Práctica 1. Web Scraping

M2.851. Tipología y ciclo de vida de los datos

Autores

Pablo Román-Naranjo Varela

Adrián Vicente Gómez

Índice

1. Contexto

EURAXESS es un sitio web oficial de la Unión Europea que sirve como repositorio de ofertas de trabajo orientadas, mayormente, al mundo de la investigación. De esta forma, numerosas empresas e instituciones publican en él sus vacantes, permitiendo a los investigadores encontrar todas estas ofertas en un solo agregador.

En este contexto, el presente trabajo persigue obtener una herramienta capaz de extraer información de este sitio web para búsquedas específicas de puestos de trabajo y para ello hará uso de herramientas de web scraping.

Los datos extraídos pueden ser de gran utilidad para aquel/aquella científico/a de datos que quiera empezar la carrera investigadora tras la finalización del máster (ofertas predoctorales), pero también para los doctores en ciencia de datos que quieran continuar su línea de investigación (ofertas posdoctorales).

2. Título

El título escogido para el dataset es: ***Data Science job offers in Euraxess.***

3. Descripción del dataset

El dataset obtenido a partir del scraping de la web EURAXESS contiene las ofertas de trabajo relacionadas con el término *Data Science* recogidas en el sitio. Cada fila del juego de datos contiene una oferta de trabajo diferente y sus atributos.

4. Representación gráfica

La herramienta obtenida para scrapear la web de Euraxess funciona como se observa en la ilustración 1.

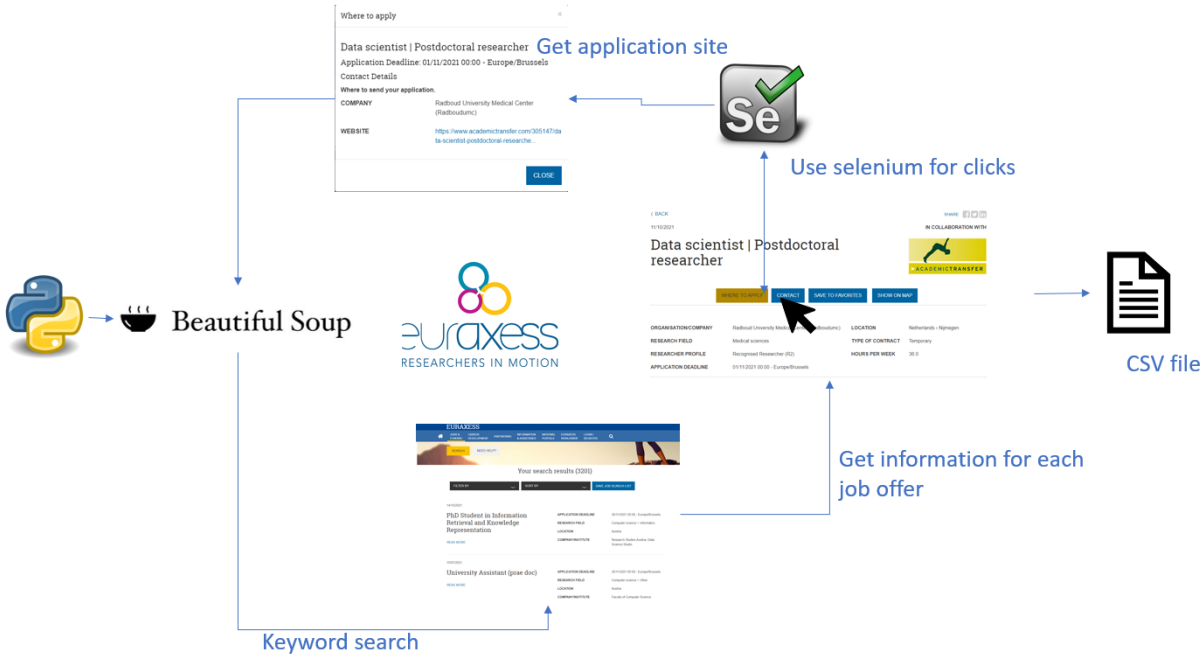

Ilustración 1. Funcionamiento de herramienta de scraping.

