Práctica 1: Web scraping

Alicia Escontrela y Beatriz Figueroa Martínez

1. Contexto.

Para aquellos interesados el mundo de las start ups, ya sea porque estén pensando en invertir como en emprender, es imprescindible estar al día sobre las tecnologías, bien informado sobre las novedades que se presentan en el mundo empresarial y las oportunidades de negocio que pueden surgir. El dataset que hemos creado nos proporciona una selección de noticias y tendencias actuales sobre emprendimiento empresarial con una orientación a la transformación digital.

Hemos elegido la revista Retina como objeto de nuestro Scraping ya que es un medio riguroso y con artículos de gran calidad relacionados con las nuevas tendencias desde una perspectiva tecnológica.

2. Definir un título para el dataset.

SCRAPING NEWS para emprendedores.

3. Descripción del dataset.

Generamos un conjunto de datos donde se detalla el título del artículo, la fecha en la que se publicó y quién es su autor. Además, aportamos el link a la noticia y la sección en la que se encuentra.

4. Representación gráfica.

5. Contenido.

El csv que generamos tiene los siguientes campos:

Title: Título del artículo
PageLink: Link al detalle del artículo
Date: Fecha de publicación
Section: Sección en la que se cataloga el artículo.
Author: Autor. En caso de no especificarse el autor, rellenamos

Recogemos todos los artículos en el apartado de emprendimiento empresarial, a lo largo de todas las páginas, empezando en el link
https://retina.elpais.com/tag/iniciativa_empresarial/a/

6. Agradecimientos.

Los datos se han obtenido de la página web de la revista Retina https://retina.elpais.com/, utilizando Python 2.7 y la librería BeautifulSoup para aplicar técnicas de web scraping.

En el conjunto de datos se incluyen los datos de los autores que han elaborado cada uno de los artículos.

Nos han servido de gran ayuda para realizar esta práctica siguientes páginas:

https://help.data.world/hc/en-us/articles/115006114287-Common-license-types-for-datasets

https://guides.github.com/activities/hello-world/

https://docs.python.org/2/library/robotparser.html

https://creativecommons.org/licenses/by-nc-sa/3.0/es/

7. Inspiración.

Este conjunto de datos permite conocer las últimas tendencias de emprendimiento en el sector de la innovación, lo cual sería de utilidad en varios contextos. Por ejemplo, en el sector de periodismo y marketing, facilita conocer las personas que están generando contenido en este sector y las últimas tendencias en el mercado.

Otra aplicación de utilidad es para PYMEs y emprendedores relacionados con el sector, que permiten conocer las tendencias y las personas con las cuales podrían contactar para establecer relaciones y participar en eventos que les permitan darse a conocer.

También es útil en el sector educativo y de investigación para facilitar la información a personas que busquen contenido relacionado con este sector.

Por otra parte, el código generado se puede extender a otras secciones dentro de la página o como referencia para aplicar a otras páginas de noticias similares para crear nuevos datasets.

8. Licencia.

La licencia seleccionada para este dataset es CCO – Public Domain License, lo cual hemos seleccionado para facilitar el uso del contenido de los artículos de esta revista a la investigación

y al emprendimiento. Teniendo en cuenta que este contenido se publica en la página web y en la revista que se puede adquirir en los kioscos, por tanto, entendemos que es de carácter público.

9. Tabla de contribuciones.

Contribuciones	Firma
Investigación previa	AER, BFM
Redacción de las respuestas	AER, BFM
Desarrollo código	AER, BFM