TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS	
Segundo Semestre	Máster en Ciencia de Datos
Práctica 1 – PRA1	

Apellidos: *Alonso Cabrera* Nombre: *Carlos A.*

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

El algoritmo presentado en esta práctica "Web Scraping ELPAIS calonsocab.py" extrae los titulares del diario online EL PAIS. Este algoritmo, en cooperación con otros algoritmos de análisis de datos, que se escapan al contenido de esta práctica, podría ejecutarse diariamente con el fin de obtener una información de actualidad en tiempo real.

Coincidiendo con las elecciones generales llevadas a cabo durante el día 10 de Noviembre de 2019, un posible uso del algoritmo desarrollado en el fichero "*Web Scraping ELPAIS calonsocab.py"* podría ser comparar los titulares sobre las elecciones de los distintos diarios online.

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

Información de actualidad

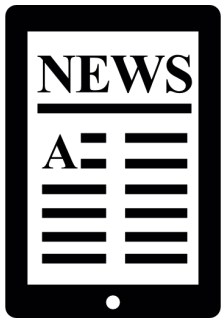
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El dataset obtenido es un conjunto de cadenas de texto separadas por el carácter ";" y almacenado en un fichero con extensión *.csv. Cada cadena supone un titular del diario online ELPAIS, que se corresponde con el titular que esté mostrando dicho diario online en el

momento de la ejecución del script “Web Scraping ELPAIS calonsocab.py”.

La dimensión del fichero será variable ya que dependerá de los titulares que el diario online esté mostrando.

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



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

El dataset contiene únicamente un atributo, que se corresponde con los titulares extraidos. El número de registros de dicho atributo se corresponderá con la cantidad de titulares extraidos, que a su vez dependerá de la cantidad de titulares que esté mostrando el diario online.

El periodo de tiempo de los datos almacenados en el fichero “calonsocab_PRA1.csv” se corresponde con el momento de la ejecución del script “Web Scraping ELPAIS calonsocab.py”. En particular, el dataset “calonsocab_PRA1.csv” contiene los titulares del 11 de Noviembre de 2019, del diario online ELPAIS.

En cuanto a la forma de extracción, en primer lugar, a través de la inspección del código fuente de la URL indicada anteriormente se ve que los titulares que quieren extraerse están contenidos en elementos <a> bajo la etiqueta 'h2' y con el atributo class='articulo-titulo'.

Y se usa un bucle para acceder a los elementos <a> contenidos en el objeto donde está almacenado en contenido de la web y se usa función para acceder al texto almacenado en el elemento <a>.

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

Agradezco a freeCodeCamp.org por animarme a realizar un web scraping extrayendo los titulares de una web.

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

El dataset presentado en esta práctica resulta interesante porque muestra los aspectos mas relevantes de la actualidad nuestra sociedad. Extrapolando, si adicionalmente extrajéramos los titulares de más diarios online, los ficheros resultantes podrían servir como entrada de un algoritmo que comparara los titulares de los diferentes medios, con el fin de buscar noticias convergentes.

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 a usar sería Released Under CC0: Public Domain License ya que se pretende que el dataset obtenido pueda ser usado por cualquier persona y para cualquier propósito.

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

Se adjunta en el enlace https://github.com/camayus/PRA1-Web-Scraping-UOC

10. Dataset. Presentar el dataset en formato CSV

Se adjunta en el enlace https://github.com/camayus/PRA1-Web-Scraping-UOC

Contribuciones	Firma
Investigación previa	**Carlos Alonso**
Redacción de las respuestas	**Carlos Alonso**
Desarrollo del código	**Carlos Alonso**

Referencias

[1]https://towardsdatascience.com/web-scraping-news-articles-in-python-9dd605799558

[2]https://github.com/vprusso/youtube_tutorials/tree/master/web_scraping_and_automation/beautiful_soup

[3]https://stackoverflow.com/questions/37289951/python-write-to-csv-line-by-line

[4]https://github.com/rafoelhonrado/foodPriceScraper/blob/master/foodPriceScraper.py

[5]https://github.com/datalifecicleuoc/web-scraping

[6]https://guides.github.com/activities/hello-world/

[7]https://www.youtube.com/watch?v=87Gx3U0BDlo

[8]https://code.tutsplus.com/es/tutorials/scraping-webpages-in-python-with-beautiful-soup-the-basics--cms-28211

[9]https://elpais.com/

[10]https://elpais.com/robots.txt

[11]https://www.geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/