Relazione Tecnico Scientifica
THE MOVIE DATASET

Julia Bui Xuan – 882385
Lorenzo Longo – 846738
Veronica Morelli – 839257

INTRODUZIONE

Cosa si può dire del successo di un film prima che venga prodotto? Esistono delle formule specifiche per raggiungerlo? È possibile predire se un film avrà una valutazione alta, considerando alcune sue caratteristiche? Sono queste le domande che ci siamo posti e ci hanno spinto alla creazione del dataset. In quanto il loro costo di produzione, nella maggior parte dei casi, supera ben oltre 100 milioni di dollari, saper rispondere a queste domande può portare informazioni molto utili e importanti per le produzioni cinematografiche, ma rappresenta anche una sorgente di grande interesse per i fans.

OBIETTIVO

L'obiettivo di questo progetto consiste nel creare un dataset contenente informazioni dei film più famosi dal 2010 fino ai tempi d'oggi. Oltre ai titoli dei film, quindi, si vogliono aggiungere caratteristiche riguardanti gli attori, con il corrispettivo ruolo interpretato, ma anche i premi vinti, le valutazioni e gli incassi ottenuti, gli studi di produzione cinematografica che li hanno prodotti e i direttori coinvolti, i paesi di produzione, i generi a cui appartengono, le parole chiave e le recensioni associate. Inoltre, le lingue disponibili, la loro durata e una lista di film simili. L'utente finale a cui questo database è destinato è colui interessato ad analizzare e visionare i film, in modo da poter fare ulteriori approfondimenti e confronti, individuando caratteristiche rilevanti o meno. La presenza delle recensioni, inoltre, può essere usata come punto di partenza per esplorazioni future come la sentiment analysis e topic extraction.

Il progetto si sviluppa nelle seguenti fasi: **Data Acquisition**, **Data Storage**, **Data Cleaning**, **Data Integration**, **Data Quality** e **Data Exploration**.

DATA ACQUISITION: API e WEB SCRAPING

Le tecniche utilizzate per l'acquisizione dati sono le seguenti: **API** e **Web Scraping**.

1. **API** (Application Programming Interface)

   La prima fonte dati è stata ottenuta tramite API. L'obiettivo é quello di avere un insieme di informazioni rilevanti a descrivere un film. In particolare, al termine di questa prima fase di scaricamento dati, si otterrà un database documentale contenente 14365 documenti, ognuno di essi relativo ad un film. Gli attributi per ogni film si possono trovare nell'*Appendice A - Descrizione Attributi Film IMDB prima della fase di Data Cleaning*. Negli step successivi il dataset subirà fasi di data cleaning, assessment data quality e infine verrà arricchito con dati provenienti da una seconda fonte dati. La fase di acquisizione dati è

eseguita in due fasi diverse, una consequenziale all'altra. La documentazione relativa all'utilizzo delle API dal sito IMDB si trova al seguente link: https://imdb-api.com/api.
Nel primo step si sfruttano le funzionalità dell'**API Advanced Search** reperibile al sito: https://imdbapi.com/api#AdvancedSearch-header.
Nel secondo step si sfruttano le funzionalità dell'**API Title** reperibile al sito: https://imdb-api.com/api#Title-header. Per utilizzare le API si è dovuto creare un account IMDB dal quale si è ottenuta una chiave, chiamata API Key. Si specifica che per ogni chiave si possono fare solo 100 richieste al giorno, di conseguenza, nel secondo step si sono dovuti creare diversi account, quindi diverse API Key per riuscire a scaricare tutti i dati in un tempo ragionevole.

