Mongo DB José Gabriel González Parra- Master Big Data y Data Science

Indice

- Indice
- Introducción
- Carga del fichero en MongoDB Compass
- Numero de películas
- Top 10 mejores peliculas según IMDB
- Top 10 mejores peliculas según la votación en la Web
- Top 10 peliculas con mayor Box Office(Recaudación)
- Top 10 actores mayor recaudación
- Top 10 directores mayor recaudación
- Número de películas por género
- Año con mayor número de películas
- Películas más largas y mas cortas del top 250
- Directores con mejor valoración en el Top 250
- Directores con mayor número de películas en el Top 250
- Conclusiones

Introducción

En el presente trabajo se procede a analizar mediante consultas en una base de datos de **MongoDB** un conjunto de datos de películas de la web IMDB este siendo su ranking de las mejores 250 películas de la historia. Para la realización de las consultas se hizo uso del GUI MongoDB Compass para realizar dichas consultas así como la importación de datos JSON, y del Shell de MongoDB.

El dataset ha sido extraído por mi persona, mediante el uso del **IDE** para **Python** Pycharm y las librerías Scrapy y BS4, puede encontrar el script de Python completo para la extracción de datos de la página web de IMDB en el siguiente link hacia Github.

Además de su extracción, los datos pertenecientes a películas han sido depurados para elimnar caracteres imnecesarios, errores de extracción asi como cambiar ciertos tipos de datos de str a int y eliminar espacios entre caracateres, en el siguiente link podra encontrar la depuración completa de los datos realizada en un Jupyter Notebook.

Los campos pertenecientes al dataset peliculas.csv son los siguientes :

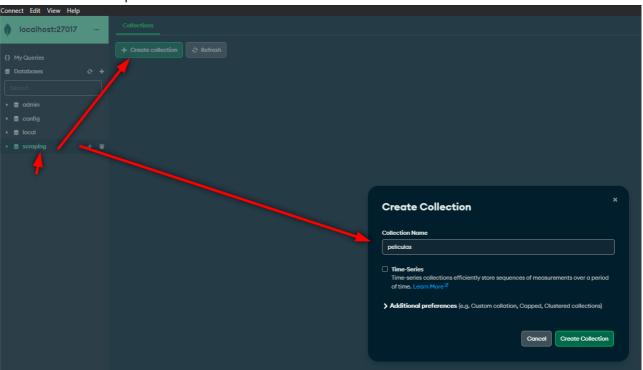
- 1. **Title** : Nombre de la película
- 2. **Year** : Año estrenado la película
- 3. Run_time : Duración estimada en minutos(min) de la película
- 4. Genre : Género de la película
- 5. **imdb_score** : Puntuación de la página IMDB
- 6. **Metascore_score** : Puntuación de la página Metacritic
- 7. Summary: Breve resumen de la película
- 8. **Director** : Nombre del director de la película
- 9. Lead_actor: Nombre del actor principal de la película
- 10. **supporting_actor** : Nombre del actor secundario de la película

- 11. tritagonist_Actor : Nombre del actor terciario de la película
- 12. Votes : Número de votos por parte del usuario para cada película
- 13. **Box_office**: Recaudación a nivel mundial expresada en Millones de dólares(\$)
- 14. Top250_position: Posición en el ranking de acuerdo a su puntuación de IMDB

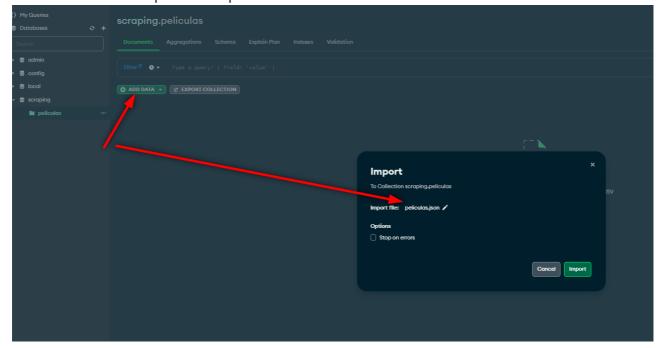
Carga del fichero en MongoDB Compass

La importación del fichero se hizo de **MongoDB Compass**, a través del siguiente proceso:

