

# Tarea NoSQL

Javier Aós Aragonés

## Introducción

A lo largo de la historia del cine se han creado numerosas obras cinematográficas de diferentes características. Algunas de estas más largas, otras más cortas, pertenecientes a géneros distintos, con valoraciones y recaudaciones diferentes etc. Dentro de esta gran variedad el espectador debe elegir la obra de la que desea disfrutar, ya sea a la hora de ir al cine, de comprar la película físicamente o incluso de elegir entre las muchas opciones que las plataformas digitales ofrecen.

Todo esto requiere de un desembolso económico previo, lo que hace que la industria del cine se convierta en un negocio y por ello que dentro del propio negocio se pueda analizar qué tipo de películas funcionan mejor atendiendo a las características de cada una.

En este trabajo se analizarán los datos de las 1000 películas mejor valoradas en IMBD. A partir de los datos brindados por esta página, podremos tener un mejor entendimiento sobre “qué hace a una película una buena película”.

## Análisis de los datos

Antes de nada, debemos importar los datos a la base de datos de MongoDB. Seleccionamos/creamos la base de datos que queremos usar.

use tarea

Una vez seleccionada, pulsamos el botón de “Import” para importar los datos. En este caso será un CSV sacado de la página web “Kaggle” que se añadirá en la entrega de la tarea.

Una vez importados los datos empezamos con las consultas.

Lo primero, vamos a ver cuántos documentos se han cargado.
















```
db.movies.find().count()
```



1000

Vemos que tenemos 1000 documentos cargados, perteneciendo cada uno de estos a una película. Ahora para saber que datos tenemos de cada película seleccionaremos una de ejemplo con sus claves.

```
db.movies.findOne()
```

Key	Value	Type
 (1) 6269139a049f334	{14 fields}	Document
 _id	6269139a049f3343d4ea4ce1	ObjectId
 Movie_Title	The Shawshank Redemption	String
 Released_Year	1994	Int32
 Runtime	142	Int32
 Genre	[ "Drama" ]	Array
 IMDB_Rating	9,3	Double
 Meta_score	80	Int32
 Director	Frank Darabont	String
 Star1	Tim Robbins	String
 Star2	Morgan Freeman	String
 Star3	Bob Gunton	String
 Star4	William Sadler	String
 Noofvotes	2.343.110 (2.3M)	Int32
 Gross	28.341.469 (28.3M)	Int32

Como podemos ver tenemos:

1. Movie\_Title: Nombre de la película
2. Released\_Year: Año en el que se estrenó
3. Runtime: Duración total
4. Genre: Género
5. IMDB\_Rating: Valoración de la película en IMBD
6. Meta\_score: Valoración de la película según críticos de cine
7. Director: Nombre del director
8. Star1, Star2, Star3, Star4: Nombre de las “estrellas”
9. Noofvotes: Número total de votos
10. Gross: Dinero total generado por la película

Una vez revisada la estructura de los datos vamos a consultar información dentro de esta. Aunque antes vamos a hacer un ejemplo de inserción, actualización y eliminación de datos como ejemplo básico.

```

db.movies.insert({
  "Movie_Title" : "Ejemplo",
  "Released_Year" : 2022,
  "Runtime" : 144,
  "Genre" : [
    "Drama",
    "Crime"
  ],
  "IMDB_Rating" : 10,
  "Meta_score" : 99,
  "Director" : "Pepe Perez",
  "Star1" : "Tim Robbins",
  "Star2" : "Morgan Freeman",
  "Star3" : "Bob Gunton",
  "Star4" : "William Sadler",
  "Noofvotes" : 244556,
  "Gross" : 3000709
})

db.movies.update({"Movie_Title" : "Ejemplo"}, {$set:{Meta_score: 90}})

db.movies.find({"Movie_Title": "Ejemplo"})

db.movies.remove({"Movie_Title": "Ejemplo"})

```

Simplemente hemos añadido un nuevo documento de ejemplo, hemos actualizado su meta score y lo hemos eliminado.