FASE 1 - API Advanced Search

Tramite l'API Advanced Search è possibile scaricare un insieme di film definendo diversi parametri. L'obiettivo del primo step è quello di ottenere un insieme di film popolari su IMDB dal 2010 al 2021. Si sceglie di scaricare 100 film (*feature film*) per ogni mese in base al punteggio di popolarità del film, di conseguenza si selezionano i 100 film più popolari in ordine decrescente (totale 1200 film per anno). Considerando che il periodo temporale è di 12 anni e che per ogni anno si scaricano informazioni su 1200 film, il dataset finale consiste in 14400 film (*Dataset 1*). Nel nostro caso sono presenti alcuni film duplicati e il dataset finale consiste in 14365 documenti. L'API utilizzata è la seguente:
https://imdb-api.com/API/AdvancedSearch/k_1234567/title_type=feature&count=100&sort=moviemeter,desc
Nell'*Appendice B - Esempio Output API Advanced Search* si mostra un esempio di output in formato json relativo ad un film. Come si può notare, l'output non restituisce sufficienti informazioni. L'API che fornisce esaustivi dati riguardo al film è l'API Title. Tale API richiede in input l'id del film. Nel *Dataset 1* ogni documento rappresenta un film e contiene un attributo chiamato *title* relativo al titolo e un attributo *id* relativo all'identificativo univoco del film. Da *Dataset 1* si estraggono gli id che verranno utilizzati in input nello step 2 per l'API Title.

FASE 2 - API Title

Come descritto precedentemente, ogni film è definito univocamente da un id. L'API Title è utilizzabile tramite l'id. L'output dell'API Title è un file json contenente numerose informazioni relative ad un film. L'API è definita come segue:
https://imdb-api.com/en/API/Title/k_12345678/tt1375666
Nell'*Appendice B - Esempio Output API Title* si mostra un esempio di output.
Si ricorda che gli id ottenuti dalla fase precedente sono 14365. Inoltre, si ricorda che ogni API Key può essere utilizzata per 100 richieste al giorno. Per ottenere informazioni su tutti i 14365 film con una sola API Key ci vorrebbero 144 giorni. Di conseguenza, per riuscire a scaricare i dati in un tempo ragionevole si sono creati 23 account, ottenendo così 23 API Key. Gli account su IMDB sono stati creati utilizzando mail di amici e famigliari. Con 23 chiavi si possono fare 2300 richieste al giorno. Di conseguenza, i dati sono scaricabili in 7 giorni.

DATA STORAGE PRIMA FONTE DATI

L'output della Fase 1 consiste in una lista di id, ognuno relativo ad un film. La lista è immagazzinata in locale in un file testo (estensione .txt). L'output della Fase 2 consiste in un unico database documentale in formato json contenente 14365 documenti, ognuno relativo ad un film. Per ottenere questo dataset completo si sfruttano le funzionalità della libreria python `pymongo` come segue. Da python ci si collega a **MongoDB** creando un database e una collezione nella quale immagazzinare i dati. Dopo questa operazione si eseguono le richieste di dati con l'API Title. Ogni volta che una richiesta viene eseguita si immagazzina il risultato su MongoDB tramite la funzione `collect.one()`. Questa procedura permette di immagazzinare informazioni in un database documentale che viene continuamente aggiornato da nuovi dati richiesti tramite API. Considerando che MongoDB limita il database ad una dimensione di 17MB, si creano due diversi database nei quali immagazzinare dati. Alla fine della fase di storage si ottiene un **dataset documentale** in formato **json**. Le prossime fasi che seguono sono il Data Cleaning e l'Assessment della Data Quality. Una volta che il dataset sarà sistemato potrà essere utilizzato nella fase di Data Integration e Data Exploration.

2. Web Scraping

La seconda fonte dati, è estratta attraverso Web Scraping. Più precisamente, questa tecnica è applicata alla pagina web *https://www.rogerebert.com*, dove sono riportate le recensioni ufficiali del critico cinematografico statunitense Rogert Ebert. Si tratta di una pagina web con "scrolling infinito"; i contenuti, infatti, sono pubblicati all'interno di una stessa pagina senza essere costretti a generarne nuove per diversi film. L'esplorazione della pagina da parte dell'utente è facilitata, mentre l'operazione di scraping è più complessa. Attraverso un'ispezione dettagliata della pagina sorgente siamo riusciti estrarre tutte le recensioni. La libreria utilizzata è `BeautifulSoup`, finalizzata all'estrazione dati da file HTML e XML. Inoltre, è stata anche installata la libreria `requests` in modo da poter eseguire richieste HTTP attraverso codice Python.
La prima iterazione viene fatta sul numero di pagine: sono stati scaricati i nomi di tutti i film con data di rilascio dal 2010, disponibili sul sito web, insieme ai corrispondenti link che portano alla loro recensione completa. In questo modo, per ogni pagina sono stati ottenuti 28 film, di cui gli ultimi 4 apparivano essere sempre gli stessi, ovvero "Top Gun: Maverick", "Interceptor", "RRR", "Eiffel". Questi ultimi infatti, appartenevano alla categoria di film con "Popular Reviews". Sono stati rimossi i duplicati.
La seconda iterazione viene fatta sui link ottenuti nel punto precedente: vengono scaricate tutte le recensioni. Esse sono composte da diversi paragrafi, per cui è stato necessario unire i *tag* relativi attraverso un'ulteriore iterazione.

