FILTRADO BASADO EN CONTENIDO

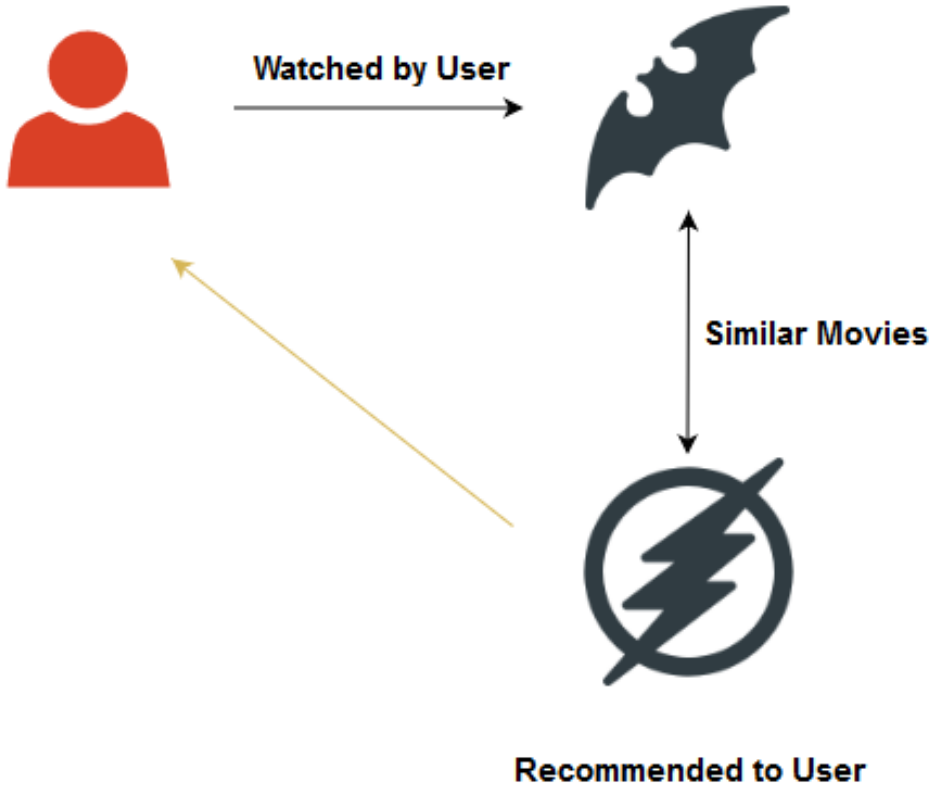
DATOS MASIVOS
DR. CHRISTIAN AGUILAR FUSTER

Juan de Jesús Aguilar Solano	1576327
Karla Cureño Vega	2085376
Abril Grisel Guevara Cedillo	1419239

