilizzare come misura di riferimento la completeness, che rappresenta il grado di copertura con cui una determinata variabile o un fenomeno è presente all'interno dei dati. Come primo passo, è stato calcolato il valore della completeness relativo al database generale.

Descrizione	**Valore**
Numero totale di documenti	2010
Numero di documenti completi	1665
Percentuale di completezza	82.84%

Tabella 6: Table Completeness

Il database contiene informazioni su un totale di 2010 film, dei quali 1665 presentano informazioni complete. Questo dato evidenzia una buona percentuale di completeness, con l'82,84% dei documenti che contengono tutte le informazioni richieste. Successivamente, entrando nello specifico, è stata calcolata l'attribute completeness.

Campo	**Percentuale di completezza**
Titolo	100.00%
Box Office Revenue	100.00%
Register	99.95%
Genere	100.00%

Tabella 7: Percentuali di completezza del database sui film per vari campi

La percentuale di completeness relativa al titolo, al genere e alla revenue è massima, il che indica che tutti i film presenti nel database non presentano valori nulli per queste variabili. La percentuale è leggermente inferiore per la variabile regista, con solo pochi film che non dispongono di informazioni su di esso. Inoltre, è stata calcolata anche l'attribute completeness relativa alle valutazioni dei tri siti.

Fonte	**Valutazione (%)**
ComingSoon	100.0
IMDb	86.56
MyMovies	100.0

Tabella 8: Attribute Completeness relativa alla variabile review

I giudizi relative a ComingSoon e MyMovies hanno una completezza massima pari al 100%, il che indica che ciascun film del database contiene le recensioni relative a queste fonti. D'altra parte, IMDb ha una completezza leggermente

inferiore, con l'86,57%, una percentuale leggermente inferiore. Tuttavia, essa indica comunque una buona e valida raccolta di informazioni, suggerendo che la maggior parte dei film nel database ha una recensione disponibile anche su IMDb.

Attualmente, questo database non viene aggiornato costantemente, il che limita la capacità di valutare una misura temporale di accuratezza. Tuttavia, per futuri sviluppi con aggiornamenti automatici che includono l'inserimento annuale di tutti i titoli rilasciati o l'aggiornamento delle recensioni associate ai film, potrebbe diventare possibile valutare anche una misura di accuratezza temporale come la currency. Quest'ultima misura quantifica quanto rapidamente i dati sono aggiornati rispetto ai cambiamenti reali nel cinema, migliorando così la pertinenza e l'affidabilità del database nel tempo.

7 Sviluppi futuri

Il progetto propone la creazione di un archivio che raccolga tutte le informazioni sui principali titoli cinematografici, con l'obiettivo di individuare e analizzare sia l'impatto economico sia l'opinione della critica per ciascun titolo. Ci sono diversi modi in cui il progetto potrebbe essere ulteriormente migliorato:

- ampliare ciclicamente il dataset con i nuovi film usciti;
- aggiornare il dataset con le nuove recensioni inserite dagli utenti;
- integrare il dataset con dati provenienti da piattaforme di streaming come Netflix e Disney+.
- aggiungere nuove nazioni come sorgenti di dati per il box office;
- migliorare il database per una maggiore effiecienza nelle queries;
- implementare una migliore gestione degli errori;
- integrare strumenti di visualizzazione dei dati per fornire approfondimenti grafici sui dati raccolti.