



Universitat Oberta
de Catalunya

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

López Ibáñez, Adrián

Busquets Aran, Nil

09/11/2020

Práctica 1

Contenido

Objetivos	3
Actividades	3
Actividad 1.....	3
Actividad 2.....	4
Actividad 3.....	4
Actividad 4.....	5
Actividad 5.....	5
Dataset	5
Extracción Datos.....	7
Actividad 6.....	11
Actividad 7.....	13
Actividad 8.....	14
Actividad 9.....	14
Actividad 10.....	14
Bibliografía	15

Objetivos

Los objetivos concretos de esta práctica son:

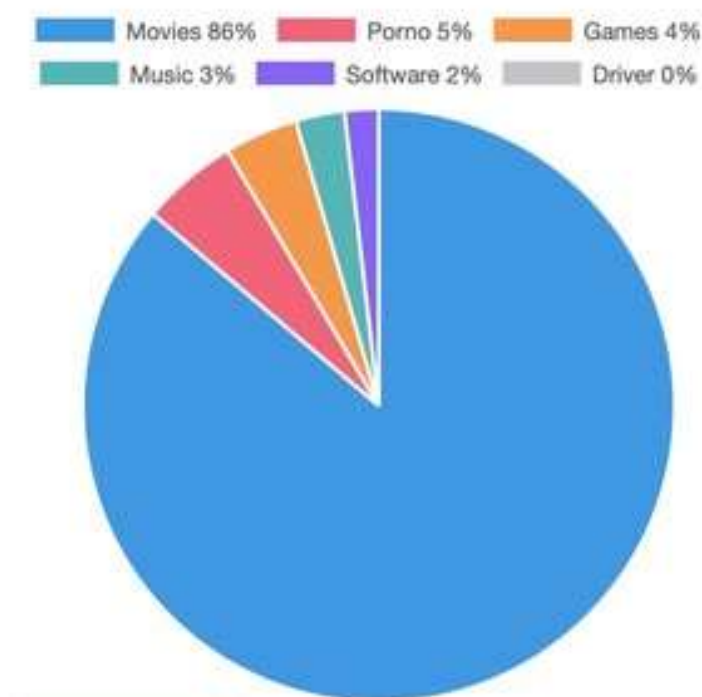
- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes que su tratamiento aporta valor a una empresa y la identificación de nuevos proyectos analíticos.
- Saber identificar los datos relevantes para llevar a cabo un proyecto analítico.
- Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos o repositorios) y mediante diferentes mecanismos (tales como queries, API y scraping).
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Actividades