Ahora sí, vamos a consultar el número de valores nulos que hay para las cuatro categorías más importantes.

```

db.movies.find({IMDB_Rating: {$exists: false}}).count()
db.movies.find({Meta_score: {$exists: false}}).count()
db.movies.find({Noofvotes: {$exists: false}}).count()
db.movies.find({Gross: {$exists: false}}).count()

```

```

0      163    5      175

```

Como podemos ver la mayoría de los valores nulos se concentran en el "Meta\_score" y el "Gross".

Vamos a consultar cuales son las 10 mejores películas según su valoración en IMBD.

```

db.movies.find({}, {_id: 0, Movie_Title: 1, IMDB_Rating: 1})
.sort({ IMDB_Rating:-1 }).limit(10)

```

	Movie_Title	IMDB_Rating
1	The Shawshank Redemption	9,3
2	The Godfather	9,2
3	The Godfather: Part II	9
4	12 Angry Men	9
5	The Dark Knight	9
6	Schindler's List	8,9
7	The Lord of the Rings: The Return of the King	8,9
8	Pulp Fiction	8,9
9	Inception	8,8
10	Fight Club	8,8

Y ahora vamos a ver las mejores valoradas según su “Meta\_score”, como muchas de estas valoraciones son 100 (el máximo), vamos a ordenarlas también por su valoración en IMDB como segundo orden.

```
db.movies.find({}, {_id: 0, Movie_Title: 1, Meta_score: 1})
.sort({ Meta_score: -1, IMDB_Rating: -1}).limit(10)
```

	Movie_Title	Meta_score
1	The Godfather	100
2	Casablanca	100
3	Rear Window	100
4	Lawrence of Arabia	100
5	Vertigo	100
6	Fanny och Alexander	100
7	Trois couleurs: Rouge	100
8	Sweet Smell of Success	100
9	Il conformista	100
10	Boyhood	100

Como podemos observar las 10 mejores películas según IMDB y según los críticos de cine no coinciden en su totalidad. Se podría tener en cuenta cualquiera de las dos valoraciones para saber si una película es mejor o peor, pero en este caso voy a centrarme en el rating de IMDB ya que es el de los usuarios finales los que al ser mayoría aportan más al negocio del cine y hacen que una película sea más “famosa”. Además, como hemos visto antes existen 163 valores nulos en el meta score.

Ahora vamos a ver que películas son las que más han recaudado o las más taquilleras.

```
db.movies.find({}, {_id: 0, Movie_Title: 1, Gross: 1})
.sort({ Gross: -1}).limit(10)
```

	Movie_Title ↕	Gross ↕
1	Star Wars: Episode VII - The Force Awakens	936.662.225 (0.94G)
2	Avengers: Endgame	858.373.000 (0.86G)
3	Avatar	760.507.625 (0.76G)
4	Avengers: Infinity War	678.815.482 (0.68G)
5	Titanic	659.325.379 (0.66G)
6	The Avengers	623.279.547 (0.62G)
7	Incredibles 2	608.581.744 (0.61G)
8	The Dark Knight	534.858.444 (0.53G)
9	Rogue One	532.177.324 (0.53G)
10	The Dark Knight Rises	448.139.099 (0.45G)

Vemos de nuevo que las películas más taquilleras no coinciden con las mejores valoradas según IMDB o los críticos.

