Tipología y ciclo de vida de los datos

Práctica 1: Web scraping

César Aguilera Padilla

Daniel Velasco Torre

Máster en Ciencia de Datos
Curso 2020-2021

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

Se ha elegido hacer web scraping al sitio web http://meneame.net, que resulta ser el mayor agregador de noticias en español.

Según Wikipedia:

"Menéame es un sitio web y red social basado en la participación comunitaria en el que los usuarios registrados envían noticias que los demás usuarios del sitio (registrados o no) pueden votar, promoviendo las más votadas a la página principal (portada) mediante la aplicación de un algoritmo que unifica varios parámetros en un único valor numérico que denomina internamente «karma». Al igual que Digg, del cual es una traducción modificada, Menéame combina marcadores sociales, el blogging y la sindicación con un sistema de publicación sin editores."

En su portada, Menéame proporciona una lista de noticias de actualidad que han sido votadas por sus usuarios. Cada noticia dispone de información interesante como puede ser el título, sus etiquetas, su categoría o su fuente.

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

Noticias de la portada de Menéame.net

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).

De la portada de Menéame se han extraído una serie de noticias, en las que cada noticia posee:

- *Un título*
- *Una entradilla*
- *Etiquetas*
- *Votos*
- *Votos negativos*
- *Número de clicks*
- *Número de comentarios*
- *Fecha de envío*
- *Fecha de publicación*
- *Categoría*
- *Karma*
- *Fuente*

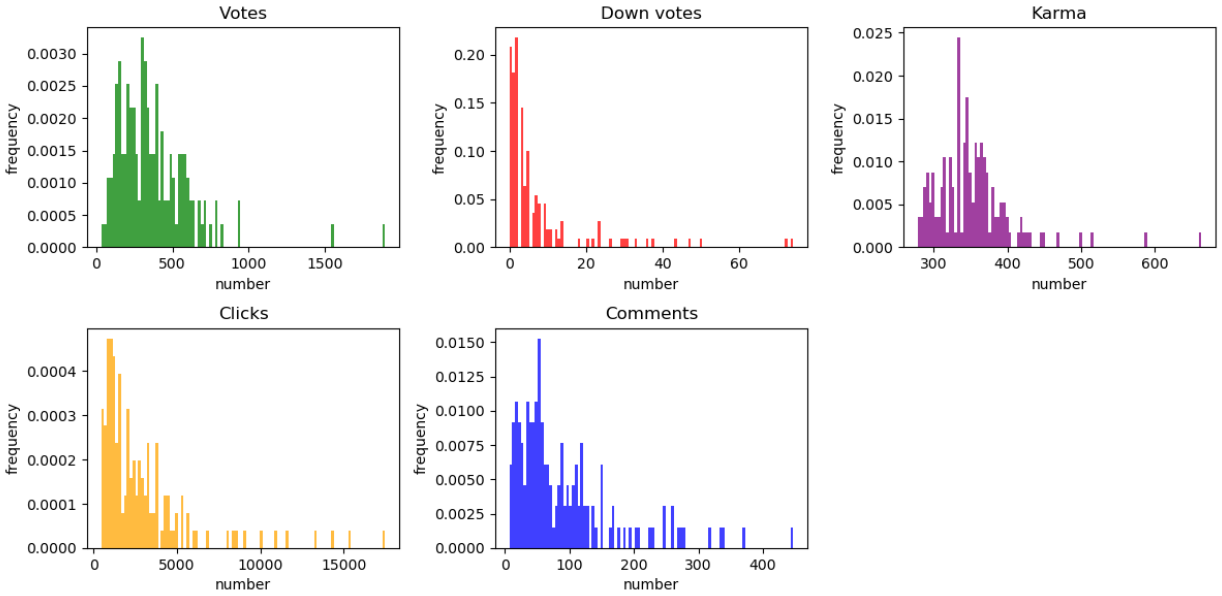
Cada uno de los ítems anteriores se corresponde con una columna para cada noticia presente en el dataset.