DATA STORAGE SECONDA FONTE DATI

Si è creato un **database relazionale** in formato **csv** contenente 6525 titoli di film con la recensione corrispondente. Il database è stato immagazzinato in locale sul computer personale.

DATA CLEANING – PRIMA FONTE DATI

Dopo aver immagazzinato i dati su MongoDB, si esegue la fase di Data Cleaning per sistemare il dataset e mantenere solo le variabili di interesse. Questa fase è necessaria a rendere il dataset utile per analisi successive.
La prima fonte dati consiste in un database documentale dove ogni documento contiene 46 attributi (*APPENDICE A - Descrizione Attributi Film IMDB prima della fase di Data Cleaning*). Prima di procedere nella fase di data cleaning sugli attributi si nota che alcuni documenti sono nulli e hanno valore dell'attributo *ErrorMessage* diverso da *None*; ciò significa che durante l'acquisizione dati ci sono stati problemi relativi a questi documenti. Considerando che i documenti con errore sono 600, si decide di eliminarli e procedere nelle fasi successive. Per quanto riguarda gli attributi si specifica che molti di essi non sono stati mantenuti perché ritenuti non rilevanti alla descrizione delle caratteristiche di un film. Per esempio, si elimina l'attributo *image* contenente il link dell'immagine copertina del film, l'attributo *id* relativo all'identificativo del film, l'attributo *imDbRatingsVotes* relativo al numero di votanti del film ecc.. In generale, gli attributi mantenuti riguardano il titolo del film, il cast, il genere, le lingue parlate, il paese o i paesi in cui il film è prodotto, la/le compagnie di produzione, la data di rilascio, la durata, i punteggi, i guadagni, la trama, una lista di parole chiave che descrivono il film, una lista di film simili ed eventuali premi vinti. In totale, si conservano 21 attributi rispetto ai 46 iniziali (*APPENDICE A - Descrizione Attributi Film IMDB dopo la fase di Data Cleaning*).
Per molti attributi che si decide di mantenere è necessario apportare delle modifiche. Per esempio, i valori degli attributi *imDbRating, runtimeMins, year* e *metacriticRating* si trasformano da stringa a valori numerici, per poter fare analisi esplorative quantitative in seguito. I valori degli attributi *genres, directors, writers, languages, companies* e *countries* si trasformano da stringa a lista. I valori degli attributi *actorList e similars* sono liste di dizionari: per entrambi si eliminano chiavi non rilevanti, come *image* e *id*. Il valore dell'attributo *boxOffice* è un dizionario formato da 4 chiavi relative al budget e a diversi guadagni ottenuti dal film; questi valori dovrebbero essere dei numeri interi, invece sono stringhe contenenti anche valori alfanumerici come il '$' che indica la valuta. Si utilizzano alcune espressioni regolari per eliminare caratteri alfanumerici, trasformando così i valori da stringhe a interi.

DATA QUALITY - COMPLETEZZA (PRIMA FONTE DATI)

Prima di procedere nella fase di Data Integration, si decide di eseguire la fase di *Data Quality*. Nel primo caso si valuta la dimensione di qualità della **completezza** relativa ai dati mancanti. Sapendo che nei database documentali lo schema è free, non si ha l'obbligo di mantenere i dati mancanti. In questo caso, si vanno a eliminare per ogni documento i valori che sono mancanti. Questa procedura riduce la dimensionalità del database rendendo così più veloci elaborazioni e analisi future.

Per gli attributi *awards*, *genres, directors, writers, languages, companies e countries, title, plot* il dato mancante si ha quando il valore associato alla chiave è una stringa vuota. Per gli attributi *contentRating*, *imDbRating, runtimeMins, year, metacriticRating, releaseDate* il dato mancante si ha se il valore è *None*. Per gli attributi *actorList, keywordList* e *similars* il dato mancante si ha con una lista vuota. Per l'attributo *boxOffice* il dato mancante si ha con un dizionario vuoto. Se l'attributo ha valore mancante si elimina.