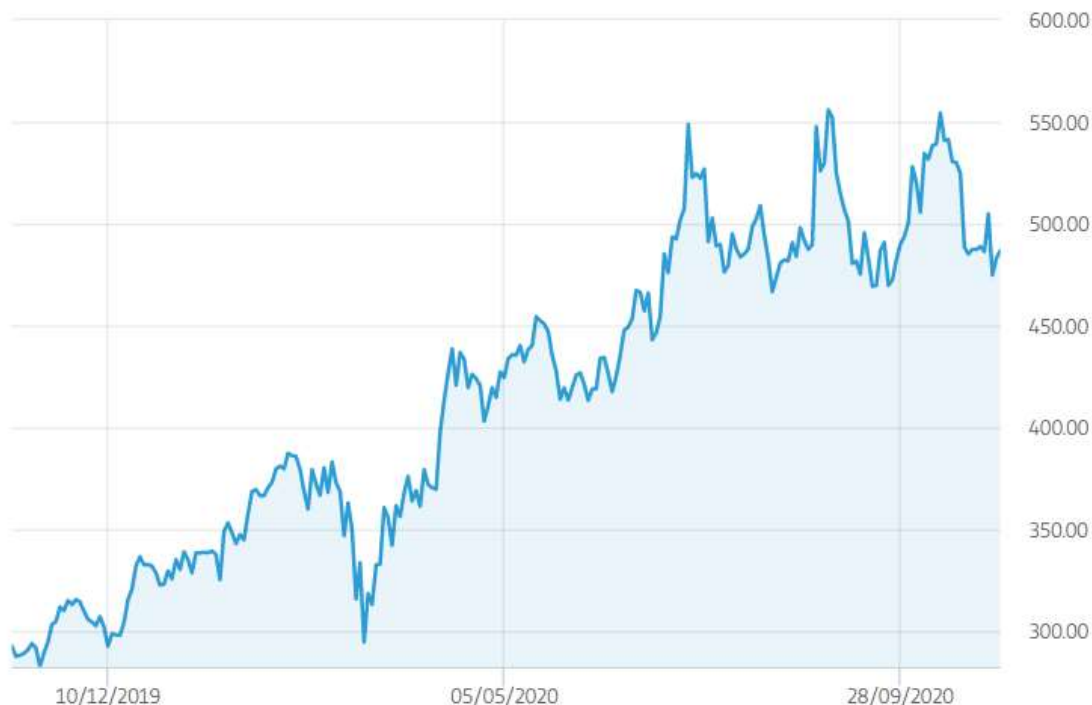
Actividad 1

Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Dada la situación actual de pandemia, hemos hecho una breve investigación de que comportamiento ha tenido la sociedad, con lo que hemos podido observar un aumento masivo de descargas de películas y series, como podemos ver en la siguiente imagen.



Dada esta información, hemos supuesto que el consumo mundial de series y películas de las diferentes plataforma habría aumentado durante el confinamiento, con lo que hemos hecho una investigación a nivel macroeconómico para verificar el aumento del uso de ciertas plataformas de series y películas online. Podemos ver en la siguiente imagen el aumento del valor de la plataforma de Netflix, esto nos confirma el aumento del uso de dichas plataformas.



Finalmente visto el potencial de recolectar estos datos en la situación actual, nos hemos dispuesto a buscar una web que contenga la información de distintas plataformas para poder hacer una comparativa, con lo que hemos encontrado la siguiente web con toda la información necesaria: <https://flixable.com/>

Actividad 2

Definir un título para el dataset. Elegir un título que sea descriptivo.

Netflix and Disney+ films and Shows in Spain, con este título estás identificando que contiene tu dataset grosso modo, series y películas de distintas plataformas.

Actividad 3

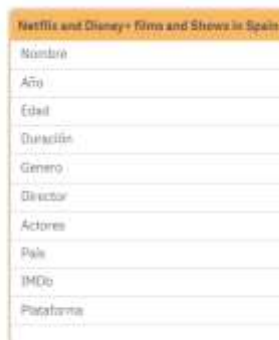
Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El datasets que se extrae, la idea principal es obtener por cada plataforma todas las películas y series que contienen seleccionando un año de inicio de los datos, año de fin de los datos y puntuación mínima deseada, con lo que podemos obtener un dataset más ligero con la información que deseamos. Entonces por cada película i/o serie tenemos la siguiente información que las completa: Nombre, Año, Edad, Duración, Genero, director, Actores, País, IMDb, Plataforma

Actividad 4

Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente

Como podemos ver en la siguiente imagen, nuestro dataset es una sola tabla, con los siguientes campos.



Nombre
Año
Edad
Duración
Género
Director
Actores
País
IMDb
Plataforma

Vista previa de datos

Nombre	Año	Edad	Duración	Género	Director
Bob Esponja: Un héroe al rescate	2020	TODOS LO	95 min	Comedias, Para toda la familia	Tim Hill
Brooklyn Nine-Nine	2019	13+	6 temporadas	Comedias, Estadounidenses, Policiacas	NA
Carmen: ¿Quién mató a María Marta?	2020	13+	1 temporada	Docuseries, Extranjeras, Policiacas	Alejandro Hartmann
Operación Feliz Navidad	2020	7+	95 min	Comedias, Para toda la familia, Románticas	Martin Wood
Outlander	2018	16+	4 temporadas	Acción, Ciencia ficción y fantásticas, Dramas	NA
Paranormal	2020	16+	1 temporada	Ciencia ficción y fantásticas, Dramas, Extranjeras	NA
Amor y anarquía	2020	13+	1 temporada	Comedias, Dramas, Extranjeras	NA

Actividad 5

Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Dataset

Como ya hemos comentado anteriormente los campos que obtenemos de nuestro web scraping son los siguientes:

Nombre: Contiene el nombre de la película y/o serie de la plataforma.