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

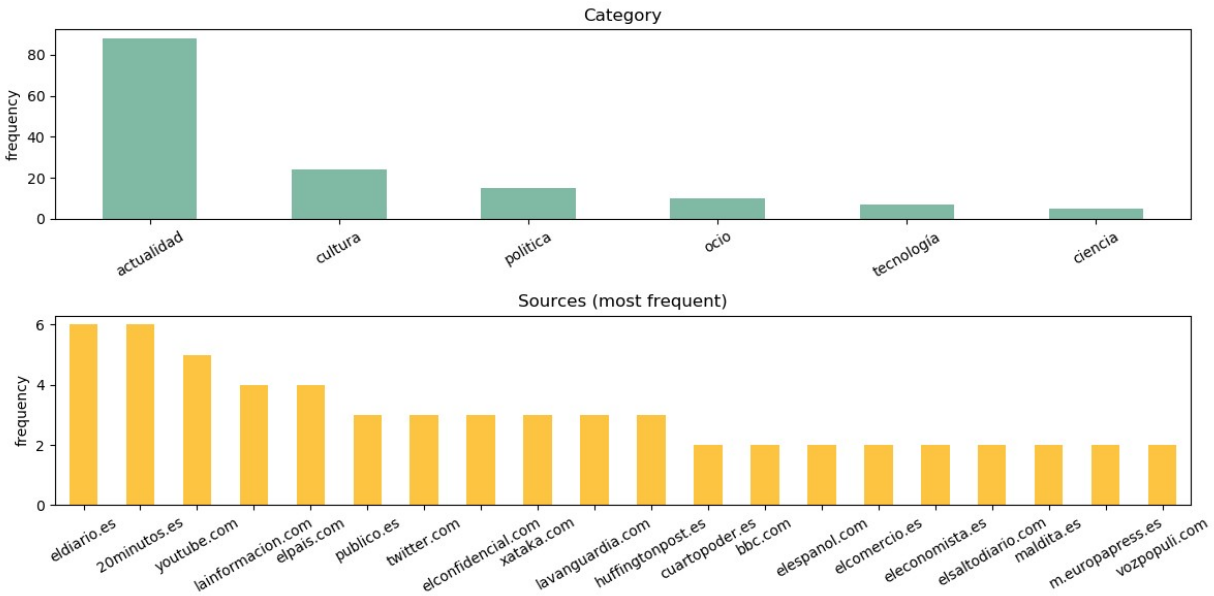
La siguientes imágenes muestran los valores más característicos de un dataset generado con MeneameScraper conteniendo 150 noticias extraídas entre las fechas 2020-10-17 y 2020-10-20.

Se recuerda al lector que los gráficos pueden ser obtenidos directamente con el programa usando la opcion *"-g"*.

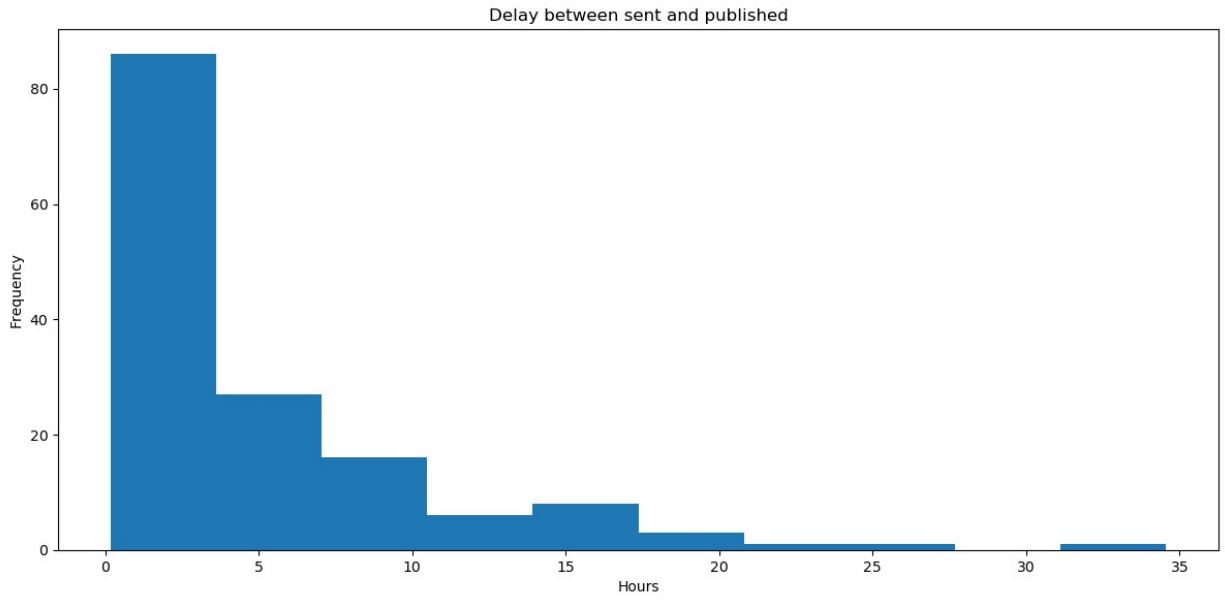
Comenzamos mostrando los histogramas de los atributos numéricos de nuestro dataset de ejemplo. En ellos vemos como para un número no muy grande de noticias (150) los histogramas muestran distribuciones tendiendo a normales, lo cual indica que valores como votos negativos, postivos, número de comentarios etc suelen tener valores similares en torno a una media para todas las noticias.



Para los valores categóricos también mostramos la frecuencia de cada uno de ellos (o los más frecuentes). En este caso también se puede destacar como la categoría *actualidad* es claramente la mas usada, seguida de *cultura* y *política*. En cuanto a las fuentes, vemos como hay infinidad y las noticias se distribuyen bastante entre ellas.



Finalmente, podemos hacer un análisis de la diferencia en tiempo entre cuando la noticia ha sido enviada por el colaborador y cuando se ha publicado. Donde se puede ver que la mayoría de noticias de publican entre 0 y 5 horas desde que se reciben.