30 de marzo de 2022

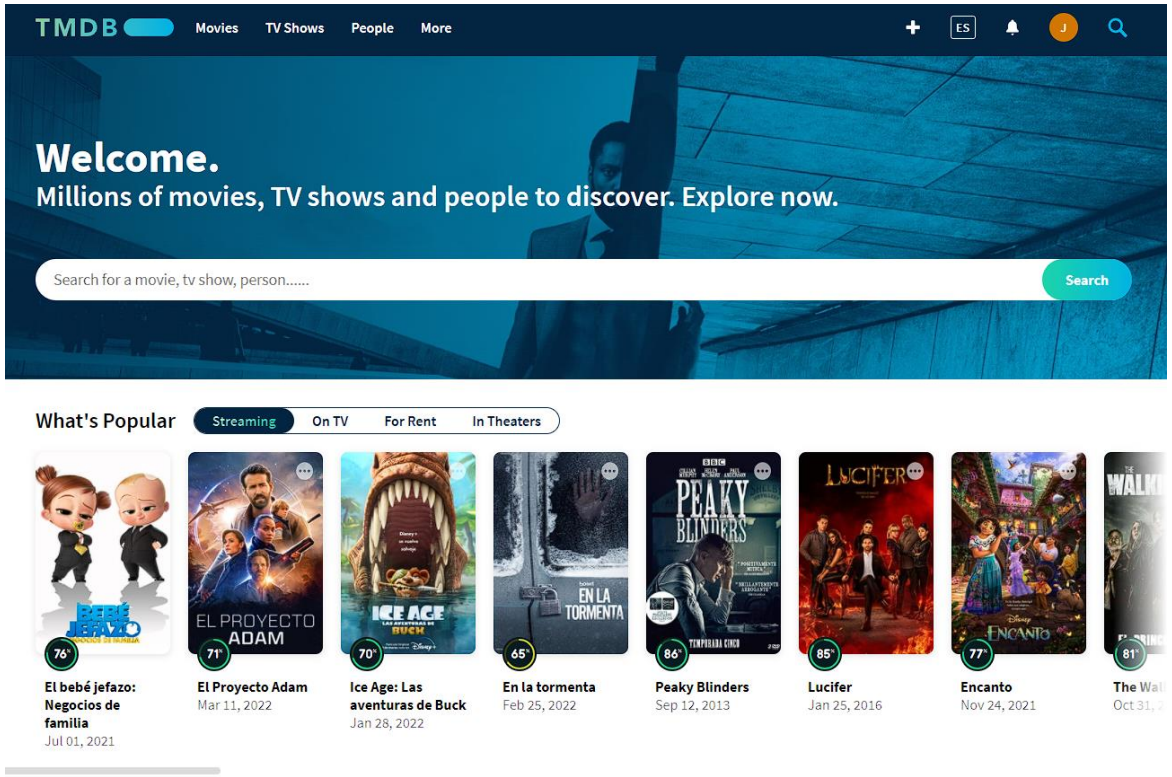
FILTRADO BASADO EN CONTENIDO

Utiliza los atributos de los objetos que el usuario ha consumido, visto o mostrado interés para sugerir nuevos objetos con atributos similares basado en el nivel de correlación entre el objeto observado y el objeto sugerido.

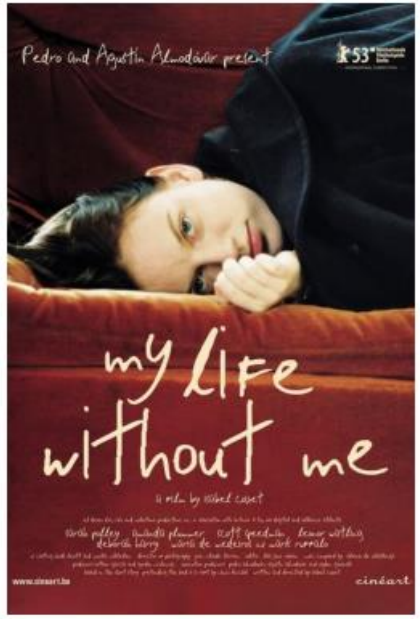


EXTRACCIÓN DE DATOS MEDIANTE API

Fueron extraídos del sitio web The Movie Data DB a través de su API.



EXTRACCIÓN DE DATOS MEDIANTE API



Identificador y nombre

Id = 20

Nombre: My Life Without Me

Sintaxis de consulta:

https://api.themoviedb.org/3/movie/20?api_key=fhk6739847hsjdf


```
{'adult': False,
 'backdrop_path': '/kZyurQjTMLHalUxs7sHgH5XeiwO.jpg',
 'belongs_to_collection': None,
 'budget': 2500000,
 'genres': [{'id': 18, 'name': 'Drama'}, {'id': 10749, 'name': 'Romance'}],
 'homepage': 'http://www.clubcultura.com/clubcine/clubcineastas/isabelcoixet/mividasinmi/index.htm',
 'id': 20,
 'imdb_id': 'tt0314412',
 'original_language': 'en',
 'original_title': 'My Life Without Me',
 'overview': 'A fatally ill mother with only two months to live creates a list of things she wants to do before she dies without
telling her family of her illness.',
 'popularity': 13.863,
 'poster_path': '/9Fa7tCEKIha1llGH7E41mxSpaF6.jpg',
 'production_companies': [{'id': 49,
   'logo_path': '/xpf5iHdvvBtsH8jBMlgIJHAET0c.png',
   'name': 'El Deseo',
   'origin_country': 'ES'},
  {'id': 77,
   'logo_path': None,
   'name': 'Milestone Productions',
   'origin_country': ''}],
 'production_countries': [{'iso_3166_1': 'CA', 'name': 'Canada'},
  {'iso_3166_1': 'ES', 'name': 'Spain'}],
 'release_date': '2003-03-07',
 'revenue': 12300000,
 'runtime': 106,
 'spoken_languages': [{'english_name': 'English',
   'iso_639_1': 'en',
   'name': 'English'}],
 'status': 'Released',
 'tagline': '',
 'title': 'My Life Without Me',
 'video': False,
 'vote_average': 5.8,
 'vote_count': 364}
```