Si calcola la percentuale di attributi (*chiavi*) mancanti per ogni documento del database. Si osserva che la maggior parte dei documenti contiene tra le 9 e le 14 chiavi su 19, quindi qualche informazione è sempre mancante. Non ci sono documenti con informazioni sempre complete.

```
Il 0.2% dei documenti contiene 3 chiavi su 19
Il 0.1% dei documenti contiene 4 chiavi su 19
Il 0.2% dei documenti contiene 5 chiavi su 19
Il 1.1% dei documenti contiene 6 chiavi su 19
Il 2.5% dei documenti contiene 7 chiavi su 19
Il 5.3% dei documenti contiene 8 chiavi su 19
Il 8.6% dei documenti contiene 9 chiavi su 19
Il 15.9% dei documenti contiene 10 chiavi su 19
Il 18.9% dei documenti contiene 11 chiavi su 19
Il 18.4% dei documenti contiene 12 chiavi su 19
Il 14.1% dei documenti contiene 13 chiavi su 19
Il 9.7% dei documenti contiene 14 chiavi su 19
Il 4.1% dei documenti contiene 15 chiavi su 19
Il 0.8% dei documenti contiene 16 chiavi su 19
```

Si calcola la percentuale di dati mancanti per ogni attributo. Si nota che gli attributi *fullcast* e *ratings* contengono sempre valori nulli, l'attributo *metacriticRating* è mancante al 99.97%. Gli attributi *title* e *year* non contengono dati mancanti. L'attributo *countries* contiene solo lo 0.27% di dati mancanti. In generale, gli attributi con una bassa percentuale di dati mancanti oltre a quelli già citati sono: *releaseDate, writers, actorList, genres* e *languages.*

```
La chiave 'title' è mancante nel 0.0% dei documenti
La chiave 'plot' è mancante nel 32.39% dei documenti
La chiave 'year' è mancante nel 0.0% dei documenti
La chiave 'releaseDate' è mancante nel 18.35% dei documenti
La chiave 'runtimeMins' è mancante nel 42.73% dei documenti
La chiave 'awards' è mancante nel 95.44% dei documenti
La chiave 'directors' è mancante nel 7.43% dei documenti
La chiave 'writers' è mancante nel 20.02% dei documenti
La chiave 'fullcast' è mancante nel 100.0% dei documenti
La chiave 'ratings' è mancante nel 100.0% dei documenti
La chiave 'actorList' è mancante nel 11.34% dei documenti
La chiave 'genres' è mancante nel 6.24% dei documenti
La chiave 'companies' è mancante nel 39.66% dei documenti
La chiave 'countries' è mancante nel 0.27% dei documenti
La chiave 'languages' è mancante nel 3.34% dei documenti
La chiave 'contentRating' è mancante nel 97.3% dei documenti
La chiave 'imDbRating' è mancante nel 83.89% dei documenti
La chiave 'metacriticRating' è mancante nel 99.97% dei documenti
La chiave 'boxOffice' è mancante nel 52.22% dei documenti
La chiave 'keywordList' è mancante nel 67.65% dei documenti
La chiave 'similars' è mancante nel 93.25% dei documenti
```

DATA QUALITY - CONSISTENZA (PRIMA FONTE DATI)

Come seconda dimensione di qualità si decide di valutare la **consistenza**.
Si elencano le verifiche di consistenza che sono state eseguite:

- L'attributo *releaseDate* è una stringa contenente la data di rilascio del film, il suo formato è YYYY-MM-DD. Si verifica che l'anno sia compreso tra il 2010 e il 2021, poichè questo è l'arco temporale scelto per scaricare i dati. Inoltre, si verifica che il mese sia compreso tra 1 e 12 e il giorno tra 1 e 31. Dato che tali valori sono in formato stringa, si trasformano in intero per poter valutare i vincoli di consistenza. L'output mostra che solo due film non rispettano i vincoli poiché l'anno in cui sono stati rilasciati è il 2022.

