Tipología y ciclo de vida de los datos

PRACTICA 1:

¿Cómo podemos capturar los datos de la web?

Juan Javier Corrales Pérez

Juan Carlos Escribano Rubio

1. Contexto

En este proyecto, hemos decidido investigar sobre la plataforma de contratación del sector público, responsabilidad de la dirección general del patrimonio del estado y dependiente del ministerio de hacienda y administraciones públicas. Dicha plataforma se ocupa de facilitar la información de las licitaciones de todos los organismos que la componen y sus resultados, a fin de poder ofrecer y garantizar la transparencia y divulgación de la información hacia los ciudadanos y usuarios.

La información que alberga corresponde a cuatro grupos de perfiles de contratantes:

- ***Sector Público Estatal*** (Administración General del Estado, Mutuas de Accidentes de Trabajo Colaboradoras de la Seguridad Social y Entidades dependientes de la Administración General del Estado)
- ***Comunidades Autónomas*** (Junta de Comunidades de Castilla-La Mancha, Generalitat Valenciana, Govern de Illes Balears, Junta de Extremadura, Gobierno de Cantabria, Gobierno de Aragón, Ciudad Autónoma de Melilla, Junta de Castilla y León, Gobierno de Canarias, Ciudad Autónoma de Ceuta, Región de Murcia y Principado de Asturias)
- ***Entidades Locales*** (Perfiles del contratante de entidades locales dados de alta en la Plataforma y Entidades dependientes)
- ***Otros*** (Instituciones públicas independientes y Universidades)

La información que se va a recolectar son las licitaciones del sector público en España, para ello se ha utilizado la página:

https://contrataciondelestado.es/

2. Título

Licitaciones y contrataciones del sector público en España

3. Descripción del dataset

La información contenida en este dataset es la relativa a las licitaciones y contrataciones realizadas por el sector público de enero a octubre del 2022.

En este dataset, se presentan los datos referentes a todas las licitaciones publicadas por cada uno de los organismos que la componen. El campo numérico correspondiente al importe hace referencia a la cantidad en euros. De los datos obtenidos no se ha realizado ningún preprocesado, por lo que se presentan tal cual han sido extraídos. El formato en el que se presenta este dataset es un fichero CSV, para facilitar su tratamiento y visualización.

4. Representación gráfica

Imagen obtenida de https://pixabay.com/ (Imagen libre de derechos de autor)

5. Contenido

En este dataset se presentan los datos de todos los expedientes sacados a licitación desde el año 2002 hasta la fecha de ejecución del script de extracción de datos. Las características de los datos extraídos son las siguientes, en orden de aparición en cada fila del dataset:

- Expediente:
  - Descripción:
    - Identificador único de expediente
  - Tipo de campo:
    - Texto
- Objeto del expediente
  - Descripción:
    - Descripción textual del expediente del contrato
  - Tipo de campo:
    - Texto
- Tipo de contrato
  - Descripción:
    - Clasificación del tipo de contrato
  - Tipo de campo:
    - Texto/Categórico
- Estado
  - Descripción:
    - Estado en que se encuentra la tramitación de la licitación a la fecha de captura de la información
  - Tipo de campo:
    - Texto/Categórico
- Importe
  - Descripción:
    - Importe en el que se licita el expediente
  - Tipo de campo:
    - Numérico
- Fecha de presentación
  - Descripción:
    - Fecha en la que el contrato se hace público
  - Tipo de campo:
    - Fecha (dd-mm-YYYY)
- Órgano de contratación
  - Descripción:
    - Descripción del órgano del sector público que ha lanzado la licitación
  - Tipo de campo:
    - Texto
- Fecha de recuperación de la información
  - Descripción:

        - Fecha en la que se ha lanzado el script de webscraping
    - Tipo de campo
        - Fecha (dd-mm-YYYY)

6. Propietario

El propietario de los datos es el Estado Español a través del Ministerio de Hacienda y Función pública.

Hemos encontrado la siguiente empresa que se encarga de recopilar y agregar información al respecto:

https://datamarket.es/#contratacion-publica-dataset

Con respecto a la recopilación de la información de manera ética y legal se han realizado las siguientes acciones:

Consulta del fichero robots.txt:

Con el fin de verificar que no se está obteniendo información de páginas que se prohíban de manera explícita en el mismo.

https://contrataciondelestado.es/robots.txt

Consulta del apartado de Aviso legal de la propia página web

Donde se indica expresamente que:

Se autoriza la reproducción total o parcial de los contenidos del Portal de la Plataforma de Contratación del Sector Público, siempre que se cite expresamente su origen.

https://contrataciondelestado.es/wps/portal/!ut/p/b1/04_Sj9Q1MTa0NDG2sDDRj9CPykssy0xPLMnMz0vMAfGjzOKdgi0sHJ0MHQ0MjEMtDBzNAgOdLV0MjAwsjYEKIoEKDHAARwNC-sP1o8BKTI2dTcK8wgLMgj3dDQw8PdxcfEINTQ3cjcygCvBY4eeRn5uqnxuVY-mp66gIABZkOm0!/dl4/d5/L2dJQSEvUUt3QS80SmtFL1o2X0FWRVFBSTkzME8xS0MwMkJOMFNHTUgzMFEz/

7. Inspiración

La motivación a la hora de analizar este conjunto de datos viene dada por el interés general de la ciudadanía sobre como destinan los recursos los organismos públicos. El conocer los tipos de contratos que salen a licitación, los presupuestos destinados a cada uno de ellos y el ámbito de ejecución nos ha parecido muy interesante.