ATRIBUTOS POR PELÍCULA

Para los propósitos de este proyecto, solo utilizaremos una porción de los atributos mostrados, concretamente éstos son:

- id: identificador de la película.
- original_title: título original de la película.
- budget: presupuesto de la película.
- genres: lista de géneros a los que pertenece.
- popularity: popularidad de la película (métrica generada por el sitio).
- release_date: fecha de lanzamiento.
- revenue: ingresos recaudados.
- runtime: duración de la película.
- vote_average: calificación media otorgada por usuarios.
- vote_count: cantidad de usuarios que calificaron la película.
- collection: atributo binario que muestra si una película pertenece a una colección.
- keywords: las palabras clave utilizadas por película (top 100 keywords).

ATRIBUTO GENRES Y KEYWORDS

- Diccionario de todos los géneros:

```
{28: 'Action', 12: 'Adventure', 16: 'Animation', 35: 'Comedy',
80: 'Crime', 99: 'Documentary', 18: 'Drama', 10751: 'Family',
14: 'Fantasy', 36: 'History', 27: 'Horror', 10402: 'Music',
9648: 'Mystery', 10749: 'Romance', 878: 'Science Fiction',
10770: 'TV Movie', 53: 'Thriller', 10752: 'War', 37: 'Western'}
```

- Binarización del género por película:

```
[{'id': 18, 'name': 'Drama'},
{'id': 10749, 'name': 'Romance'}]
```

```
{'Action': 0, 'Adventure': 0,
'Animation': 0, 'Comedy': 0, 'Crime':
0, 'Documentary': 0, 'Drama': 1,
'Family': 0, 'Fantasy': 0, 'History':
0, 'Horror': 0, 'Music': 0,
'Mystery': 0, 'Romance': 1, 'Science
Fiction': 0, 'TV Movie': 0,
'Thriller': 0,
'War': 0, 'Western': 0}
```

- Obtención del top 100 palabras clave (keywords)
  - Muestreo de 100,000 películas

```
[short film, woman director, based on novel or book, murder, musical, concert, silent film,
biography, sports, stand-up comedy, lgbt, christmas, world war ii, revenge, family, love, anime,
philippines, based on true story, martial arts, friendship, romance, coming of age, softcore, black
and white, kidnapping, wrestling, opera, new york city, ghost, based on play or musical, sequel,
police, politics, serial killer, horror, pre-code, found footage, prison, holiday, erotic movie,
drugs, vampire, zombie, rape, parent child relationship, death, gay interest, dance, remake, high
school, art, dark comedy, monster, slasher, gay, religion, stop motion, dog, gore, time travel,
supernatural, lost film, marriage, alien, football (soccer), gangster, detective, superhero, nazi,
africa, suicide, japan, "rock n roll", nature, mockumentary, racism, cartoon, fairy tale, road trip,
satire, pregnancy, spy, small town, mystery, infidelity, investigation, thriller, noir, los angeles,
california, avant-garde, school, robbery, dutch cabaret, london, england, 1970s, sibling
relationship, surrealism, anthology, france]
```

- Binarización de palabras clave por película:
  - Resultado similar al género pero con las 100 keywords como clave de cada par clave-valor.