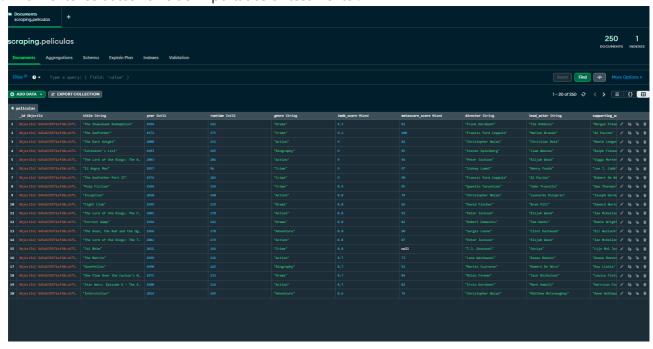
1. Se crea la colección películas :



2. Dentro de la colección películas se procede a añadir datos :



3. Finalmente los datos han sido importados exitosamente:



Numero de películas

Procedemos a contar el número de películas que contiene el Dataset de IMDB el portal web de IMDB dedicado a la información así como a la puntuación de películas, series y videojuegos.

```
db.peliculas.find().count()

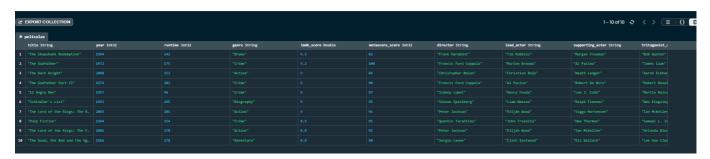
> db.peliculas.find().count()

< 250
scraping>
Volver al indice
```

Top 10 mejores peliculas según IMDB

Ahora se procede a consultar cual sería para la web IMDB el top 10 de las mejores películas de la historia de acuerdo a su puntuación en IMDB :

```
db.peliculas.find().sort({top_250_position : 1}).limit(10)
```

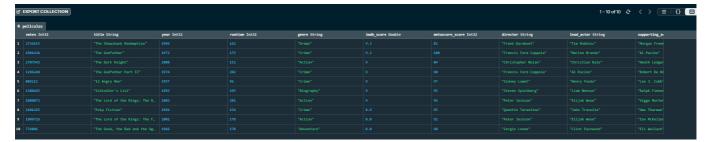


Volver al indice

Top 10 mejores peliculas según la votación en la Web

Una vez visto la clasificación del top 10 de acuerdo a su puntuación en IMDB, veamos ahora el top 10 de las películas mas votadas para observar las diferencias entre la popularidad de la película y su puntuación :

db.peliculas.find().sort({votes : -1}).limit(10)

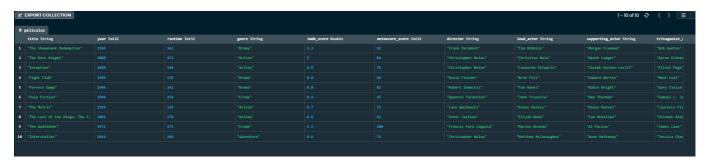


Volver al indice Podemos apreciar como ciertas películas que ciertas películas tales como: **The Shawshank Redemption,The Dark Knight,Pulp Fiction, El señor de los anillos, El padrino**,son peliculas con alta puntuación y altamente populares con las mayores cantidades de votos, sin embargo también podemos apreciar que hay películas muy populares como **Interestellar** que aunque están bien valoradas no son a priori parte del top 10 de IMDB. Volver al indice

Top 10 peliculas con mayor Box Office(Recaudación)

Ahora pasando al ámbito monetario, consultemos en la base de datos **películas** en MongoDB cual es del dataset las 10 películas con mayor recaudación a nivel mundial :

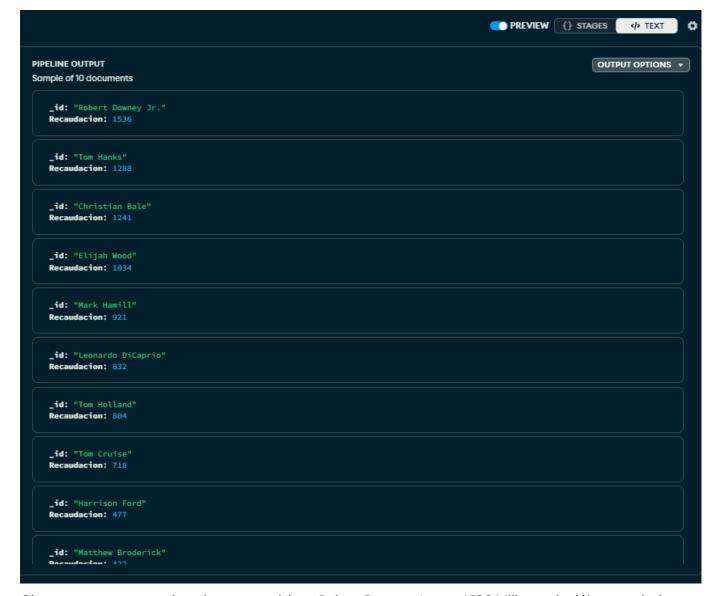
db.peliculas.find().sort({votes : -1}).limit(10)