Ahora vamos a ver las películas con un mayor número de votos.

```
db.movies.find({}, {_id: 0, Movie_Title: 1, Noofvotes: 1, IMDB_Rating: 1})
.sort({ Noofvotes: -1}).limit(10)
```

	Movie_Title ↕	IMDB_Rating ↕	Noofvotes ↕
1	The Shawshank Redemption	9,3	2.343.110 (2.3M)
2	The Dark Knight	9	2.303.232 (2.3M)
3	Inception	8,8	2.067.042 (2.1M)
4	Fight Club	8,8	1.854.740 (1.9M)
5	Pulp Fiction	8,9	1.826.188 (1.8M)
6	Forrest Gump	8,8	1.809.221 (1.8M)
7	The Matrix	8,7	1.676.426 (1.7M)
8	The Lord of the Rings: The Fellowship of the Ring	8,8	1.661.481 (1.7M)
9	The Lord of the Rings: The Return of the King	8,9	1.642.758 (1.6M)
10	The Godfather	9,2	1.620.367 (1.6M)

Como podemos ver, en este caso, he añadido el campo “IMDB\_Rating” para poder visualizarlo ya que muchas de las películas más votadas coinciden con las mejor valoradas en IMDB, por lo que aquí podría haber una correlación interesante entre estas dos variables.

Vamos a consultar cuales son las 10 películas más largas.

```
db.movies.find({}, {_id: 0, Movie_Title: 1, Runtime: 1, IMDB_Rating: 1})
.sort({ Runtime: -1}).limit(10)
```

	Movie_Title ◄	Runtime ◄	IMDB_Rating ◄
1	Gangs of Wasseyapur	321	8,2
2	Hamlet	242	7,7
3	Gone with the Wind	238	8,1
4	Once Upon a Time in America	229	8,4
5	Lawrence of Arabia	228	8,3
6	Lagaan: Once Upon a Time in India	224	8,1
7	The Ten Commandments	220	7,9
8	Ben-Hur	212	8,1
9	Swades: We the People	210	8,2
10	The Irishman	209	7,9

Vemos que a priori no existe una correlación entre la duración y la valoración.

Otro dato interesante puede ser el resultado de agrupar el número de películas estrenadas cada año.

```
db.movies.aggregate(
[
  {$unwind: "$Released_Year"},
  {$group: {_id: "$Released_Year", Total: {$sum: 1}}},
  {$sort: {Total: -1}},
  {$limit: 10}
])
```

	_id ◄	Total ◄
1	2014	32
2	2004	31
3	2009	29
4	2013	28
5	2016	28
6	2001	27
7	2007	26
8	2006	26
9	2015	25
10	2012	24

El año en el que más películas se han estrenado ha sido 2014, seguido muy de cerca por 2004.

Podemos hacer lo mismo con los directores.

```
db.movies.aggregate(
[
  {$unwind: "$Director"},
  {$group: {_id: "$Director", Total: {$sum: 1}}},
  {$sort: {Total: -1}}
  {$limit: 10}
])
```

	_id	Total
1	Alfred Hitchcock	14
2	Steven Spielberg	13
3	Hayao Miyazaki	11
4	Martin Scorsese	10
5	Akira Kurosawa	10
6	Stanley Kubrick	9
7	Billy Wilder	9
8	Woody Allen	9
9	Clint Eastwood	8
10	Quentin Tarantino	8

Aquí vemos que Alfred Hitchcock es el que más películas ha dirigido también seguido muy de cerca por Steven Spielberg.

Ahora vamos a ver que nota media tiene cada director según sus películas.

```
db.movies.aggregate(
[
  {$unwind: "$Director"},
  {$group: {_id: "$Director", Nota_Media: {$avg: "$IMDB_Rating"}, Total: {$sum: 1}}},
  {$sort: {Nota_Media: -1}}
  {$limit: 10}
])
```

	_id	Nota_Media	Total
1	Frank Darabont	8,95	2
2	Lana Wachowski	8,7	1
3	Irvin Kershner	8,7	1
4	Masaki Kobayashi	8,6	1
5	George Lucas	8,6	1
6	Thomas Kail	8,6	1
7	Sudha Kongara	8,6	1
8	Roberto Benigni	8,6	1
9	Fernando Meirelles	8,6	1
10	Olivier Nakache	8,5	1