ATRIBUTOS POR PELÍCULA

Película: My Life Without Me

budget	popularity	release_date	revenue	runtime	vote_average	vote_count	Action	Adventure	Animation	Comedy	Crime	Documentary	Drama	Family	Fantasy	History	Horror	Music	Mystery	Romance	Science Fiction	TV Movie	Thriller	War	Western
2500000	13.604	3/7/2003	12300000	106	5.8	364	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0

collection	short film	woman director	based on novel or book	murder	musical	concert	silent film	biography	sports	stand-up comedy	lgbt	christmas	world war ii	revenge	family	love	anime	philippines	based on true story	martial arts	friendship	romance	coming of age	softcore	black and white
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

kidnapping	wrestling	opera	new york city	ghost	based on play or musical	sequel	police	politics	serial killer	horror	pre-code	found footage	prison	holiday	erotic movie	drugs	vampire	zombie	rape	parent child relationship	death	gay interest	dance	remake	high school
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

art	dark comedy	monster	slasher	gay	religion	stop motion	dog	gore	time travel	supernatural	lost film	marriage	alien	football (soccer)	gangster	detective	superhero	nazi	africa	suicide	japan	rock 'n' roll	nature	mockumentary	racism
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

cartoon	fairy tale	road trip	satire	pregnancy	spy	small town	mystery	infidelity	investigation	thriller	noir	los angeles, california	avant-garde	school	robbery	dutch cabaret	london, england	1970s	sibling relationship	surrealism	anthology	france
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

CONJUNTO DE DATOS FINAL

id	budget	original_title	popularity	release_date	revenue	runtime	vote_average	vote_count	Action	Adventure	...
2	0	Ariel	8.165	10/21/1988	0	73	6.8	157	0	0	...
3	0	Varjoja paratiisissa	8.509	10/17/1986	0	74	7.2	158	0	0	...
5	4000000	Four Rooms	14.441	12/9/1995	4257354	98	5.7	2127	0	0	...
6	21000000	Judgment Night	13.336	10/15/1993	12136938	110	6.5	230	1	0	...
8	42000	Life in Loops (A Megacities RMX)	2.352	1/1/2006	0	80	7.5	18	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
285854	0	Reunited	0.600	1/26/2010	0	90	4.3	2	0	0	...
285855	0	Der Schrecken der Garnison	0.877	4/23/1931	0	86	0.0	0	0	0	...
285856	0	Big Muddy	2.304	9/4/2014	0	104	4.9	6	0	0	...
285857	0	ダロス	2.340	12/21/1983	0	83	6.3	9	0	0	...
285858	0	Corbo	3.649	4/17/2014	0	119	6.0	16	0	0	...