Por motivos de legibilidad, **box_office** está representado en millones de dólares, por ende en primer lugar se posiciona la película : **"Avengers: Endgame"** con 858, es decir que tuvo una recaudación a nivel mundial de 858.000.000 \$. Volver al indice

Top 10 actores mayor recaudación

Un aspecto fundamental al momento de financiar una película es tener en cuanta el retorno de la inversión de dicha película, para ello cuanta mucho tener un actor taquillero es decir, que sea mundialmente conocido y llene las salas de cine, por eso consultaremos en la base de datos cual es el actor con mas recaudación en sus películas :

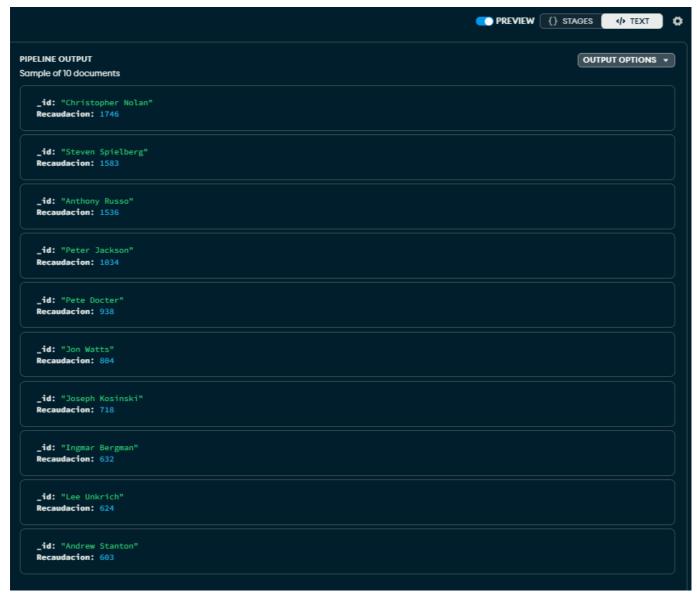


Observamos como en primer lugar se posiciona Robert Downey Jr. con 1536 Millones de dólares es decir : 1.536.000.000 \$, lógico teniendo en cuanto su participación en las películas de Marvel **"Los Vengadores"**.

Volver al indice

Top 10 directores mayor recaudación

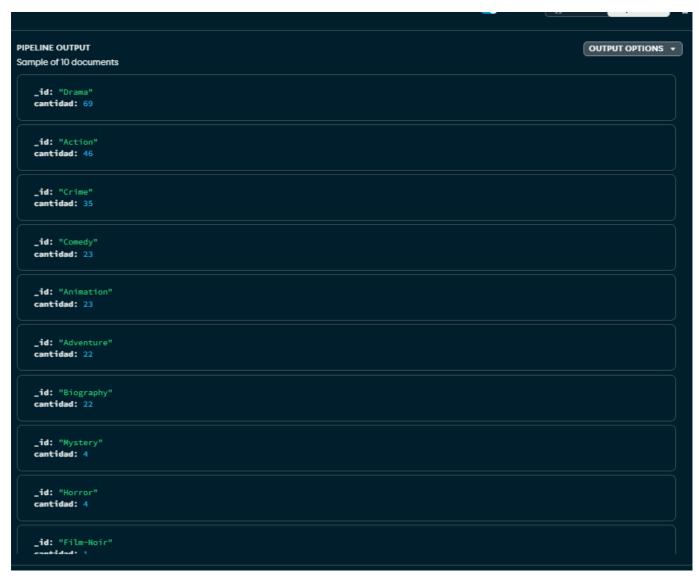
De la misma manera que tomamos en cuanta la recaudación total de cada actor, es de suma importancia de la misma manera observar la recaudación total por directores :



Volver al indice Podemos concluir de esta forma que para obtener el máximo ROI se debería contratar a **Cristopher Nolan** como director debido a que es el director con mayor recaudación a nivel mundial con 1746 Millones de dólares y a **Robert Downey Jr.** como actor principal con una recaudación de 1536 Millones de dólares.