Año: Este año corresponde al año de rodaje de la película.

Edad: Se refiere al público al que va orientado la película o serie.

Duración: Tenemos la duración en minutos de la película, en el caso de las series tenemos el número de temporadas.

Género: El género cinematográfico es el tema general de una película que sirve para su clasificación

Director: El nombre del director/es de la película o serie.

Actores: El nombre de los actores principales de la película, en este caso no aparecen los actores secundarios.

País: De que país es la película, de todas formas, puedes ser de un país, pero mundialmente distribuida.

IMDb: Nos proporciona la puntuación media por parte de los usuarios, IMDb es una base de datos en línea de información relacionada con películas, programas de televisión, videos caseros, videojuegos y contenido en streaming en línea.

Plataforma: Con ello sabemos si la plataforma se trata de Netflix o Disney +

A continuación, se muestran un seguido de imágenes con lo que se puede ver una pequeña muestra de los datos que contiene el dataset:

Nombre	Año	Edad	Duraci...	Genero	Director
10 cosas que odio de ti	1999	9	98 min	Comedia, Comedia romántica, De adolescente a adulto	Gil Junger
Annie	1999	AL	91 min	Drama, Familiar, Musical	Rob Marshall
De ladrón a policía	1999	13+	94 min	Acción, Comedias	Les Mayfield
Deep Blue Sea	1999	16+	105 min	Acción, Ciencia ficción y fantásticas, Terror	Renny Harlin
Doble traición	1999	13+	106 min	Thrillers	Bruce Beresford
Doug: La película	1999	6	80 min	Animación, Comedia, Familiar	Maurice Joyce
Dudley de la montaña	1999	7+	83 min	Comedias, Para toda la familia	Hugh Wilson
Eyes Wide Shut	1999	16+	159 min	Clásicas, Dramas, Premiadas	Stanley Kubrick

Actores	País	IMDb	Plataforma
4Minute, B1A4, BtoB, ELSIE, EXID, EXO, Got7, INFINITE, KARA, Shinee, Sistar, VIXX, Nine Muses, BTS, Secret, Topp Dogg	Corea del Sur	6.1/10	Netflix
50 Cent, Adewale Akinnuoye-Agbaje, Joy Bryant, Omar Benson Miller, Tory Kittles, Terrence Howard, Ashley Walters, Marc John Jefferies, Viola Davis, Sullivan Walker, Serena Reeder, Bill Duke,	Estados Unidos, Canadá	5.4/10	Netflix
A.J. LoCascio, Sendhil Ramamurthy, Fred Testa, Jake Johnson, Lauren Lapkus, Zachary Levi, BO Wong, David Gunning	Estados Unidos	5.7/10	Netflix
A.J. Trauth, Spencer Breslin, Lalaine, Sally Stockwell, Peter Feeney, Tim Reid	Nueva Zelanda, Estados Unidos	6.1/10	Disney
Aaditi Pohankar, Vijay Varma, Vishwas Kini, Kishore Kumar G., Shivaní Rangole, Suhita Thatte, Sandeep Dhabale, Saqib Ayub	India	6.5/10	Netflix
Aamir Khan, Ashutosh Gowariker	India	8.7/10	Netflix
Aamir Khan, Darsheel Safary, Tanay Chheda, Tisca Chopra, Vipin Sharma, Girija Oak, M.K. Raina	India	8.4/10	Netflix
Aamir Khan, Gracy Singh, Rachel Shelley, Paul Blackthorne, Kulbhushan Kharbanda, Raghuvir Yadav, Yashpal Sharma, Rajendranath Zutshi, Rajesh Vivek, Aditya Lakhia	India	8.1/10	Netflix
Aamir Khan, Monica Dogra, Kriti Malhotra, Prateik, Aasha Pawar, Jyoti Pawar, Norma Lobo, Kiku Gidwani, Danish Hussain, Jehan Maneekshaw	India	7.0/10	Netflix

