Descripción Práctica 1

Contexto:

Este conjunto de datos recopila diferentes contenidos periodísticos publicados por el medio generalista español eldiario.es durante los últimos 5 años aproximadamente.
Entre estos contenidos podemos encontrar tanto noticias como artículos de opinión relativos a los **cambios de las ciudades**, la **vida digital**, el **medio ambiente**, el **maltrato animal** y la **creación cultural**, que componen los 5 focos/secciones que han sido elegidos arbitrariamente dentro del medio.

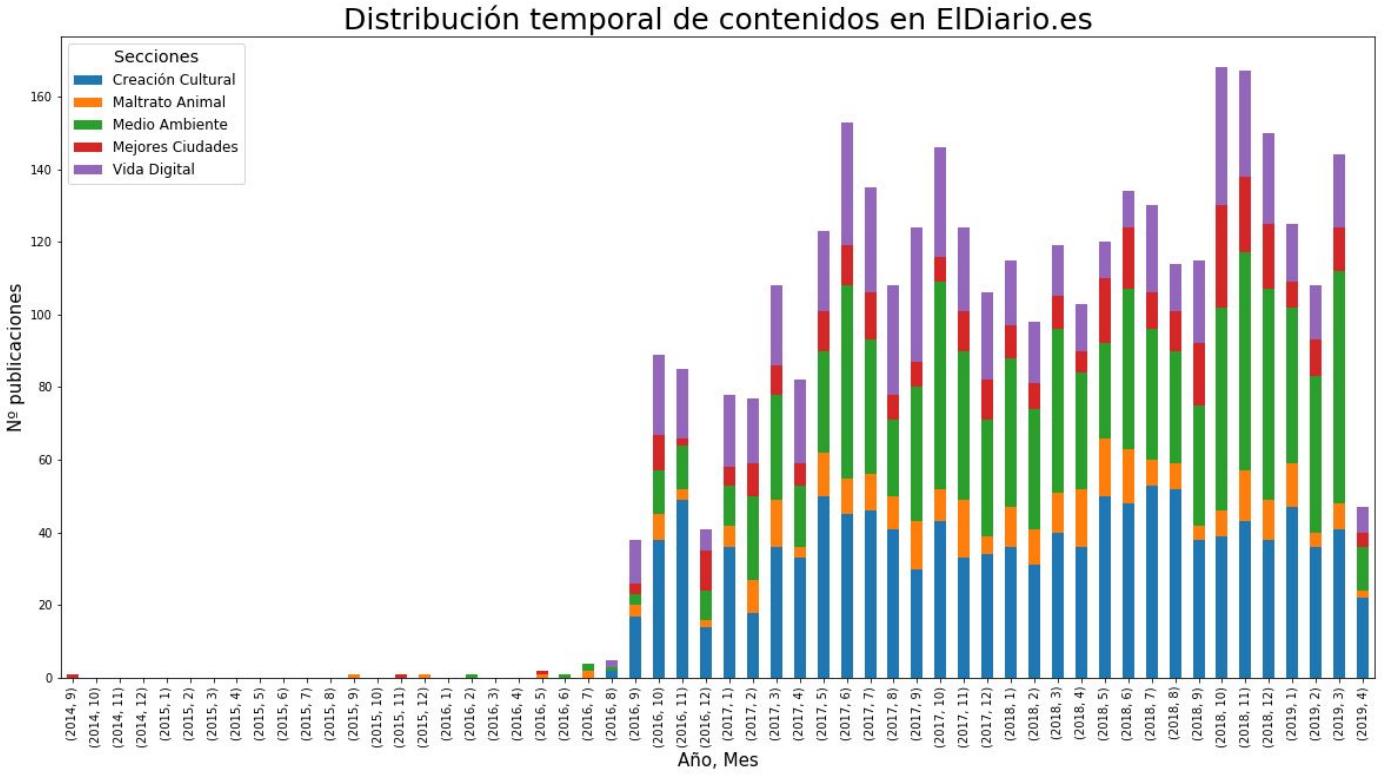
Título del dataset:

Conjunto de datos relativos a contenidos periodísticos publicados en un medio español generalista puramente digital para su uso en proyectos de procesamiento natural, clasificación de textos por tema, etc.

Descripción del dataset

El conjunto de datos recopilado como parte de la práctica de la asignatura de Tipología y Ciclo de Vida de los Datos contiene diferentes características de diferentes contenidos publicados por eldiario.es en 5 de sus secciones entre los años 2015 hasta el 2019.
Entre sus variables se encuentran el título, los autores, la fecha y la sección en la que se publicó.

Representación gráfica:



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 3591 instancias obtenidas el día 14 de abril de 2019 a las 21:00 (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.

Inspiración:

Se ha decidido realizar esta práctica debido al interés de contar con un dataset que incluyera información sobre contenido periodístico (noticias o artículos de opinión) de un medio español nativo digital 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, realizar resúmenes de textos automáticos usando redes neuronales recurrentes, etc.

También se podría utilizar los artículos de opinión que contiene para realizar análisis de sentimiento o minería de opinión, tratar de clasificarlos haciendo uso de algoritmos como **Kmeans** o **MeanShift** y observar si se forman diferentes clusters asociados a cada uno de los temas que tratan.

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. Tiene las siguientes características:

- **Cualquier persona es libre de:**
  - **Compartir** — copiar y redistribuir el material en cualquier medio o formato
  - **Adaptar** — remezclar, transformar y construir a partir del material para cualquier propósito, incluso comercialmente.
- **Bajo los siguientes términos:**
  - **Atribución** — se debe dar crédito de manera adecuada, brindar un enlace a la licencia, e indicar si se han realizado cambios. Puede hacerlo en cualquier forma razonable, pero no de forma tal que sugiera que cualquier persona o su uso tienen el apoyo de la licenciante.
  - ***CompartirIgual*** — Si remezcla, transforma o crea a partir del material, debe distribuir su contribución bajo la misma licencia del original.

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