Como vemos la mayoría son directores con solo una película, por lo tanto, vamos a realizar la misma consulta, pero indicando que el director tenga al menos 5 películas dirigidas, para que así tengamos una muestra algo mayor para cada director.

```
db.movies.aggregate(
[
  {$unwind: "$Director"},
  {$group: {_id: "$Director", Nota_Media: {$avg: "$IMDB_Rating"}, Total: {$sum: 1}}},
  {$match: {Total: {$gte: 5}}}
  {$sort: {Nota_Media: -1}}
  {$limit: 10}
])
```

	_id	Nota_Media	Total
1	Christopher Nolan	8,4625	8
2	Francis Ford Coppola	8,4	5
3	Peter Jackson	8,4	5
4	Charles Chaplin	8,3333	6
5	Sergio Leone	8,2667	6
6	Stanley Kubrick	8,2333	9
7	Akira Kurosawa	8,22	10
8	Quentin Tarantino	8,175	8
9	Martin Scorsese	8,17	10
10	Billy Wilder	8,1444	9

Ahora sí, podemos ver que en la consulta se reconocen algunos directores bastante conocidos.

Podemos consultar las películas dirigidas por el director de cine con mayor nota media, que es "Christopher Nolan".

```
db.movies.aggregate(
  {$match: {Director: "Christopher Nolan"}}
  {$sort: {IMDB_Rating: -1}}
  {$project: {_id: 0, Director: 0, Star1: 0, Star2: 0, Star3: 0, Star4: 0}}
)
```

	Movie_Title	Released_Year	Runtime	Genre	IMDB_Rating	Meta_score	Noofvotes	Gross
1	The Dark Knight	2008	152 min	Action Crime Drama	9	84	2.303.232 (2.3M)	534.858.444 (0.53G)
2	Inception	2010	148 min	Action Adventure Sci-Fi	8,8	74	2.067.042 (2.1M)	292.576.195 (0.29G)
3	Interstellar	2014	169 min	Adventure Drama Sci-Fi	8,6	74	1.512.360 (1.5M)	188.020.017 (0.19G)
4	The Prestige	2006	130 min	Drama Mystery Sci-Fi	8,5	66	1.190.259 (1.2M)	53.089.891 (53.1M)
5	The Dark Knight Rises	2012	164 min	Action Adventure	8,4	78	1.516.346 (1.5M)	448.139.099 (0.45G)
6	Memento	2000	113 min	Mystery Thriller	8,4	80	1.125.712 (1.1M)	25.544.867 (25.5M)
7	Batman Begins	2005	140 min	Action Adventure	8,2	70	1.308.302 (1.3M)	206.852.432 (0.21G)
8	Dunkirk	2017	106 min	Action Drama History	7,8	94	555.092 (0.56M)	188.373.161 (0.19G)

Ahora vamos a ver los mejores directores según su recaudación.

```
db.movies.aggregate(
[
  {$unwind: "$Director"},
  {$group: {_id: "$Director", Recaudación: {$sum: "$Gross"}}},
  {$sort: {Recaudación: -1}}
  {$limit: 10}
])
```



	_id	Recaudación
1	Steven Spielberg	2.478.133.165 (2.5G)
2	Anthony Russo	2.205.039.403 (2.2G)
3	Christopher Nolan	1.937.454.106 (1.9G)
4	James Cameron	1.748.236.602 (1.7G)
5	Peter Jackson	1.597.312.443 (1.6G)
6	J.J. Abrams	1.423.170.905 (1.4G)
7	Brad Bird	1.099.627.795 (1.1G)
8	Robert Zemeckis	1.049.446.456 (1.0G)
9	David Yates	978.953.721 (0.98G)
10	Pete Docter	939.382.131 (0.94G)

Vemos que algunos coinciden con los que mejor nota media tienen en IMDB.

