Ranking de películas mejor valoradas desde 2014

Juan Herranz Martín

11 de noviembre de 2019

Contexto

Se desea llevar a cabo una recolección de información de las películas mejor valoradas por los usuarios de una plataforma web desde el año 2014. Se trata de la plataforma *filmaffinity*, web de referencia de información cinematográfica, con un gran número de títulos (tanto clásicos como estrenos), sus valoraciones y sus críticas.

Descripción

El conjunto de datos extraído contiene información referente a películas publicadas en los últimos 5 años, organizadas también por el género al que pertenecen. Para cada género y año se obtiene información de hasta 30 películas con la mejor valoración obtenida hasta la fecha de captura.

Representación gráfica

A continuación se muestran las 10 primeras películas (justo a sus atributos) que contiene el conjunto de datos:

	Título	Año	Género	Puntuación	N_Votos	Director/a
2	Redada asesina 2 (The Raid 2)	2014	AC	7,0	7.199	Gareth Evans
3	Big Hero 6	2014	AC	7,0	34.883	Chris Williams, Don Hall
4	Ataque a los Titanes: Sin lamentos (parte 1)	2014	AC	7,0	259	Tetsurō Araki
5	Guardianes de la galaxia	2014	AC	6,9	57.376	James Gunn
6	X-Men: Días del futuro pasado	2014	AC	6,9	44.328	Bryan Singer
7	Kingsman: Servicio secreto	2014	AC	6,8	36.843	Matthew Vaughn
8	Ataque a los Titanes, la película. Parte 1. El arco y la flecha escarlatas	2014	AC	6,8	718	Tetsurō Araki
9	Al filo del mañana	2014	AC	6,8	38.354	Doug Liman
10	La LEGO película	2014	AC	6,7	23.195	Phil Lord, Christopher Miller
11	Kenshin, el guerrero samurái 2: Infierno en Kioto	2014	AC	6,7	1.488	Keishi Ohtomo

Local_carátula	Web_carátula
Pictures/the_raid_2_berandal-549665158-msmall.jpg	https://pics.filmaffinity.com/the_raid_2_berandal-54966
Pictures/big_hero_6-799861615-msmall.jpg	https://pics.filmaffinity.com/big_hero_6-799861615-msr
Pictures/shingeki_no_kyojin_kui_naki_sentaku_1-753889741-msmall.jpg	https://pics.filmaffinity.com/shingeki_no_kyojin_kui_naki
Pictures/guardians_of_the_galaxy-595487268-msmall.jpg	https://pics.filmaffinity.com/guardians_of_the_galaxy-59
Pictures/x_men_days_of_future_past-252906568-msmall.jpg	https://pics.filmaffinity.com/x_men_days_of_future_past
Pictures/kingsman_the_secret_service-485444831-msmall.jpg	https://pics.filmaffinity.com/kingsman_the_secret_servic
Pictures/shingeki_no_kyojin_zenpen_guren_no_yumiya_attack_on_titan_part_i_crimson_bow_and_arrow-977564146-msmall.jpg	https://pics.filmaffinity.com/shingeki_no_kyojin_zenpen_
Pictures/edge_of_tomorrow-632023834-msmall.jpg	https://pics.filmaffinity.com/edge_of_tomorrow-632023
Pictures/the_lego_movie-819614387-msmall.jpg	https://pics.filmaffinity.com/the_lego_movie-819614387
Pictures/ruroni_kenshin_kyoto_taika_hen_ruroni_kenshin_the_great_kyoto_fire-395445708-msmall.jpg	https://pics.filmaffinity.com/ruroni_kenshin_kyoto_taika_

Contenido

Los datos se han recogido mediante *web scraping* con la ayuda de la librería BeautifulSoup, en el periodo temporal que comprende desde el inicio de 2014 hasta octubre de 2019. Cada registro corresponde a una película, y los campos que se han recogido son:

- **Título**: Nombre de la película.
- **Año**: Año del estreno de la película.
- **Género**: El género cinematográfico al que pertenece.
- **Puntuación**: La calificación media obtenida hasta la fecha.
- **N_votos**: El número de votos o valoraciones que ha recibido la película.
- **Director/a**: El director/directora/directores de la película.
- **Local_portada:** URL local de acceso a la imagen descargada de la portada de la película
- **Web_portada:** URL web de acceso a la imagen de la portada de la película.

Agradecimientos

Los datos se han recogido desde la web de datos cinematográficos online FilmAffinity (https://www.filmaffinity.com/es/main.html): principal web prescriptora de cine en Internet con una completa base de datos. Para ello, se han llevado a cabo técnicas de *Web Scraping* en Python sobre las páginas HTML de esta web que contenían la información que se deseaba obtener.

Inspiración

Este conjunto de datos ofrece varias posibilidades de análisis. Por una parte, puede resultar interesante para estudiar el éxito de cierto género o cierto director/a en base a las valoraciones de los usuarios en los últimos años, y con ello poder elaborar modelos predictivos sobre los posibles éxitos futuros.

Por otra parte, haciendo uso del contenido gráfico obtenido, podría entrenarse un algoritmo de *machine learning* de reconocimiento de imágenes con redes neuronales que fuera capaz de clasificar las películas por género según su portada o encontrar patrones para la elaboración de una nueva portada.

Licencia

La licencia seleccionada para este conjunto de datos es CC BY-SA 4.0 License y los motivos de esta selección tienen que ver con las posibilidades que ofrece el conjunto de datos y las condiciones de la licencia:

- El beneficiario de la licencia tiene el derecho de copiar, distribuir, exhibir y representar la obra y hacer obras derivadas siempre y cuando reconozca y cite la obra de la forma especificada por el autor o el licenciante.

- El beneficiario de la licencia tiene el derecho de distribuir obras derivadas bajo una licencia idéntica a la licencia que regula la obra original.

De esta forma, se garantiza el reconocimiento del autor del conjunto de datos de manera adecuada cuando se realizan cambios sobre éste. Además, las obras derivas del conjunto original, se distribuirán con los mismos términos iniciales que rigen ese conjunto.

Esta licencia también permite el uso comercial, que es lo idóneo para este conjunto de datos, fomentando el desarrollo de diversos proyectos y análisis junto a la propagación de la autoría del conjunto.

Código fuente

El código fuente en Python puede encontrarse en:
https://github.com/jherranzma/BestFilms/blob/master/codigo_best_films.py

Dataset

El dataset generado, en formato CSV, puede encontrare en:
https://github.com/jherranzma/BestFilms/blob/master/films.csv

Recursos

- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Tutorial de Github: https://guides.github.com/activities/hello-world