5. Contenido

El dataset se obtuvo el día 4 de noviembre de 2021 para los resultados obtenidos de la búsqueda del término "Data Science". Los atributos descriptores del dataset son:

- **Job Offer Title**: Título descriptivo de la oferta de trabajo. Ejemplo: Data scientist | Postdoctoral researcher.
- **Researcher Profile**: Perfil requerido para los solicitantes de la oferta, en caso de haber varios perfiles se separan por comas. Ejemplo: Recognised Researcher (R2).
- **Company**: Compañía o institución que ofrece el trabajo. Ejemplo: Radboud University Medical Center (Radboudumc).

- **Fields**. Campos de investigación en los que se engloba la oferta. Ejemplo: Physics.
- **Hours/Week**: Jornada laboral en horas (por semana). Ejemplo: 36.
- **Country**: País donde se ofrece el trabajo. Ejemplo: Netherlands.
- **City**: Ciudad donde se ofrece el trabajo. Ejemplo. Nijmegen.
- **Where to Apply**: Dirección web o email en la que solicitar el trabajo. Ejemplo: https://www.academictransfer.com/en/305147/data-scientist-postdoctoral-researcher/apply/#apply
- **More info**: Dirección web de Euraxess donde se encuentra la oferta. Ejemplo: https://euraxess.ec.europa.eu/jobs/695497

6. Agradecimientos

Los datos recogidos mediante esta herramienta de web scraping pertenecen, en su totalidad, a EURAXESS, una iniciativa europea que da soporte a las personas que quieren comenzar o continuar su carrera de investigación.

Aunque tras una búsqueda exhaustiva no hemos encontrado ningún análisis de este tipo en EURAXESS, sí hemos podido observar análisis similares con datos procedentes de otro portal de empleo, como lo es Glassdoor. Estos análisis se pueden encontrar en:

1. Repositorio *'Scraped Job Data'* de picklesueat.
2. Repositorio *'Glassdoor Jobs Data-Analysis'* de Atharva-Phatak

Tras consultar el archivo robots.txt de EURAXESS, se verificó que la ruta que íbamos a rastrear era válida según los criterios de esta web, tal y como se puede comprobar en el siguiente fragmento extraído del archivo robots.txt:

```
User-agent: *
Crawl-delay: 10
.
.
.
# Allow page parameter with keywords=
Allow: /*?keywords=&page=*
```

De igual manera, nos aseguramos que nuestro script no saturaba de peticiones el servidor web mediante espaciado automático de peticiones HTTP, usando la función *sleep()*.

7. Inspiración

No siempre es fácil encontrar oportunidades laborales si estás interesado en empezar o continuar una carrera investigadora en un campo determinado, como por ejemplo la ciencia de datos. En este sentido, tener un dataset actualizado con ofertas de trabajo sobre tu campo de interés simplificaría esta búsqueda. Este dataset pretende responder a preguntas tales como: a) en qué país hay más demanda de empleo de científico de datos o b) en qué campo de la investigación se requiere, en mayor medida, de científico de datos.

En comparación con los análisis presentados en el apartado anterior, y debido a la diferente naturaleza de los portales de empleo que se han rastreado (Glassdoor VS EURAXESS), nuestro análisis se enfoca por completo a ofertas de trabajo dirigidas a la investigación. Estas ofertas, por norma general, no se llegan a publicar en portales de empleo más generalistas, como Glassdoor, o se quedan diluidas entre otras ofertas y son difíciles de encontrar.

8. Licencia

La licencia escogida para el dataset y el código que genera el mismo es *CC BY-NC-SA 4.0 License*. Se ha escogido esta licencia ya que permite la distribución y copia del material, además de su transformación en todos los términos excepto en propósitos comerciales.

El hecho de no permitir el uso con propósitos comerciales es para no entrar en conflictos legales con los administradores de EURAXESS.

9. Código

El código utilizado para la generación del dataset puede consultarse en el siguiente repositorio de Github.

10. Dataset

El dataset generado el día 4 de noviembre de 2021 con la herramienta desarrollada para el término "*Data Science*" puede encontrarse en Zenodo (doi:10.5281/zenodo.5636238)

11. Contribuciones

Contribuciones	Firma
Investigación Previa	PRNV; AVG
Redacción de las respuestas	PRNV; AVG
Desarrollo del código	PRNV; AVG

PRNV: Pablo Román-Naranjo Varela

AVG: Adrián Vicente Gómez

12. Videopresentación

El video presentando el proyecto puede encontrarse en el siguiente enlace:

https://drive.google.com/file/d/1vUG5Zh4NLBoWAm_ZuRq8G0AjKeCMw3YU/view?usp=sharing