Número de películas por género

Ahora observemos el número de peliculas por cada género en el top 250 películas de IMDB :



Observamos con el el género con mayor número de peliculas es Drama con 69, seguido de acción con 46. Volver al indice

Año con mayor número de películas

Prosigamos con averiguar cual sería el año con mayor número de películas dentro del top 250 :

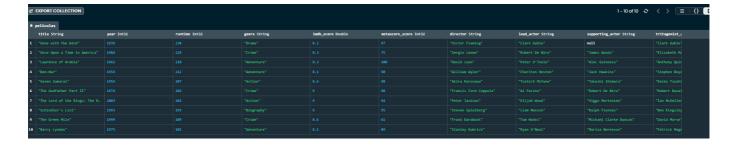


Tenemos al año 1995 como el año con mayor número de películas en el top 250 de IMDB con películas mundialmente conocidas como **Toy Story, Casino, S7even** entre otros. Volver al indice

Películas más largas y mas cortas del top 250

Observamos a continuación las películas mas largas dentro del top 250 :

```
db.peliculas.find().sort({runtime : -1}).limit(10)
```



Observemos a continuación las películas mas cortas dentro del top 250 :

```
db.peliculas.find().sort({runtime : 1}).limit(10)
```



Volver al indice

Directores con mejor valoración en el Top 250

Observemos cuáles son los directores con mejores valoración por la crítica de la pagina Metacricit :

```
PREVIEW {} STAGES $\lfloor \text{$\psi}$ TEXT
PIPELINE OUTPUT
                                                                                                                      OUTPUT OPTIONS +
Sample of 10 documents
   _id: "Orson Welles"
   valoracion: 100
   id: "Michael Curtiz"
   valoracion: 100
   _id: "Yasujirô Ozu"
   valoracion: 100
   _id: "Stanley Donen"
   valoracion: 99
   _id: "Fritz Lang"
   valoracion: 98
   _id: "John Huston"
   valoracion: 98
   _id: "Joseph L. Mankiewicz"
   valoracion: 98
   _id: "Carl Theodor Dreyer"
   valoracion: 98
   _id: "Guillermo del Toro"
   valoracion: 98
   _id: "Charles Chaplin"
```

Volver al indice

Directores con mayor número de películas en el Top 250

Una vez observado la valoración, centremonos ahora en otra métrica importante para determinar que tan reconocido es un director, el número de películas dentro del top 250 :

```
db.peliculas.aggregate([
    {$group: {
        _id: "$director",
        numero_peliculas : {$sum : 1}
    }},
    {
        $sort: {
            numero_peliculas: -1
        }}
])
```

```
PIPELINE OUTPUT
                                                                                                                     OUTPUT OPTIONS •
Sample of 10 documents
   _id: "Christopher Nolan"
   numero_peliculas: 7
   _id: "Stanley Kubrick"
  numero_peliculas: 7
   _id: "Akira Kurosawa"
   numero_peliculas: 7
   _id: "Martin Scorsese"
   numero_peliculas: 7
   _id: "Steven Spielberg"
  numero_peliculas: 7
   _id: "Alfred Hitchcock"
   numero_peliculas: 6
   _id: "Billy Wilder"
   numero_peliculas: 5
   _id: "Quentin Tarantino"
  numero_peliculas: 5
   _id: "Charles Chaplin"
   numero_peliculas: 5
   _id: "Hayao Miyazaki"
        o policular: 4
```

Volver al indice

Conclusiones

La realización del análisis del dataset de Peliculas utilizando la base de datos NoSQL, MongoDB y su GUI MongoDB Compass se obtuvo de manera eficaz, con resultados perfectamente legibles, con diferentes opciones de lectura y distintos formatos. Gracias a una fuente de datos conocida como son películas mundialmente conocidas ha sido sencillo realizar consultas pertinentes e interesantes así como su interpretación. Destacar la Interfaz de **MongoDB Compass** que además de ser agradable de visualizar, posee variadas opciones de ayuda con botones de ubicación permanente y funcionamiento impecable sin problemas de lag ni de reinicio.

Volver al indice