- L'attributo *imDbRating* deve essere compreso tra 0 e 10, poiché è su questo intervallo di valori che si esprime la valutazione del film. Nessun film presenta valori fuori da questo range. Un'operazione analoga è stata effettuata per l'attributo *imDbRating* incluso nell'attributo *Similars*. Anche in questo caso il vincolo è rispettato.

- L'attributo *runtimeMins* esprime la durata del film in minuti. "L'Academy of Motion Picture Arts and Sciences, l'American Film Institute, e il British Film Institute definiscono un lungometraggio (feature film) come un film di 40 minuti minimo". Poiché i film sono stati scaricati con il criterio di essere lungometraggi (feature film), si imposta la condizione che essi abbiano una durata maggiore o uguale a 40 minuti. L'output mostra che ci sono 49 film che hanno durata inferiore ai 40 minuti. E' utile quindi indagare il motivo per cui tali film sono stati scaricati nonostante la condizione imposta nell'API.

- L'attributo *budget($)* contenente nell'attributo *boxOffice* esprime il budget stabilito per la produzione di un film. Si vuole vedere quanti e quali film hanno un budget molto basso e minore di 10000 dollari. Un budget così basso potrebbe destare qualche sospetto sul film e sulla sua natura. L'output mostra che 2868 film hanno un budget minore di 10000 dollari.

- L'attributo *boxOffice* è un dizionario contenente massimo 4 chiavi, 3 di esse riguardano i seguenti attributi: openingWeekendUSA relativo al guadagno in dollari nel primo weekend dopo l'uscita del film, grossUSA relativo al guadagno totale in dollari negli Stati Uniti (USA), cumulativeWorldwideGross relativo al guadagno totale in dollari in tutto il mondo. Si vuole verificare che i valori associati a questi attributi siano coerenti con il loro significato. Ciò significa che devono essere rispettate queste disuguaglianze:
  openingWeekendUSA < grossUSA < cumulativeWorldwideGross
  Dal momento che non tutti i film presentano valore non mancante per gli attributi precedentemente citati, si verifica la consistenza anche a coppie di attributi. Dall'output si osserva che nessun film presenta un'incongruenza.

DATA CLEANING E COMPLETEZZA (SECONDA FONTE DATI)

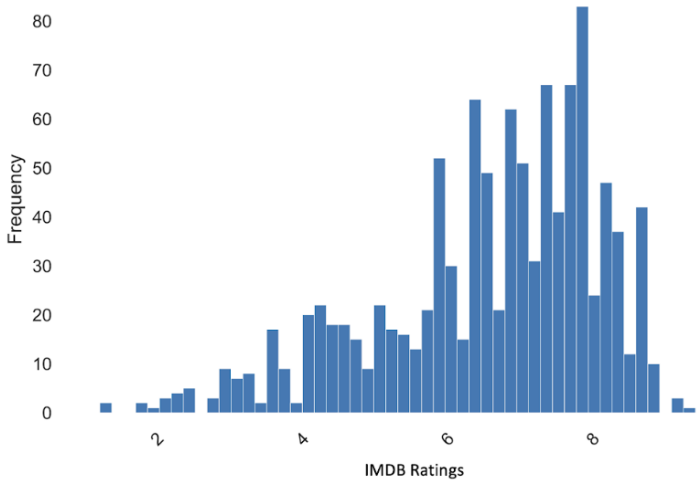
Il database ottenuto dalla seconda fonte dati è un database relazionale in formato csv, composto dalle variabili *Title* e *Review*. Il database non presenta alcun dato mancante. L'unica sistemazione necessaria riguarda il testo nelle recensioni. Dall'operazione di scraping compaiono nel testo dei caratteri da eliminare come *\n* e *\xa0*. Tramite la funzione `replace()` si eliminano tali caratteri.

DATA INTEGRATION

Dopo le fasi di *Data Cleaning* e *Assessment della Data Quality* si passa alla fase di *Data Integration*. Si ricorda che per la prima fonte dati (API) si ha un database documentale in formato json contenente dati riguardanti 13765 film, mentre per la seconda fonte dati (scraping) si ha un database relazionale in formato csv contenente 6525 recensioni. Si integrano i due database in un nuovo database documentale in formato json che verrà immagazzinato su MongoDB. Nel nuovo database per i documenti che hanno una recensione ci sarà un nuovo attributo denominato recensione, relativo alla recensione.
Come attributo sul quale fase Data Integration si considera il **titolo** di ogni film. In totale si eseguono 232 match. I documenti non contengono più solo le caratteristiche principali del film, ma anche una recensione. La recensione offre all'utente una descrizione completa e dettagliata del film in tutte le sue sfaccettature.
Il dataset è pronto per fare le prime analisi di *Data Exploration*.

