PEC1: WEB SCRAPING

Tipología y ciclo de vida de los datos

Resolución de preguntas con respecto a la extracción del dataset de la práctica

Sergio Beltrán Nuez (sbeltrann@uoc.edu)

Borja López Gómez (blopezgomez@uoc.edu)

Contexto

La información recolectada que se introduce en este documento se encuentra dentro del contexto de la asignatura Tipología y Ciclo de Vida de los Datos de la UOC. En esta práctica, el objetivo es aplicar los fundamentos y conceptos aprendidos del bloque de Web Scraping.

Para ello, se ha elegido la página web del portal de transparencia de Newtral (https://transparentia.newtral.es/). En esta página web se muestra información sobre los sueldos públicos actuales de los cargos políticos de España a nivel nacional, autonómico y local.

La información de los sueldos de estos cargos políticos está accesible de forma pública a través del Boletín Oficial del Estado (BOE) y de los equivalentes boletines autonómicos. A pesar de ello, en la página web de Newtral ya están todos estos datos centralizados en un único portal, por lo que se ha considerado una página web ideal para obtener la mayor cantidad de datos con una única lógica de scraping.

Título

El título elegido para este dataset es "**Sueldos de los políticos españoles**" (Spanish politicians salaries).

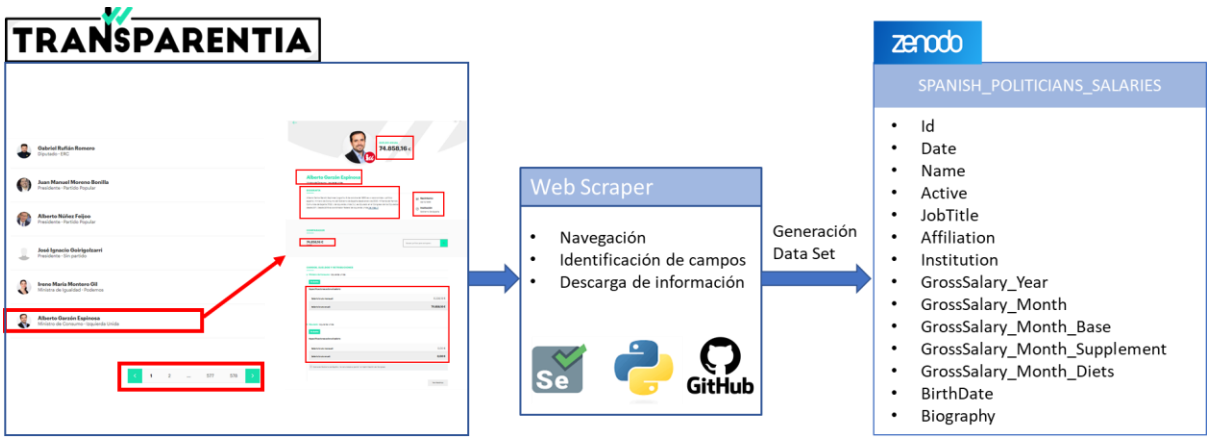
Descripción del dataset

El conjunto de datos extraído representa los sueldos actuales de los políticos de toda España. Incluye políticos tanto a nivel nacional como autonómico y local.

Para cada político (fila del dataset), se incluye información sobre el sueldo bruto anual, así como un desglose de cuánto es su sueldo base, complementos y dietas. En el apartado de "Contenido" de este documento se explican el resto de los datos que también se han scrapeado de la página web.

Como posibles líneas futuras (adicionales a esta práctica), si este proceso de scraping se ejecutara de forma recurrente (una o dos veces al año), se podría tener un histórico de cómo han ido evolucionando los sueldos de los políticos españoles.

Representación gráfica



Contenido

En esta sección se realiza una pequeña descripción de cómo se ha implementado el proceso de web scraping, y qué se ha conseguido obtener.

¿Cómo se han recogido los datos?

Inicialmente se hizo un análisis de cómo estaba hecha la página web de Newtral, y no parecía estar hecha con ningún framework Javascript que pudiese complicar el proceso de scraping. Pero al hacer el primer intento de obtención de contenido con **Requests y BeautifulSoup**, el resultado estaba vacío, sólo aparecía el contenido estático de la página web, y no los datos de los políticos que realmente se esperaban.

Estudiando más en profundidad la página web y cómo obtiene los datos (utilizando el inspector de Network de Chrome/Firefox), vimos que la lista de políticos y los datos de cada político se cargan a través de llamadas a un API REST hechas de forma dinámica con Javascript. Por lo tanto, **ha sido necesario utilizar Selenium** en el algoritmo de scraping, de forma que este contenido dinámico ya estuviera cargado.

En cuanto a los **mecanismos de protección contra scraping**, creamos un algoritmo inicial utilizando también Selenium que estuviera navegando por la web durante horas. No se observó ningún tipo de ban ni ocultación de datos, por lo que se intuye que no tienen implementado ningún firewall o protección contra un número de peticiones abusivo como puede ser un proceso de scraping.

Para la obtención de todos los datos, encontramos una URL inicial de la que partir: https://transparentia.newtral.es/busqueda-avanzada?name=&inactive=false

A partir de esta URL, aparecen más de 800 páginas de políticos tanto en activo como inactivos, teniendo 20 políticos por página. De esta forma, es posible obtener la información de cada uno de ellos, y mediante la paginación que la propia página web presenta se ha ido navegando hacia la página siguiente hasta finalizar el proceso.

En el algoritmo de scraping hemos implementado condiciones de salida para llevar a cabo pruebas y comprobar la extracción de datos sin necesidad de tener que procesar las 800 páginas disponibles. A continuación, se detallan los campos del dataset que se han extraído de la página web.

