PRA 1 - Tipología y ciclo de vida de los datos

Maria Dolores Moyano Guerrero y Víctor Cáncer Castillo

Índice

Contexto

Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

Estos datos se han recogido para practicar el *web scraping* en la asignatura de *Tipología y ciclo de vida de los datos* del Máster de ciencia de datos de la UOC.

Cómo (futuros) científicos de datos hemos tenido la curiosidad de estudiar cómo está el mercado laboral actualmente en varias ciudades europeas y americanas. Además hemos querido averiguar en qué lugares el trabajo de científico de datos está más reconocido por las empresas y por lo tanto mejor remunerados.

Para ello hemos obtenido los sueldos que se ofrecen por diferentes empresas utilizando la web `https://glassdoor.es/`, donde los trabajadores pueden informar de su sueldo de manera anónima. Por otro lado hemos extraído datos de la web `https://datosmacro.expansion.com/` dónde hay múltiples datos económicos, entre ellos el salario medio, lo cual nos puede mostrar si el trabajo del científico de datos está mejor/peor remunerado que el resto de trabajos en ese pais o ciudad.

Figura 1: *Captura de las páginas webs utilizadas para el estudio*

Título

Definir un título que sea descriptivo para el dataset.

El objetivo de los datasets es poder hacer una comparación de los sueldos para un científico de datos en determinadas ciudades con el sueldo medio de los países donde radican dichas ciudades.

Los dos datasets principales que extraemos son: SMI.csv (Salario Medio interprofesional) y dataset_sueldos_clean.csv (Suelos de cientificos de datos tras un preproceso de datos).

Descripción del dataset

Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

Para las ciudades descritas se recogen los datos:

- *Sueldos_DataScientist_vs_SMI*: Comparación de los sueldos de las distintas ciudades con el Sueldo Medio.
- *SMI.csv*: Contiene los sueldos medios de un listado de países.
- *dataset_sueldos.csv*: Sueldos extraídos para científicos de datos en distintos periodos.
- *dataset_sueldos_clean.csv*: Sueldos de un científico de datos tratados para que todos tengan un sueldo anual.

Representación gráfica

Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

Los datos obtenidos en los dos datasets recogen, de manera resumida, los datos siguientes:

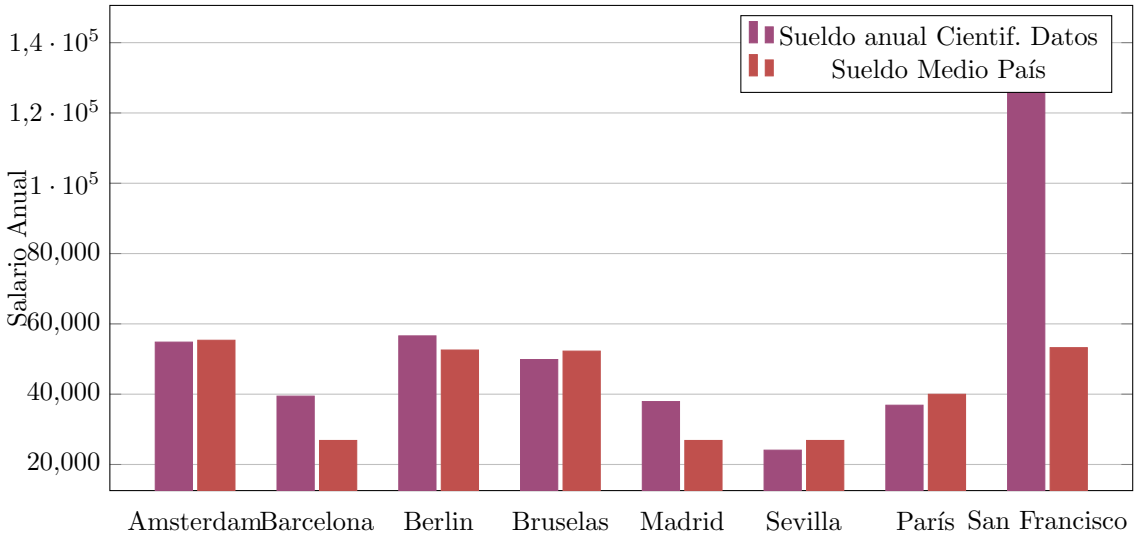

Figura 2: *Comparación de los sueldos de científico de datos en las ciudades estudiadas*

Por otro lado creamos el siguiente esquema que relaciona los datos de los datasets:

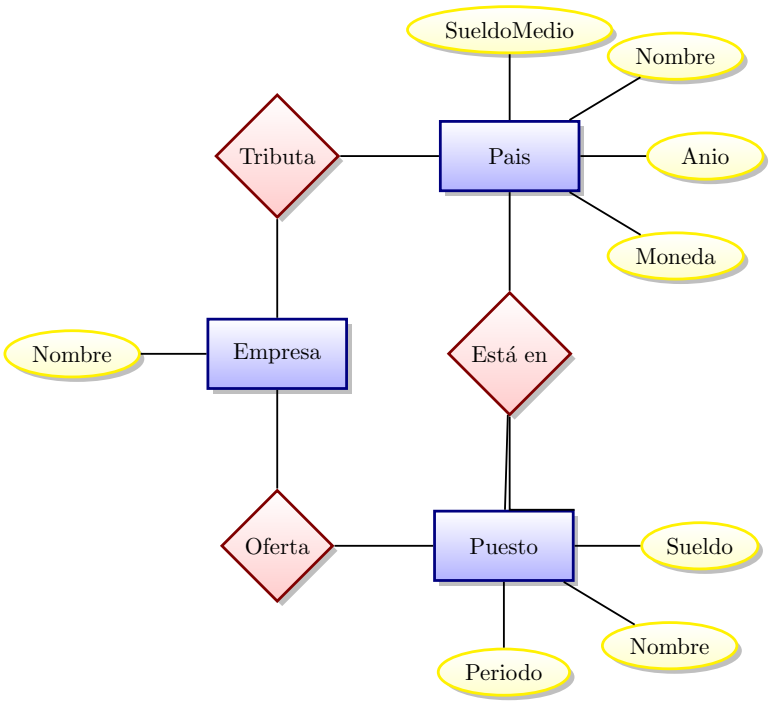

Figura 3: *Diagrama de las entidades principales del proyecto*

Contenido

Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido.

Dataset_sueldos.csv: se ha obtenido de la página *https://glassdoor.es/* el 5/4/2022. No ha sido necesario recoger los datos en varios días debido a la naturaleza del proyecto.

- **Ciudad:** Ciudad donde está radicado el empleo.
- **Empresa:** Empresa que proporciona el empleo.
- **Sueldo:** Sueldo bruto correspondiente al empleo, ciudad y empresa.
- **Periodo:** Nos ofrece información sobre si el sueldo es por hora, por mes o anual.

Dataset_sueldos_clean.csv: se ha obtenido de la página *https://glassdoor.es/* el 5/4/2022. No ha sido necesario recoger los datos en varios días debido a la naturaleza del proyecto.

- **Ciudad:** Ciudad donde está el empleo.
- **Empresa:** Empresa que proporciona el empleo.
- **Sueldo:** Sueldo bruto correspondiente al empleo, ciudad y empresa.
- **Sueldo Anual:**. Sueldo ajustado a un año (algunas empresas no lo proporcionan anualizado).

SMI.csv: se ha obtenido de la página *https://datosmacro.expansion.com/* el 5/4/2022. No ha sido necesario recoger los datos en varios días debido a la naturaleza del proyecto.

- **País:** País que se corresponde al sueldo medio.
- **Año:** Año de cálculo del sueldo medio.
- **SalMed Local:** Salario medio en moneda local.
- **Moneda:** Moneda del salario local.
- **Salmed $:** Salario medio en dólares.
- **Salmed €:** Salario medio en euros.
- **Var:** Variación del salario medio.

Sueldos_DataScientist_vs_SMI: Se han calculado a partir de los datos extraídos en las páginas descritas.

- **Ciudad:** Ciudad donde está el empleo.
- **Sueldo anual:** Sueldo anual de los empleos descargados.
- **Suelvo vs Sueldo Medio:** Ratio entre el sueldo del puesto de Científico de Datos y el sueldo medio del país.

Agradecimientos

Agradecemos a los propietarios de las páginas *https://glassdoor.es/* y *https://datosmacro.expansion.com/*, por los datos de los que nos hemos podido alimentar en este proyecto de Web Scraping.

En cuanto a los principios éticos/legales hemos intentado seguir lo que se indicaba en los ficheros *robots.txt* pero desgraciadamente eso limitaba mucho los proyectos que queríamos desarrollar, por lo que finalmente hemos tenido que acceder a parte de la web que no permitían en dicho fichero.

Inspiración

En el actual contexto, donde las comunicaciones permiten buscar trabajo en distintos destinos sin tener que desplazarse, y además es posible trabajar en remoto, los sitios webs que permiten buscar trabajo y comparar sueldos son una herramienta muy útil a la hora de escoger el puesto más conveniente. Los datos recogidos son interesantes porque permiten comparar los sueldos de los científicos de datos en caliente, utilizando el factor de corrección del sueldo medio del país, para poder analizar si efectivamente esos puestos de trabajo son valorados en la ciudad donde se demandan y así tener un mapa claro de la situación, para prestar ayuda a los demandantes de empleo de esta disciplina.

Este proyecto podría ampliarse a la búsqueda de trabajo para otra tipología de puestos, simplemente, modificando la búsqueda.

Licencia

Se ha escogido la licencia **CC BY-SA 4.0** ya que se ajusta a los requerimientos que nosotros creemos convenientes. Las clausulas que debemos cumplir para ello se podrían resumir en:

- Se debe dar crédito a los propietarios originales indicando los cambios realizados. De esta manera la autoría de dichos datos queda respetada por nuestra parte.
- Permitimos el uso de los datos generados y cualquier uso de éstos debe trabajar bajo esta licencia, permitiendo que su uso sea lo más amplio posible.
- Se permite su uso comercial, permitiendo que estos datos se pudieran utilizar por parte de empresas.

Código

Para hacer tanto scraping como pre-procesado de datos hemos utilizado Python. El código está disponible en GitHub: `https://github.com/mdmoyano/web_scraping/src`.

Dataset

El Dataset está publicado en Zenodo, disponible en `https://doi.org/10.5281/zenodo.6408001`

Vídeo

Cada uno de los integrantes del grupo hará un vídeo cuyo enlace será enviado al profesor.

Contribuciones	**Firma**
Investigación previa	VCC, MMG
Redacción de las respuestas	VCC, MMG
Desarrollo del código	VCC, MMG

Índice de figuras