Ahora veremos los mejores actores principales según su recaudación.

```
db.movies.aggregate(
  [
    { $unwind: "$Star1",
      { $group: { _id: "$Star1", Recaudación: { $sum: "$Gross" } } },
      { $sort: { Recaudación: -1 } }
    ],
    { $limit: 10 }
  ]
)
```

	_id	Recaudación
1	Tom Hanks	2.493.097.454 (2.5G)
2	Joe Russo	2.205.039.403 (2.2G)
3	Leonardo DiCaprio	1.877.321.752 (1.9G)
4	Daniel Radcliffe	1.835.901.034 (1.8G)
5	Christian Bale	1.351.591.432 (1.4G)
6	Robert Downey Jr.	1.150.720.327 (1.2G)
7	Elijah Wood	1.035.942.020 (1.0G)
8	Daisy Ridley	936.662.225 (0.94G)
9	Mark Hamill	922.340.616 (0.92G)
10	Craig T. Nelson	870.022.836 (0.87G)

Otra consulta interesante podría ser los géneros mejor valorados en IMDB, mostrando también el total de películas que contiene cada género.

```
db.movies.aggregate(
[
  {$unwind: "$Genre"},
  {$group: {_id: "$Genre", Nota_Media: {$avg: "$IMDB_Rating"}, Total: {$sum: 1}}},
  {$sort: {Nota_Media: -1}},
  {$limit: 10}
])
```

	_id	Nota_Media	Total
1	War	8,0137	51
2	Western	8	20
3	Film-Noir	7,9895	19
4	Sci-Fi	7,9776	67
5	Mystery	7,9677	99
6	Drama	7,9594	724
7	Crime	7,9545	209
8	History	7,9536	56
9	Adventure	7,952	196
10	Action	7,9487	189

Vemos que no por tener más películas publicadas el género es mejor. Además, también se puede observar que la nota media de cada género no varía demasiado, por lo que se entiende que no existe un género de preferencia muy por encima de los demás.

También podríamos consultar los géneros que más han recaudado.

```
db.movies.aggregate(
[
  {$unwind: "$Genre"},
  {$group: {_id: "$Genre", Recaudación: {$sum: "$Gross"}}},
  {$sort: {Recaudación: -1}},
  {$limit: 10}
])
```

	_id	Recaudación
1	Adventure	28.162.383.500 (28.2G)
2	Drama	27.286.692.442 (27.3G)
3	Action	22.032.503.614 (22.0G)
4	Comedy	12.244.862.183 (12.2G)
5	Sci-Fi	9.029.610.220 (9.0G)
6	Animation	8.573.378.214 (8.6G)
7	Crime	6.952.527.041 (7.0G)
8	Thriller	6.442.895.382 (6.4G)
9	Fantasy	6.080.093.435 (6.1G)
10	Biography	5.298.587.721 (5.3G)

Aquí si que podemos ver una buena diferencia entre los géneros que más han recaudado (aventura, drama y acción) frente a los que menos han recaudado (thriller, fantasía y

biografía). Por lo tanto, vemos que existe una preferencia de género en los espectadores atendiendo a la recaudación.

## Conclusiones

A lo largo del análisis hemos podido ver como dependiendo del criterio u objetivo que elijamos la concepción de “buena película” puede cambiar mucho. Hemos visto como una buena película según su recaudación es distinta a una buena película según los espectadores o incluso los críticos de cine. También hemos visto como ocurría lo mismo con el género de la película o los directores y actores principales.

Por tanto, como conclusión, podemos decir que una película puede considerarse buena dependiendo de cuanto satisfaga los criterios “objetivo”. En este caso quizás el criterio más importante para la industria del cine sea la recaudación y en ese caso hemos visto que existen algunos géneros preferencia, así como actores y directores, que consiguen una mayor recaudación que los demás. Cabe mencionar que este ha sido un análisis muy superficial del tema y para sacar conclusiones más precisas se necesitaría realizar un análisis en profundidad.