¿Qué campos que incluye el dataset?

El objetivo principal del proceso de scraping era el de obtener el salario de cada uno de los políticos en España. Durante el desarrollo del scraper, hemos encontrado más información relevante que también hemos incluido en el proceso, ya que consideramos que podría ser de relevancia para un posterior proceso de análisis o minería de datos.

Para cada político, se han recogido los siguientes campos:

- Id: un identificador único para cada político asignado por nosotros. Va desde 1 hasta el fin.
- Date: fecha en la que se extrae la información.
- Name: nombre del político.
- Active: campo booleano para saber si está en activo o no.
- JobTitle: puesto que ocupa actualmente.
- Affiliation: partido político al que está afiliado.
- Institution: institución donde lleva a cabo su labor.
- GrossSalary_Year: sueldo bruto anual.
- GrossSalary_Month: sueldo bruto mensual.

- GrossSalary_Month_Base: desglose de qué parte de su sueldo bruto mensual corresponde a sueldo base.
- GrossSalary_Month_Supplement: desglose de qué parte de su sueldo bruto mensual corresponde a suplementos.
- GrossSalary_Month_Diets: desglose de qué parte de su sueldo bruto mensual corresponde a dietas.
- BirthDate: fecha de nacimiento del político.

Agradecimientos

Los propietarios de este conjunto de datos son:

- Sergio Beltrán Nuez https://github.com/beltransrg
- Borja López Gómez https://github.com/blopez

No se ha hecho uso de ningún análisis existente anteriormente.

Con respecto a los principios éticos y morales, se ha intentado acceder a las secciones de Aviso Legal y Política de Privacidad y se han dado dos circunstancias:

1) Los hipervínculos desde la página inicial de la página web están rotos (falta una barra).
2) Al poner la barra de forma manual, llevan a una página de error 404 "político no encontrado", por lo que en el momento de hacer este scraping no hay constancia de ninguna política que lo prohíba.

Inspiración

Creemos que es un tema muy interesante, no sólo desde la perspectiva técnica del reto de scraping de datos, sino porque es un tema de interés para la población en general. Los políticos tienen unos sueldos altos, que además son pagados con dinero público que sale del bolsillo de los contribuyentes.

Además, como hemos visto durante los últimos años, la corrupción en la política es algo muy común. Por lo tanto, que existan datos públicos y transparentes sobre el sueldo de los representantes públicos es un paso adelante.

Por ello, gracias a la extracción de datasets como este, podrían implementarse procesos posteriores de análisis de datos para encontrar tendencias en partidos políticos, regiones, grupos de edad, etc. Seguro que daría mucho de lo que hablar.

Licencia

La decisión de utilizar una licencia pública, como Open Database License (ODbL) para el dataset y CC0: Public Domain Licensepara el código fuente es para poder hacer un uso público y permitir a otros usuarios la posibilidad de compartir, modificar y utilizar libremente tanto la base de datos como el código.

El **ODbL** se creó con el objetivo de permitir a los usuarios compartir sus datos libremente sin preocuparse por los problemas relacionados con los derechos de autor o la propiedad. Permite a los usuarios utilizar libremente los datos de la base de datos, incluso en otras bases de datos; editar los datos existentes en la base de datos; y añadir nuevos datos a la base de datos. La licencia establece los derechos de los usuarios de la base de datos, así como el procedimiento correcto para atribuir el mérito a los datos, y cómo realizar cambios o mejoras en los datos, simplificando así el intercambio y la comparación de datos.

Por su parte **CC0** permite publicar el código de la forma más completa posible en el dominio público, de modo que otros puedan basarse en ellas, mejorarlas y reutilizarlas libremente para cualquier fin sin restricciones en virtud de la legislación sobre derechos de autor o bases de datos.

Código

El código desarrollado para la extracción del conjunto de datos ha sido realizado en Python, y está accesible en Github a través del siguiente enlace: https://github.com/beltransrg/politiciansSalaries

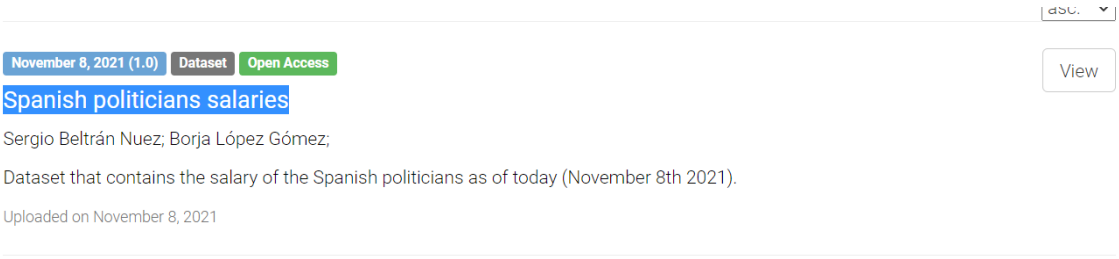
En el propio Readme del repositorio de Github se explica cómo está estructurado dicho código fuente.

Dataset

El dataset generado recoge un conjunto de atributos asociados a diferentes cargos públicos. Se compone de una clave natural que es el nombre y una clave primaria autogenerada de manera automática en base al orden en el que está incluido en la página. El dataset se compone de un total de 16941 registros y 14 atributos asociados a cada registro.

El dataset se ha publicado en el portal Zenodo bajo el nombre Spanish politicians salaries, en la siguiente url:

https://zenodo.org/search?page=1&size=20&q=5655212



Contribuciones

Contribuciones	Firma
Investigación previa	SNN, BLG
Redacción de las respuestas	SNN, BLG
Desarrollo del código	SNN, BLG