DATA EXPLORATION

L'esplorazione dei dati è implementata attraverso la libreria `pandas_profiling` e le funzioni di MongoDB. Tutte le analisi fatte sono commentate nel notebook relativo alla Data Exploration. In questa sezione si presentano i risultati ottenuti più interessanti.
La figura sottostante mostra un esempio di analisi esplorativa fatta sulla variabile *IMDB ratings*.



Per la variabile *IMDB ratings*, la distribuzione presenta un'asimmetria negativa, come si può notare dalla figura sottostante. La variabile presenta una media di 6.56 con una standard deviation di 1.53. Il valore minimo, corrispondente a 1.2, appartiene al film "Intrinsic Leverage" di Marcus D. Clark. Quest'ultimo appare nel dataset solo una volta, non avendo diretto altri film. Al contrario, il valore massimo, corrispondente a 9.4, appartiene al film "Fanny Pey" di Alex Amadei. Quest'ultimo ha diretto anche altri film come "Beneath Glass Trees" (con punteggio IMDB mancante) e "Kaptin Boom" (con punteggio IMDB pari a 8.3).

Per quanto riguarda la variabile *budget* di *BoxOffice*, il numero di film che presentano un valore con incasso minore di 100$ sono 241. Il valore minimo è pari a 1$ ed appartiene a 38 film. Il valore massimo invece, è pari a $100,000,000 ed appartiene ad un solo film dal titolo "The Game of Freedom" di Vladimir Dukelic.

SVILUPPI FUTURI

In generale si ritiene soddisfacente il lavoro eseguito per creare il database finale. Uno sviluppo futuro per rendere il database più completo è quello di andare acquisire nuove recensioni provenienti da altri siti ufficiali, non solo il sito di Roger Ebert. Inoltre, in futuro il dataset può essere arricchito con nuove informazioni relative a film sempre più attuali.

APPENDICE A

Descrizione Attributi Film IMDB prima della fase di Data Cleaning

1. **ActorList**: lista di dizionari, dove ogni dizionario contiene due chiavi, contenenti informazioni riguardo il nome dell'attore e il ruolo interpretato.
2. **Awards**: lista di stringhe, contenente informazioni riguardo la classifica dei "Top Rated Movie", il numero di Oscar vinti, i premi vinti e le nomine totali;
3. **BoxOffice**: dizionario in riferimento al botteghino. Esso contiene 4 chiavi, dove viene specificato il budget (che include tutti i costi relativi allo sviluppo, produzione e post-produzione del film), il Worldwide Cumulative Box Office (l'importo totale di denaro pagato dalle persone di tutto il mondo per guardare un film), grossUSA (l'importo totale raccolto attraverso la vendita di biglietti sugli sportelli dei cinema o online durante un determinato periodo di tempo negli USA.), OpeningweUSA (definito come le ricevute al botteghino dal venerdì alla domenica negli USA);
4. **Companies**: lista di studi di produzione cinematografica;
5. **CompanyList**
6. **ContentRating**: i certificati di rating dati ai titoli, necessari per ammettere o escludere dalla visione gli spettatori più giovani;
7. **Countries**: lista di paesi. IMDB definisce il paese di un titolo come il luogo o i luoghi in cui hanno sede le società di produzione per quel titolo, e quindi dove ha avuto origine il finanziamento;
8. **CountryList**
9. **Directors**: lista di direttori;
10. **DirectorList**
11. **ErrorMessage**: questo attributo indica se durante lo scaricamento dati ci sono stati dei problemi. Il valore dell'attributo è *None* se non ci sono stati problemi;
12. **fullTitle**: Titolo integrale;
13. **Genres**: lista di generi ai quali appartiene il film;
14. **GenreList**
15. **Id**: codice alfanumerico che identifica univocamente il film;
16. **imdbRating**: media delle valutazioni date dagli utenti di IMDB. Il punteggio non consiste in una media algebrica ma in una media ponderata: la valutazione degli utenti più attivi pesa di più rispetto a quella di utenti meno attivi;
17. **imdbRatingVotes**: numero di utenti IMDB che ha valutato il film;
18. **Image**: link all'immagine di copertina del film su IMDB;
19. **Images**
20. **keyWordList**: lista di parole chiave, ovvero parole allegate a un titolo per descrivere qualsiasi oggetto, concetto, stile o azione notevole che si svolge durante un titolo. Lo scopo principale delle parole chiave è consentire ai visitatori di cercare e scoprire facilmente i titoli;
21. **Keywords**: stringa contenente tutte le parole chiave;
22. **Languages**: lista delle lingue parlate nei titoli nel database;
23. **LanguageList**
24. **metacriticRating**: punteggio che converte ogni recensione dal sito web "Metecritic" in una percentuale;

