SERGIO POSTIGO / ENDIKA MOMOITIO

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

Se ha recogido información de la página web del Boletín Oficial del Estado (BOE) porque cada día se publica información de interés para muchos ámbitos y sectores. Esté Boletín incorpora los acuerdos a los que ha llegado el propio Gobierno de España, más ahora con la crisis del Covid-19, y además incorpora mucha información de dominio público cómo formalizaciones de contrato, convenios colectivos, concurso de acreedores...

2, Definir un titulo para el dataset. Elegir un titulo que sea descriptivo.

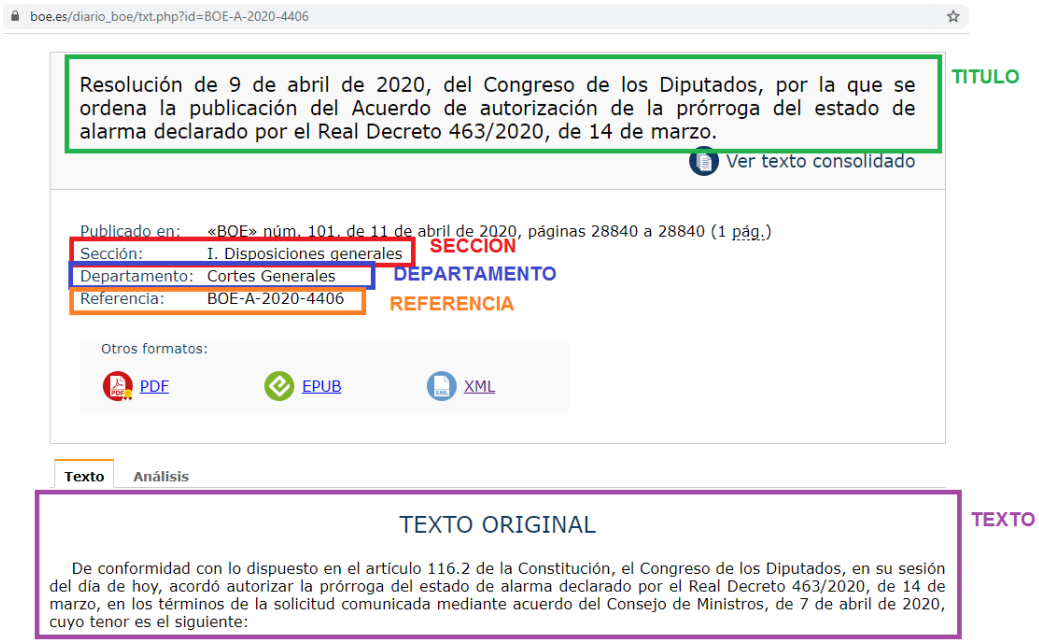
El título para el dataset es simplemente “BOE_FechaActual.csv”, donde “FechaActual” es el día actual en formato “YYYYMMDD”. Por ejemplo “BOE_20200413.csv” para exportar la información del Boletín Oficial del Estado del Domingo 13 de Abril de 2020.

3. Descripcion del dataset. Desarrollar una descripcion breve del conjunto de datos que se ha extraido (es necesario que esta descripcion tenga sentido con el titulo elegido).

El dataset es una exportación de la información de interés de cada artículo publicado en BOE del día en el que se realiza la exportación. Los campos están separados por el símbolo pipe (“|”) para evitar que la exportación del texto altere los campos debido a los delimitadores. El proceso está ideado para que el script se ejecute todos los días y genere un fichero cada día. Esta información podría analizarse con técnicas de NLP (Natural Lenguage Processing) para extraer entidades o conceptos destacados.

4. Representacion grafica. Presentar una imagen o esquema que identifique el dataset visualmente

En la siguiente imagen es posible ver las diferentes zonas según el dataset generado



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

La información que contiene el dataset es la siguiente:
- Referencia. Valor a partir del cual se puede acceder a la página original. Sería equivalente con una clave primaria si cargamos los datos en una base de datos.
- Departamento. Valor del departamento que ha publicado la información en el BOE.
- Sección. Valor de la sección a la que va dirigida la información.
- Título. Resumen de la información extraída del artículo del BOE
- Texto. Información extraída del artículo del BOE en el que aparece detallada la información.

Los datos son del día de ejecución. Está preparado para ejecutarlo cada día y así almacenar la información relevante del BOE de manera diaria.

Para recoger esta información se ha utilizado una técnica de scraping a través de la librería BeautifulSoup de Python. La página web no dispone de API pública de extracción de información.

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

Agradecer al Ministerio de Interior que gestiona el BOE la facilidad para acceder a los datos a través del mecanismo de scraping. Disponen de una URL que facilita todo el mapa web diario para poder acceder fácilmente a la información.

7. Inspiracion. Explique por que es interesante este conjunto de datos y que preguntas se pretenden responder.

La información e insights que se pueden extraer del BOE son casi infinitos… Algunos pueden ser:
- Detección de formalizaciones de contratos y servicios para saber que empresas ganan contratos y/o dinero de contratos públicos.

- Detección precoz de empresas que entran en concurso de acreedores.
- Análisis NLP para extraer entidades, analizar sentimiento y muchas más técnicas de extracción de información del lenguaje humano.
- Monitorización de empresas públicas para saber que contratos licitan y por los importes.
- Análisis en pliegos públicos por ejemplo, para saber que descuentos se realizan en los pliegos públicos
- …

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

CCO: Public Domain License. Es una práctica de un Master Universitario por lo que no se pretende licenciar de manera estricta ningún código y por eso se ha elegido que sea un script/documento de Domino Público.

9. Codigo. Adjuntar el codigo con el que se ha generado el dataset, preferiblemente

El código se puede obtener en el mismo repositorio GitHub en el que se ha encontrado este fichero.

10. Dataset. Publicacion del dataset en formato CSV en Zenodo con una pequena descripcion

La publicación del dataset en formato CSV se puede encontrar en la siguiente url:

```
https://doi.org/10.5281/zenodo.3751881
```

11. Entrega. Presentar el trabajo con el DOI del dataset en Github.

https://github.com/SergiP/TCVD_Practica1

Contribuciones	**Firma**
Investigacion previa	SP / EM
Redaccion de las respuestas	SP / EM
Desarrollo codigo	SP / EM