Tipología y Ciclo de Vida de los Datos

Webscraping

PRAC 1

Ana Blanes Martínez
Xavier Castilla Carbonell

Índice

1. Contexto.

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

Las redes sociales juegan cada vez más un papel más importante en nuestras vidas y con ellas se han creado figuras muy influyentes que son vistas por miles sino millones de personas. Estos "influencers" tienen la capacidad de generar tendencias y corrientes de opinión dada su alta visibilidad. En este contexto es interesante conocer cuales son las figuras más influyentes de cada una de las redes sociales más populares del momento.

SocialBlade nos presenta que plataformas sociales son más influyentes en el momento (con mayor afluencia de visitas) y nos presenta en una tabla por cada plataforma con los datos de los usuarios con mayor seguimiento.

Nuestro estudio permite conocer a los usuarios más influyentes de Internet lo que sirve de punto de partida para realizar otros estudios sobre estas personalidades, estudios que pueden determinar las tendencias del momento, el nivel de influencia de estos usuarios sobre su audiencia, estudio de mercado para publicidad, etc.

Además, los ficheros de datos .csv que se generan se actualizan cada cierto tiempo con lo que también tenemos un registro actualizado de que usuarios son los más influyentes de cada plataforma en todo momento.

2. Definir un título para el dataset.

Elegir un título que sea descriptivo.

El título escogido es **TopInfluencers:** como hemos comentado en el punto anterior la página procesada es SocialBlade, la cual ofrece un conjunto de estadísticas e informes para aquellos usuarios registrados con información relevante y de interés sobre redes sociales; además, en abierto y accesible, ofrece a modo de ejemplo un listado del top 25 de canales ordenado en función de su número de seguidores, lo que definimos como los influencers top. Esta es la información que procesamos y por tanto la razón del título de dataset.

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).

Con el objetivo de ofrecer un conjunto de datos que sea la base de futuros estudios ofrecemos varios datasets, uno con la información concreta de cada plataforma (para estudios de dicha plataforma) y otro con la información en común de todas las plataformas (para estudios de redes sociales en general):
TopInfluencers{$nombreplataforma}.csv contiene los datos de los usuarios más influyentes de una de las plataformas, los nombres de las variables de estos datos están ligados a las nomenclaturas de cada plataforma.
TopInfluencers.csv contiene los datos comunes de los usuarios de todas las plataformas sociales y las variables son las comunes entre todas las plataformas.

Los datos de nuestro dataset se actualizan hora y todas las variables son propensas al cambio, si se quiere hacer un estudio exhaustivo se deberá almacenar un snapshot del dataset para evitar su actualización.

Los datos del dataset son siempre numéricos o caracteres, no hay imágenes ni otros tipos de datos. Si se puede dar el caso de que existan NA en variables que no están informadas, como se desean tratar estos datos queda a disposición de quien los use, por lo demás no hace falta hacer limpieza de los datos a menos que se busque normalizar alguno de ellos.

Qué tipos de datos se pueden ver en cada uno de los ficheros se describe más adelante.

4. Representación gráfica.

Presentar una imagen o esquema que identifique el dataset visualmente.

A continuación mostramos la representación gráfica del dataset escogido. El formato utilizado para ello es una tabla, ya que de esta manera se puede ver fácilmente para cada una de las plataformas cuyos datos son procesados, los atributos que tienen asociados. De igual manera podemos ver dónde aparecen cada uno de los atributos.

	Youtube	Twitch	Twitter	Instagram	Facebook	Dailymotion	Mixer	Common
Rank	X	X	X	X	X	X	X	X
Grade	X	X	X	X	X	X	X	X
Display name	X		X	X				X
Videos	X							
Subscribers	X							
Views	X	X						
Username		X	X	X	X	X	X	X
Last game		X						
Followers		X	X	X		X	X	X
Tweets			X					
Following			X	X				
Media				X		X		
Category					X			
Likes					X			
Talking about					X			
Vidviews						X		
Channel views							X	
Level							X	
Latest game							X	

5. Contenido.

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

Para la propuesta escogida se generan varios datasets con sus ficheros csv correspondientes: uno por cada plataforma que es procesada más uno que contiene los datos comunes a todas ellas; aunque todos comparten una estructura parecida algunos presentan algunas particularidades, por lo que vamos a detallarlos uno a uno. De cara a facilitar la explicación de los campos que son compartidos y los que no, se van a presentar primero los campos para, a continuación, explicar el significado de cada uno de ellos.

TopInfluencers.csv: Rank, Grade, Name, Followers/Suscribers

TopInfluencersYoutube.csv: Rank, Grade, Display name, Videos, Subscribers, Views

TopInfluencersTwitchtv.csv: Rank, Grade, User name, Last game, Views, Followers

TopInfluencersTwitter.csv: Rank, Grade, User name, Display name, Tweets, Followers, Following

TopInfluencersInstagram.csv: Rank, Grade, User name, Display name, Media, Followers, Following

TopInfluencersFacebook.csv: Rank, Grade, User name, Category, Likes, Talking about

TopInfluencersDailymotion.csv: Rank, Grade, User name, Display name, Media, Followers, Vidviews

TopInfluencersMixer.csv: Rank, Grade, User name, Followers, Channel views, Level, Latest game

- **Rank**: Es la posición en el ránking. Para cada una de las plataformas se indica la posición de un canal en función del número de suscriptores/seguidores.
- **Grade**: según explica SocialBlade, el número de suscriptores/seguidores no es lo único que debe contar a la hora de puntuar un canal: también el número medio de visitas así como su comparación con otros sites determinan la puntuación que viene dada en esta columna y que puede tener los valores A/B/C (++/+/-/—) …
- **User name**: El nombre del usuario propietario del canal
- **Display name**: El nombre mostrado del usuario en el canal
- **Followers**: Número de seguidores que presenta un canal en concreto.
- **Subscribers**: Número de suscriptores que presenta un canal en concreto.
- **Following**: Número de canales que son seguidos por un canal en concreto.
- **Una serie de campos específicos** para cada plataforma donde se registran los datos correspondientes; los nombres son intuitivos y permiten saber qué datos contienen: Videos Views, Last game, Tweets, Media, Category, Likes, Talking about, Media, Vidviews, Channel views, Level.