25. **OriginalTitle**: titolo originale;
26. **Plot**: breve descrizione della trama per aiutare a informare le persone degli eventi che si svolgono nel titolo;
27. **PlotLocal**
28. **PlotLocalIsRTL**
29. **posters**
30. **ratings**
31. **releaseData**: data di uscita, registra quando un titolo è stato rilasciato al pubblico in un determinato paese;
32. **runtimeMins**: durata in minuti del film;
33. **runtimeMinStr**
34. **Similars**: lista di film simili;
35. **StarList**: lista degli attori più popolari che hanno recitato nel film. Si decide di mantenere solo la lista degli attori;
36. **Stars**
37. **Tagline**: Sottotitolo;
38. **title**: Titolo;
39. **trailer**: link al trailer del film;
40. **tvEpisodeInfo**
41. **tvSeriesInfo**
42. **type**: tipo di dato. In questo caso tutti i film hanno valore 'movie' per l'attributo type;
43. **wikipedia**
44. **writerList**
45. **Writers**: lista di sceneggiatori;
46. **Year**: anno di produzione.

Descrizione Attributi Film IMDB dopo la fase di Data Cleaning

1. ActorList
2. Awards
3. Box-office
4. Companies
5. ContentRating
6. Countries
7. Directors
8. FullCast
9. Genres
10. imdbRating
11. keyWordList
12. Languages
13. metacriticRatingPlot
14. Plot
15. ratings
16. releaseData
17. runtimeMins

18. Similars
19. Title
20. Writers
21. Year

APPENDICE B

Esempio Output API Advaced Search


```
{'id': 'tt3446122',
 'image': 'https://imdb-api.com/images/original/nopicture.jpg',
 'title': 'Dementia: Alone',
 'description': '(2010)',
 'runtimeStr': None,
 'genres': 'Horror',
 'genreList': [{'key': 'Horror', 'value': 'Horror'}],
 'contentRating': None,
 'imDbRating': None,
 'imDbRatingVotes': None,
 'metacriticRating': None,
 'plot': None,
 'stars': 'Carla Saunders, Wardell Richardson, Christine Hameed, Don Cano, Donna Hutchinson',
 'starList': [{'id': 'tt3446122', 'name': 'Carla Saunders'},
  {'id': 'tt3446122', 'name': 'Wardell Richardson'},
  {'id': 'tt3446122', 'name': 'Christine Hameed'},
  {'id': 'tt3446122', 'name': 'Don Cano'},
  {'id': 'tt3446122', 'name': 'Donna Hutchinson'}]}
```


Esempio Output API Title


```
{'title': 'Dementia: Alone',
 'year': 2010,
 'directors': ['Carla Saunders'],
 'writers': ['Carla Saunders'],
 'actorList': [{'name': 'Wardell Richardson', 'asCharacter': 'OFFICER COX'},
  {'name': 'Christine Hameed', 'asCharacter': 'On set Voice overs'},
  {'name': 'Don Cano', 'asCharacter': 'OFFICER JACOBS'},
  {'name': 'Donna Hutchinson', 'asCharacter': 'Dina'},
  {'name': 'Jassa Deen', 'asCharacter': 'Antagonist Role'}],
 'genres': ['Horror'],
 'companies': ['Promises Entertainment'],
 'countries': ['USA'],
 'languages': ['English'],
 'boxOffice': {'budget($)': 12000}}
```


Si consideri che ci sono alcuni attributi mancanti.