Los datos que hemos extraído se encuentran películas y series desde inicios de 1920 hasta fecha de hoy 2020.

Extracción Datos

Como breve explicación de cómo están distribuidos los datos dentro de nuestra web tenemos que: tiene todas las películas y series de diferentes plataformas como Netflix y Disney+



Como podemos ver en la imagen la web nos proporciona una información muy reducida en su página principal, que es nombre de la película o serie, el año de la serie y su puntuación de IMDb, pero dentro de cada título expuesto en la pantalla principal te proporcionan más información como actores, director, duración etc... Podemos ver un ejemplo de una serie a continuación:



Para la extracción de datos, inicialmente parametrizamos la URL en función de la plataforma de la que vamos a extraer los datos y después se le añaden otros parámetros que se piden al usuario una vez se ejecuta el programa, que son: a que año comienzan los datos, a que año terminan y la puntuación mínima de la que queremos las películas/series. Seguidamente, mediante el paquete Selenium, abrimos la web de donde queremos extraer los datos y scroleamos hasta abajo del todo, ya que se trata de una página web extensible y no fija.

#Abrir Firefox

```
driver = webdriver.Firefox(executable_path="geckodriver.exe")
driver.get(url)
driver.maximize_window()
```

#Desplazamiento pagina despacio

```
time.sleep(1)
iter=1
while True:
    scrollHeight = driver.execute_script("return document.docum
entElement.scrollHeight")
    Height=250*iter
    driver.execute_script("window.scrollTo(0, " + str(Height) + ");")
    if Height > scrollHeight:
        print('Final de la pagina')
        break
    time.sleep(1)
    iter+=1
```


Se añaden tiempos de espera(`time.sleep(1)`), para prevenir que nos detecten como bot. Una vez toda la pagina cargada guardamos las puntuaciones en una lista, ya que es el dato que tenemos a la pagina principal.

```
#Leer html una vez se ha recorrido toda la página
body = driver.execute_script("return document.body")
source = body.get_attribute('innerHTML')

#Parsear con BeautifulSoup
soup = BeautifulSoup(source, 'lxml')

#Guardar clasificación imdb en lista
an = soup.find_all('div', class_='card-description')
annoMedia = list()
for i in an:
    annoMedia.append(i.text)
annoMedia = [x.replace('\n', '') for x in annoMedia]
annoMedia = [x.replace(' ', '') for x in annoMedia]
anno = list()
for word in annoMedia:
    anno.append(word[:4])
listImdb = list()
for word in annoMedia:
    listImdb.append(word[-6:])
listImdb = [x.replace(' ', '') for x in listImdb]
cambiar = [len(set(i)) == 1 for i in zip(anno, listImdb)]
i=0
for word in cambiar:
    if word == True:
        listImdb[i] = 'NA'
    i += 1
```

A continuación, lo que se hace es recoger todas las URLs de las películas / series que nos quedan una vez filtradas.

```
#Encontrar todos los enlaces a peliculas y series en la página
titulosDivs = soup.findAll('div', attrs={'class' : 'card-header card-
header-image'})
titulosEnlace = list()
for div in titulosDivs:
    titulosEnlace.append('https://es.flixable.com' + div.find('a')['
ref'])
```

