Descripción Práctica 1

Contexto, descripción del dataset e inspiración:

Se ha decidido realizar esta práctica debido al interés de contar con un dataset que incluyera información sobre el contenido (noticias o artículos de opinión) que permita realizar análisis usando NLP sobre los temas que tratan, los autores que más escriben en las secciones (llamadas focos en el medio elegido), búsqueda de tendencias en los temas o fechas de publicación, etc.

Para ello se ha elegido el medio generalista español **eldiario.es** y se han elegido arbitrariamente 5 focos/secciones:

- **Mejores Ciudades**
- **Vida Digital**
- **Creación Cultural**
- **Medio Ambiente**
- **Maltrato Animal**

Título del dataset:

- **eldiario_news.csv**

Representación gráfica:

Pendiente

Contenido:

El dataset presenta los siguientes 8 atributos:

- **headline:** Titular del contenido.
- **type:** Tipo de contenido (news o opinion).
- **date** Fecha de publicación del contenido.
- **section:** Sección en la que se ha publicado el contenido.
- **author:** Autor/es del contenido.
- **authorInfo:** Información opcional sobre el autor (si estuviera disponible).
- **location:** Información opcional sobre la ubicación de la noticia (si estuviera disponible).
- **url:** Dirección del contenido.

El dataset se compone de 3564 instancias obtenidas el día 6 de abril de 2019 a las 21:30 (hora local de Madrid) que engloba fechas de publicación de los años 2015-19.

Estas noticias se han recogido iterando sobre los directorios de noticias de cada una de las secciones/focos. Para evitar trampas de araña (spider trap), se ha definido manualmente la profundidad máxima para cada uno de los focos.

Agradecimientos:

Al medio de comunicación en cuestión (eldiario.es) por no excluir en el fichero robots.txt los directorios utilizados y por usar una licencia Creative Commons (CC-BY-SA), que me ha permitido realizar esta tarea de investigación de forma legal.

Licencia Dataset:

Se ha decidido elegir la licencia CC BY-SA 4.0 License debido a que es la versión actual que utiliza el medio de comunicación en cuestión (3.0) para cumplir con el término de uso que obliga a compartir bajo la misma licencia cualquier transformación de la obra original.

Además esta licencia abierta es la más correcta para compartir material que no sea software, como en este caso es el dataset creado.

Código:

El código utilizado se puede consultar en el presente repositorio en el fichero **Practica1_webScrapping.ipynb**.

Dataset:

El dataset creado se puede consultar en el presente repositorio en el fichero **eldiario_news.csv**.

Integrantes del grupo:

Debido a que trabajo y a que mis horarios me complican coordinarme con otras personas, el 20 de marzo el profesor colaborador en el aula Diego Pérez me autorizó por email realizar la práctica de forma individualizada sin penalización.

Contribuciones	Firma
Investigación Previa	GLRA
Redacción de las respuestas	GLRA
Desarrollo Código	GLRA