175188 rows × 128 columns

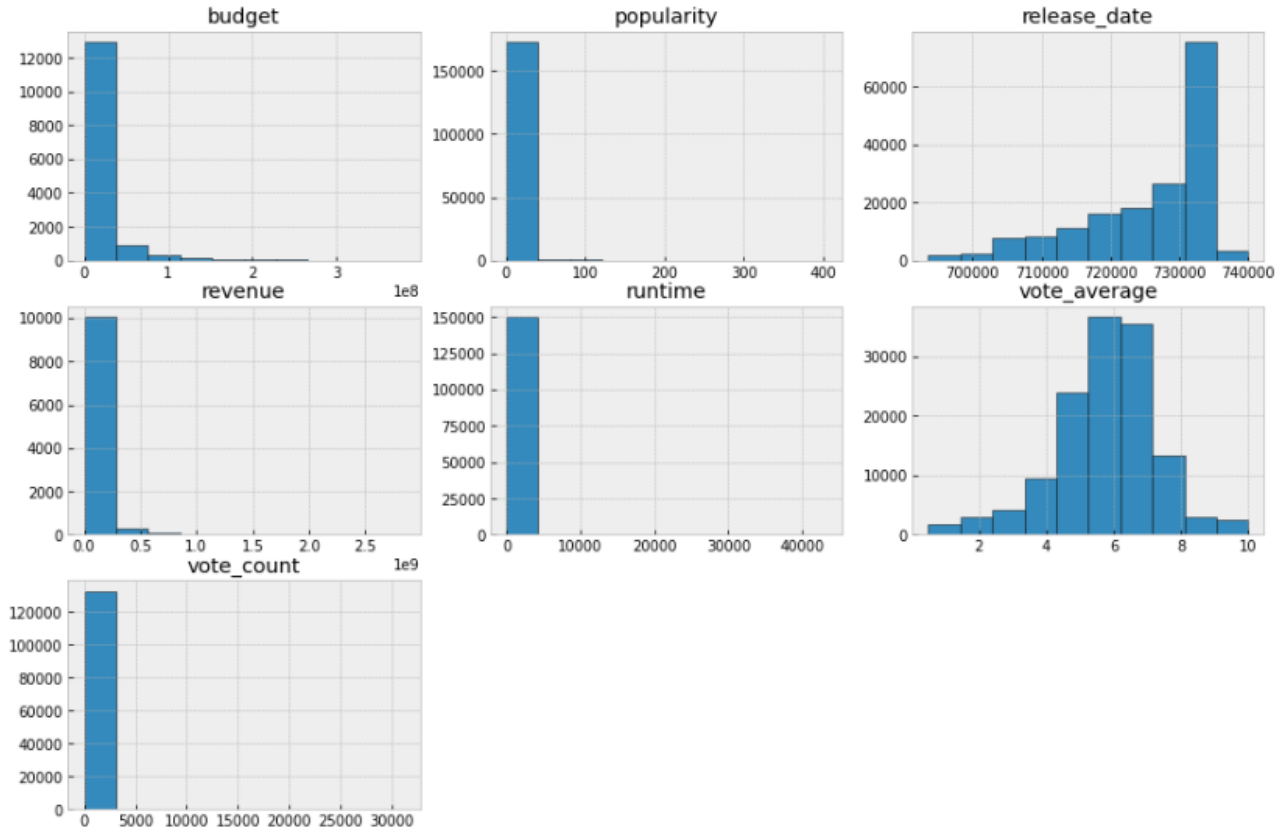
PREPROCESAMIENTO DE DATOS

1. Eliminación de películas con fecha antes de 1900
2. Eliminación del título de la película
3. Transformación de la fecha en formato ordinal
4. Reemplazo de los valores 0 a NaN

id	budget	popularity	release_date	revenue	runtime	vote_average	vote_count	Action	Adventure	Animation	...
2	NaN	8.165	726031.0	NaN	73.0	6.8	157.0	0	0	0	...
3	NaN	8.509	725296.0	NaN	74.0	7.2	158.0	0	0	0	...
5	4000000.0	14.441	728636.0	4257354.0	98.0	5.7	2127.0	0	0	0	...
6	21000000.0	13.336	727851.0	12136938.0	110.0	6.5	230.0	1	0	0	...
8	42000.0	2.352	732312.0	NaN	80.0	7.5	18.0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
285854	NaN	0.600	733798.0	NaN	90.0	4.3	2.0	0	0	0	...
285855	NaN	0.877	705030.0	NaN	86.0	NaN	NaN	0	0	0	...
285856	NaN	2.304	735480.0	NaN	104.0	4.9	6.0	0	0	0	...
285857	NaN	2.340	724265.0	NaN	83.0	6.3	9.0	0	0	1	...
285858	NaN	3.649	735340.0	NaN	119.0	6.0	16.0	0	0	0	...

174399 rows × 127 columns

ANÁLISIS DESCRIPTIVO DE LOS DATOS

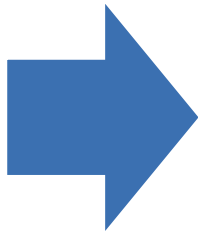


NORMALIZACIÓN DE LOS DATOS

Reasignación de outliers:

x Función `MinMaxScaler` de la biblioteca `scikitlearn`

✓ Asignar a valores por encima del percentil 90 el valor correspondiente a dicho percentil