Seguidamente se crean todas las listas vacías en las que se va a almacenar la información que queremos recoger y por cada URL que hemos obtenido, entramos dentro de ella y recogemos la información restante.

```

for url in titulosEnlace:
    t0 = time.time()
    response = requests.get(url)
    response_delay = time.time() - t0
    time.sleep(10 * response_delay)

    soup = BeautifulSoup(response.content, "html.parser")

    #Título de la película
    pel = soup.find('h1', class_='title text-left')
    listTitPeli.append(pel.text)
    #año
    anyo = soup.find('span', class_='mr-2')
    listAnyo.append(anyo.text)
    #edad recomendada
    edad = soup.find('span', class_='border border-secondary mr-2 px-1')
    listEdad.append(edad.text)
    #duracion
    duracion = soup.find_all('span')[11].text #Span en la posición 11
    (no tiene clase)
    listDuracion.append(duracion)
    #Generos
    if plataforma == 'Netflix':
        genero = soup.find_all('span')[13].text
    else:
        genero = soup.find_all('span')[14].text
    listGenero.append(genero)
    #Director
    if plataforma == 'Netflix':
        director = soup.find_all('span')[15].text
    else:
        director = soup.find_all('span')[16].text
    listDirector.append(director)
    #Actores
    if plataforma == 'Netflix':
        actores = soup.find_all('span')[17].text
    else:
        actores = soup.find_all('span')[18].text
    listActores.append(actores)
    #País
    if plataforma == 'Netflix':
        pais = soup.find_all('span')[19].text
    else:
        pais = soup.find_all('span')[20].text
    listPais.append(pais)
    #Fecha añadido
    if plataforma == 'Netflix':
        fecha = soup.find_all('span')[25].text
    else:
        fecha = soup.find_all('span')[26].text
    listFecha.append(fecha)

```

Finalmente creamos el dataframe y extraemos los datos en formato CSV.

```
#Extracción a dataframe
df = pd.DataFrame({'Nombre': listTitPeli, 'Año': listAnyo, 'Edad': listEdad, 'Duración': listDuracion, 'Genero': listGenero, 'Director': listDirector, 'Actores': listActores, 'País': listPais, 'Incorporación': listFecha, 'IMDb': listImdb, 'Plataforma': plataforma})
print(df)
if os.path.isfile('catalogo.csv') == False:
    df.to_csv('catalogo.csv', index=False, mode='a', encoding="utf-8")
else:
    df.to_csv('catalogo.csv', header=False, index=False, mode='a', encoding="utf-8")
```

Actividad 6

Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Nuestro propietario de datos es la siguiente web: <https://es.flixable.com/>

A la hora de analizar el contenido de robots.txt, ya que es en este archivo donde la mayor parte de páginas web indican las restricciones, hemos obtenido que hay una exclusión de todos los robots para la carpeta /dev.

```
User-agent: *
Disallow: /dev/
```

Mediante la siguiente petición obtenemos los datos del propietario:

```
import whois
print(whois.whois('https://es.flixable.com'))
```

Y nos devuelve la siguiente información del propietario:

```
{
  "domain_name": [
    "FLIXABLE.COM",
    "flixable.com"
  ],
  "registrar": "NAMECHEAP INC",
  "whois_server": "whois.namecheap.com",
  "referral_url": null,
  "updated_date": [
    "2020-08-03 06:22:06",
    "2020-08-03 06:22:07.030000"
  ],
  "creation_date": "2016-09-02 19:50:09",
  "expiration_date": "2021-09-02 19:50:09",
  "name_servers": [
    "DNS1.REGISTRAR-SERVERS.COM",
    "DNS2.REGISTRAR-SERVERS.COM",
    "dns1.registrar-servers.com",
    "dns2.registrar-servers.com"
  ],
  "status": "clientTransferProhibited https://icann.org/epp#clientTrans-ferProhibited",
  "emails": [
    "abuse@namecheap.com",
    "2930c32f51e24a5db9be7d377ff43a9a.protect@whoisguard.com"
  ],
  "dnssec": "unsigned",
  "name": "WhoisGuard Protected",
  "org": null,
  "address": "P.O. Box 0823-03411",
  "city": "Panama",
  "state": "Panama",
  "zipcode": null,
  "country": "PA"
}
```