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

El dataset que se da como ejemplo contiene 150 instancias de noticias de las cuales se han extraido los siguientes campos (tal y como se explica en el punto 3):

- *Title*: Título de la noticia *(String)*
- *Paragraph*: Párrafo donde se resume la noticia *(String)*
- *Tags*: etiquetas separadas por comas (,) relacionados con la noticia *(String)*
- *Votes*: Número absoluto de votos positivos de la noticia *(Integer)*
- *Down-votes*: Número absoluto de votos negativos de la noticia *(Integer)*
- *Clicks*: Número absoluto de clicks hacia la noticia *(Integer)*
- *Comments*: Número absoluto de comentarios en la noticia *(Integer)*
- *Sent-date*: Fecha de envío de la noticia por el colaborador *(String)*
- *Pub-date*: Fecha de publicación de la noticia *(String)*
- *Category*: etiqueta única indicando la categoría de la noticia *(String)*
- *Karma*: numero de karma acumulado por la noticia en relación a las personas que han votado *(Integer)*
- *Source*: Fuente de la noticia *(String)*

Para ello se han listado todas las URLs de las noticias comprendidas en las fechas dadas como atributo (17-10-2020 y 20-10-2020) y de cada una de ella se ha obtenido su contenido mediante web scraping.

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

A la profesora Mireia Calvo González por revisar la práctica antes de la entrega final.

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

El análisis de los datos de una página tan conocida como Menéame.net puede proveer de un flujo constante de información y conocimiento sobre como la sociedad acepta o rechaza ciertas noticias.

Menéame.net además de contener todo tipo de noticias y actuar como un concentrador de todo tipo de fuentes de información, permite una interacción con el usuario final (lector) que puede interactuar con la noticia y con el resto de lectores. Esto permitiría por ejemplo extraer datos muy concretos de noticias con determinadas etiquetas como por ejemplo política y realizar perfiles de tipo de lectores o predecir qué clases de noticias se valoran más de un medio con respecto a otro.

Como ejemplo, de los datos obtenidos mediante nuestra aplicación de web scrapping se podría extraer la siguiente información:

- ¿Cuál es la temática más/menos valorada de noticias?
- ¿Cuál es la temática que más hace interactuar al usuario?
- ¿Cuál atrae más al usuario? ¿Cual tiene más clicks?
- ¿Alguna de las fuentes esta mejor valorada que otra simplemente por el hecho de ser quién es? ¿Alguna menos?
- ¿Hay alguna correlación entre cierto contenido y el número de votos/clicks (análisis del lenguaje natural)?

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

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

La licencia que hemos escogido es la Creative Commons BY-NC-SA 4. Entre otras, por que:

1. Queremos que el dataset se pueda compartir, copiar y distribuir en cualquier medio o formato
2. Permitimos que se adapte y transforme el dataset para cualquier propósito, incluso comercial
3. Queremos que se nos atribuya la autoría y que se nos de crédito de manera adecuada, obligando a que las nuevas distribuciones de este dataset contengan un enlace a la licencia e indiquen los cambios que se hayan realizado
4. Queremos que cualquier transformación o modificación del dataset se distribuya bajo la misma licencia

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

Repositorio con el código: https://github.com/Cs4r/MeneameScraper

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

Link al repositorio en Zenodo: https://zenodo.org/record/4122041

11. Contibruciones.

Contribuciones	Firma
Investigación previa	C.A.P y D.V.T
Desarrollo de código	C.A.P y D.V.T
Redacción de las respuestas	C.A.P y D.V.T

12. Tecnologías.

- Lenguaje de programación utilizado: Python 3
- Bibliotecas de Python utilizadas (de más a menos importantes):
  - *Requests*: Para descargar el contenido que se va a *scrapear* (Para hacer peticiones *HTTP*)
  - *BeautifulSoup*: Para hacer web scraping sobre el contenido descargado
  - *Pandas*: Para leer, escribir y tratar los datos extraídos
  - *MatplotLib*: Para crear gráficos estadísticos a partir de los datos
  - *Time*
  - *DateTime*
  - *ArgParse*
  - *Sys*
  - *Os*

13. Buenas prácticas que se han seguido para el desarrollo del web scraping.

La web https://www.meneame.net/ impide el web scraping devolviendo un error 403 (Forbidden, Prohibido) cuando se hace uso de las cabeceras HTTP por defecto de la biblioteca *BeautifulSoup*.

Para solventar este problema y poder hacer web scraping a Menéame.net, hemos falseado las cabeceras HTTP antes de hacer la petición GET. Las cabeceras utilizadas son las siguientes:

```
HEADERS = {
    "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8,application/signed-exchange;v=b3",
    "Accept-Encoding": "gzip, deflate, br",
    "Accept-Language": "en-US,en;q=0.9,es;q=0.8",
    "Cache-Control": "no-cache",
    "dnt": "1",
    "Pragma": "no-cache",
    "Upgrade-Insecure-Requests": "1",
    "User-Agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.108 Safari/537.36"
}
```

Se puede ver como definimos las cabeceras en la línea 12 del fichero *main.py* y hacemos uso de ellas en la línea 48.