Como hemos indicado en el punto anterior existe una empresa que se encarga de recopilar esta información ampliada para proporcionar servicios de información a otras empresas que estén interesadas en presentarse a licitaciones por lo que este sería otro campo útil que explorar, aunque para esto habría que ampliar la información recopilada por el script para incluir la información de detalle de los expedientes que incluye diversos documentos relacionados de interés.

Una vez que se dispone de estos datos, nos permite poder realizar análisis y dar respuesta entre otros a los siguientes supuestos:

- Los tipos de necesidades que cubren los servicios públicos en España mediante contratación a empresas privadas.
- La cuantía del presupuesto que gastan diferentes servicios públicos en dicha contratación.
- El volumen de licitaciones por organismo

8. Licencia

Para realizar la elección del tipo de licencia tenemos el objetivo de mantener el único requisito que indica el propietario y que la información puede ser utilizada en cualquier ámbito donde resulte de utilidad. Con estas consideraciones en mente hemos decidido utilizar: CC BY-SA 4.0

https://creativecommons.org/licenses/by-sa/4.0/

9. Código

Esta extracción de datos ha sido realizada mediante técnicas de web scraping en lenguaje Python utilizando de manera destacable las librerías:

Selenium: Por su facilidad para realizar la navegación y con el fin de conocer esta librería.

BeautifulSoup: Para aprovechar su potencia y facilidad para parsear los objetos html

Multiprocessing: Para acelerar la descarga de información lanzando scripts de recuperación de información en paralelo.

Proceso de recuperación de la información:

Lo que hacemos es navegar hasta el buscador de licitaciones y buscamos las licitaciones que existen para un número determinado de meses de un año. Una vez obtenido el resultado, extraemos la información descrita anteriormente y vamos navegando entre las diferentes páginas de resultados obtenidos hasta que completamos la extracción de cada una de ellas. Por último, todos esos datos extraídos han sido almacenados en un fichero CSV para cada uno de los meses que abarca el periodo de tiempo

Características que han dificultado la extracción de información:

Hemos encontrado las siguientes características que dificultan la extracción de la información y que han determinado la elección de las librerías:

Carga dinámica de la información de la tabla

Al pulsar en el botón buscar o en el botón siguiente para pasar a la siguiente página selenium detectaba que la página había cargado cuando en realidad todavía no había cargado la información de la tabla por lo que hemos tenido que indicar a selenium que esperará hasta que cargará un elemento de la tabla para poder continuar con el proceso de webscraping

También hemos detectado dos características en la web que parecen diseñadas para dificultar el proceso de webscraping:

Lentitud en la carga de la información y pocos resultados por página

Cada página de resultados contiene 20 resultados y tarda en función del número de resultados totales entre 5 y 10 segundos. Esto hace que para descarga el año 2022 completo que devuelve 38569 páginas el tiempo de descarga estimado para toda la información sea de alrededor de 107 horas.

Para sortear esta dificultad hemos tomado dos medidas: preparar el script para que descargue por meses y lanzar varios procesos en paralelo descargando los meses que necesitamos, de esta manera conseguimos que el proceso estimado de descarga dure alrededor de 2 horas. (La ejecución real tardo más de 3 horas)

Para realizar la ejecución de varios procesos de webscraping hemos utilizado la librería multiprocessing, esta librería nos permite lanzar tantos procesos en paralelo como necesitamos con la única salvedad de controlar que no exceda las capacidades de procesado ni de ram de la máquina desde la que se ejecuta.

Tiempos de espera muy largos entre páginas que suceden de manera aleatoria:

Al realizar la ejecución del script completo hemos observado que de manera aleatoria saltaba un timeout exception en la función de selenium que se encarga de esperar a que cargue la tabla con la información, mientras el resto de los procesos seguían descargando la información sin experimentar ningún retardo:

```
WebDriverWait(navegador, 1000).until(ec.text_to_be_present_in_element
                ((By.CSS_SELECTOR, "[id$=textfooterInfoNumPagMAQ]"), str(contador)))
```

Al lanzar varias veces el proceso hemos observado que este fallo no es repetible ni predecible por lo que la única solución que hemos encontrado para sortearlo es lanzar por separado el proceso que fallaba para descargar la información al completo.

El código se encuentra ubicado dentro del repositorio Git https://github.com/jce1/Practica1

10. Dataset

El DOI al dataset en formato CSV en Zenodo es: https://doi.org/10.5281/zenodo.7324628

11. Vídeo

La URL de Google Drive donde se encuentra publicado el video es:
https://drive.google.com/file/d/11Cv_jpekDVo03T8U2Px-HYCsr5nR7uq6/view?usp=share_link

12. Contribuciones

Contribuciones	Firma
Investigación previa	jjcorrales jcescribano
Redacción de las respuestas	jjcorrales jcescribano
Desarrollo del código	jjcorrales jcescribano
Participación en el vídeo	jjcorrales jcescribano