Actividad 7

Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Como previamente hemos explicado, en la situación actual en la que nos encontramos el uso de estas plataformas a aumentado muchísimo debido al tener que permanecer mucho más tiempo cerrados en nuestras viviendas y el cierre del ocio en nuestro país. La idea principal era recoger el dataset de todas las películas y series que se han incorporado en estas dos plataformas, Netflix y Disney+ y hacer un breve estudio de si hay una diferencia notable o no entre el año anterior de la pandemia y el año actual. Unas de las preguntas que queríamos responder son:

¿Se han publicado más películas en 2020 que 2019 debido a la pandemia?

Hemos generado el siguiente gráficos para cada una de las plataformas y podemos ver que no se han publicado más películas/series hasta noviembre.



A que publico se han enfocado las películas/series este año 2020?

Se puede observar una ligera reducción para el publico de 13+, pero por lo demás es similar



Actividad 8

Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

- o Released Under CC0: Public Domain License
- o Released Under CC BY-NC-SA 4.0 License
- o Released Under CC BY-SA 4.0 License
- o Database released under Open Database License, individual contents under Database Contents License
- o Other (specified above) o Unknown License

Licencia Creative Commons Attribution 4.0 International

Con esta licencia usted es libre de:

Compartir — copiar y redistribuir el material en cualquier medio o formato

Adaptar — remezclar, transformar y construir a partir del material para cualquier propósito, incluso comercialmente.

Debido a la dificultad de extracción de los datos no se veía reflejada la idea de colocar una licencia de dominio público, pero como los datos extraídos son de una página web de un tercero, la licencia más optima es una que deje al publico disfrutar de los datos, pero, por nuestra parte, tienen el deber de hacer referencia a nuestro trabajo.

Actividad 9

Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

- GitHub → <https://github.com/adrianlope/WebScrappingFlixable>

Actividad 10

Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

- Zenodo → <https://zenodo.org/record/4249736#.X6VFOmhKguU>
- DOI → 10.5281/zenodo.4249736

Bibliografía

Sabán, A., 2020. *Además Del Uso De Netflix, La Cuarentena También Ha Disparado Las Descargas Torrent*. [online] Genbeta.com. Available at: <<https://www.genbeta.com/intercambio-de-ficheros/uso-netflix-descargas-torrent-tambien-se-han-disparado-durante-cuarentena>> [Accessed 4 November 2020].

eToro. 2020. [online] Available at: <<https://www.etoro.com/markets/nflx/chart>> [Accessed 4 November 2020].

Creativecommons.org. 2020. *Creative Commons — Atribución 4.0 Internacional — CC BY 4.0*. [online] Available at: <<https://creativecommons.org/licenses/by/4.0/deed.es>> [Accessed 6 November 2020].

Selenium-python.readthedocs.io. 2020. *2. Getting Started — Selenium Python Bindings 2 Documentation*. [online] Available at: <<https://selenium-python.readthedocs.io/getting-started.html>> [Accessed 6 November 2020].

Crummy.com. 2020. *Beautiful Soup Documentation — Beautiful Soup 4.9.0 Documentation*. [online] Available at: <<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>> [Accessed 2 November 2020].

Fredgibbs.net. 2020. *Extract, Transform, And Save CSV Data • Fredgibbs.Net*. [online] Available at: <<http://fredgibbs.net/tutorials/extract-transform-save-csv.html>> [Accessed 3 November 2020].

Contribuciones	Firma
Investigación previa	Adrián López Ibáñez, Nil Busquets Aran
Redacción de las respuestas	Adrián López Ibáñez, Nil Busquets Aran
Desarrollo código	Adrián López Ibáñez, Nil Busquets Aran