Según indica la página SocialBlade las estadísticas las actualizan una vez al día, sin embargo si se detecta que alguna plataforma/canal está recibiendo poco tráfico estas actualizaciones pueden darse con menos frecuencia, mostrándose el resultado tras varios días. De manera que para la ejecución del código de webscraping implementado en este trabajo se puede optar por ejecutar el proceso una vez al día, y así se garantiza que los datos aparezcan sincronizados con respecto a SocialBlade.

La forma de recogerlos es mediante un conjunto de programas sencillos y específicos que son ejecutados a través de un programa principal. Cada uno de estos programas se encargan de contactar con la página web de una de las plataforma y procesar los datos contenidos en ella, para luego volcarlos en un fichero csv. Aparte, se realiza una nueva llamada sobre todas las webs de las plataformas para obtener aquellos datos que son parte del fichero común, ya que son de columnas que comparten todas las platformas.

6. Agradecimientos.

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

SocialBlade es un website que, haciendo uso de las diferentes herramientas proporcionadas por diferentes plataformas/redes sociales a través de API's, etc. y mediante el correspondiente acuerdo legal con las mismas y aceptación de sus condiciones y términos de uso, proporciona un valor añadido a estos datos mediante su organización y elaboración de estadísticas de interés basados en dichos datos. Por tanto los datos mostrados son propiedad de SocialBlade, al ser resultado de un procesamiento y manipulación única por su parte en la que a los datos originales se les proporciona un valor añadido.

Referente a análisis anteriores similares al que se presenta en esta práctica, simplemente decir que el propio Social Blade se dedica precisamente al manejo de los mismos datos, esto es, presentar estadísticas y datos sobre estas redes sociales aunque obviamente a un nivel profesional e infinitamente superior al aquí ofrecido. Por tanto, análisis anteriores sobre los mismos datos que ofrecemos (top influencers en redes sociales), podemos mencionar a SocialBlade como compañía que accede, procesa, maneja y ofrece estos datos.

7. Inspiración.

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

Se han tenido en cuenta dos aspectos principales para llevar a cabo un proyecto de webscraping sobre este site:
- Los datos obtenidos resultado del proceso son de interés.
- La frecuencia de actualización es lo suficientemente alta como para que merezca la pena un desarrollo ad-hoc para disponer de estos datos actualizados.

Las páginas específicas que se han procesado presentan en abierto unos datos de gran interés, como son estadísticas correspondientes a las plataformas/redes sociales más importantes que existen en estos momentos; estos datos permiten por ejemplo saber cuáles son los canales más visitados, el número de visitas, etc. y permite realizar estudios sobre la evolución del ránking, suscriptores, etc. En estos tiempos en los que la subsistencia de los canales se debe en gran parte a la existencia de patrocinadores, anunciantes, etc. es fundamental saber la repercusión de los canales y su impacto en el público, por lo que el conocimiento de estadísiticas precisas sobre ellos puede ser significante en el éxito o el fracaso de un canal.

Relacionado con ello tenemos también el hecho de que estos datos solo son interesantes si se encuentran actualizados, y por tanto es fundamental que los procesos se ejecuten con la frecuencia adecuada, sin por supuesto sobrecargar los servidores, para garantizar que los datos proporcionados sean útiles de cara a un propósito determinado.

8. Licencia.

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

Como se ha explicado previamente, SocialBlade procesa datos proporcionados por un conjunto plataformas para mostrar estadísticas y datos relacionados de interés. Este site proporciona un acceso premium de pago donde se aportan más estadísticas de valor para el usuario.

Se han revisado los términos y condiciones legales proporcionadas por SocialBlade y en ellas menciona que los datos pueden ser utilizados, incluyendo pantallazos e imágenes, siempre y cuando se mencionen referencias al site SocialBlade. Es cierto que el site tiene una parte pública abierta con un conjunto de datos que no presentan restricciones en el fichero robots.txt y otra de acceso restringido mediante registro y pago en el que no están permitidas el botting, crawling y scraping. De manera que teniendo en cuenta que las páginas que son procesadas en este trabajo son accesibles sin restricciones pero que se trata de un site comercial que explota económicamente los datos a través de un registro con pago, se puede inferir que el tipo de licencia para los datos en abierto accedidos es **Released Under CC BY-NC-SA 4.0 Licensem**, ya que permite la distribución de estos datos públicos accesibles siempre y cuando se indique la referencia a la compañía y no se haga un uso comercial de ellos.

9. Participantes:

Contribuciones	Firma
Investigación previa	XCC, ABM
Redacción de las respuestas	XCC, ABM
Desarrollo código	XCC, ABM

10. Recursos:

- Subirats, L., Calvo, M. (2018). ***Web Scraping***. Editorial UOC.
- Lawson, R. (2015). ***Web Scraping with Python***. Packt Publishing Ltd. Chapter 2. Scraping the Data.