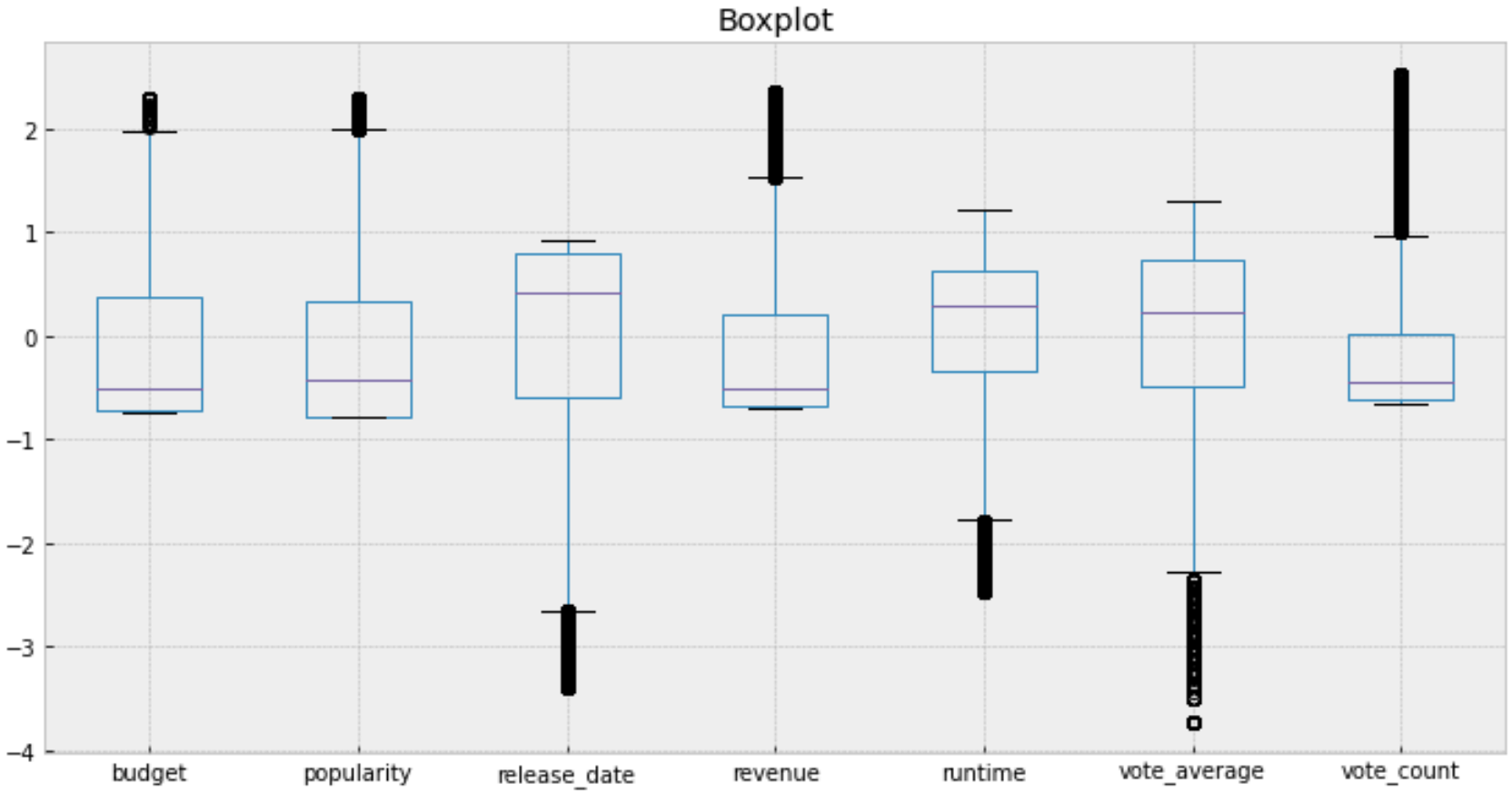


Escalamiento de datos:

- Función `StandardScaler` de la biblioteca `scikitlearn`

$$z_i = \frac{x_i - \mu}{\sigma}$$

BOXPLOT DE DATOS NORMALIZADOS



CONJUNTO DE DATOS NORMALIZADO

id	budget	popularity	release_date	revenue	runtime	vote_average	vote_count	Action	Adventure	Animation	...
2	NaN	2.293564	0.021899	NaN	-0.244393	0.794846	2.521822	0	0	0	...
3	NaN	2.293564	-0.055239	NaN	-0.213398	1.081410	2.521822	0	0	0	...
5	-0.436829	2.293564	0.295295	-0.600407	0.530483	0.006795	2.521822	0	0	0	...
6	0.847295	2.293564	0.212909	-0.422166	0.902424	0.579923	2.521822	1	0	0	...
8	-0.735803	0.168989	0.681093	NaN	-0.027428	1.296334	0.066497	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
285854	NaN	-0.783238	0.837049	NaN	0.282523	-0.996179	-0.622717	0	0	0	...
285855	NaN	-0.632686	-2.182164	NaN	0.158543	NaN	NaN	0	0	0	...
285856	NaN	0.142900	0.931819	NaN	0.716454	-0.566333	-0.450414	0	0	0	...
285857	NaN	0.162467	-0.163443	NaN	0.065557	0.436641	-0.321186	0	0	1	...
285858	NaN	0.873920	0.931819	NaN	1.181380	0.221718	-0.019655	0	0	0	...

174399 rows × 127 columns

MÉTRICA UTILIZADA EN EL MODELO

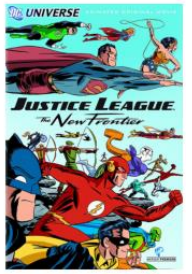
Distancia Euclidiana NaN

$$d(\vec{u},\vec{v}) = \sqrt{w\sum(u_i - v_i)^2}$$

Donde *w* es el peso y se define como: $w = \frac{\text{Dimensión del vector}}{\text{Coordenadas presentes}}$

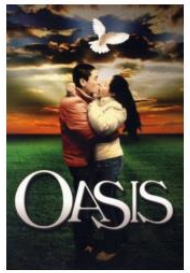
PRUEBA DEL MODELO

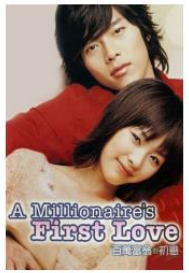
"Batman: Gotham Knight (2/2/2014)"	Lista de recomendaciones
	Star Wars: The Clone Wars Superman: Unbound All Star Superman Thor: Tales of Asgard The Invincible Iron Man Battle for Terra Justice League: The Flashpoint Paradox 銃夢 -GUNNM The Transformers: The Movie Justice League: The New Frontier

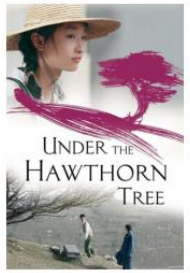


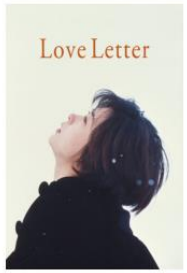
PRUEBA DEL MODELO

"The Notebook"	Lista de recomendaciones
	오아시스 君に届け 파이란 백만장자의 첫사랑 甜蜜蜜 山楂树之恋 ラブレター Stellet Licht Tabu Welcome









PARÁMETROS PARA MEJORAR EL MODELO

Asignación de pesos a las características del conjunto:

- Determinar un nivel de importancia para cada característica y realizar el escalamiento acorde al peso asignado

Eliminación/Incorporación de características al conjunto de datos:

- Eliminar fecha, algunas palabras clave, géneros, etc.
- Añadir al conjunto idioma original, país origen, etc.

CONCLUSIONES

- La extracción de datos se ha llevado a cabo de manera exitosa mediante el uso de la API de "The Movie DB", siendo este un proceso arduo y tardado ya que el sistema no permite obtener los datos mediante una sola consulta, de modo que éste se convierte en un proceso iterativo.
- El resultado del modelo basado en las pruebas realizadas parece ser bastante bueno, es decir, proporciona películas con características muy similares.
- El modelo puede mejorarse incorporando atributos o bien manipulando la selectividad mediante la asignación de pesos por atributo, esto hace que un atributo particular tenga una mayor importancia para el modelo.
- Se puede concluir que el modelo puede hacerse más complejo añadiendo restricciones y haciéndolo más selectivo teniendo en cuenta algunos criterios como región o país, idioma, actores, compañías productoras, etc.

BIBLIOGRAFÍA

1. De 2013 a 2014: INEGI. Módulo sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares. Consultado https://www.inegi.org.mx/temas/ticshogares/
2. Fkih, F. (2021). Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2021.09.014
3. Paialunga, P. (2022, 22 enero). Hands-on Content Based Recommender System using Python. Medium. Recuperado 27 de marzo de 2022, de https://towardsdatascience.com/hands-on-content-based-recommender-system-using-python-1d643bf314e4
4. Shankhar, B. S. (2021, 14 diciembre). Hyperparameters tuning in practice: Pandas vs. Caviar. Medium. Recuperado 28 de marzo de 2022, de https://medium.com/optimizing-hyperparameters/hyperparameters-tuning-in-practice-pandas-vs-caviar-82ab9763d8af
5. Buitinck et al.(2013). API design for machine learning software: experiences from the scikit-learn project.

¡GRACIAS!