Tipología y Ciclo de Vida de los Datos

Web Scraping

Rosa María Miranda Castro
Máster Universitario en Ciencia de Datos

Práctica 1 (25% nota final)

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes por un proyecto analítico y usar las herramientas de extracción de datos. Para hacer esta práctica tendréis que trabajar en grupos 2 personas. Tendréis que entregar un solo fichero con el enlace Github (https://github.com) donde haya las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos de vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Podéis mirar estos ejemplos como guía:

- Ejemplo: https://github.com/rafoelhonrado/foodPriceScraper
- Ejemplo complejo: https://github.com/tteguayco/Web-scraping

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para resolverlo.
- Capacidad para aplicar las técnicas específicas de web scraping.

Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- Saber identificar los datos relevantes que su tratamiento aporta valor a una empresa y la identificación de nuevos proyectos analíticos.
- Saber identificar los datos relevantes para llevar a cabo un proyecto analítico.
- Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos o repositorios) y mediante diferentes mecanismos (tales como queries, API y scraping).
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar la capacidad de búsqueda, gestión y uso de informacióny recursos en el ámbito de la ciencia de datos.

Descripción de la Práctica a realizar

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

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

2. Definir un título para el dataset. Elegir un título que sea descriptivo.
El nombre del dataset es RankingEmpresas

3. 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).
El dataset RankingEmpresas, contiene el ranking de las Compañías que conserva la posición del ranking general pero que se ordena en base al tamaño de la Compañía (Activos), como lo define el Código Orgánico de la Producción, Comercio e Inversiones:

4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido

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

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

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

- ○Released Under CC0: Public Domain License
- ○Released Under CC BY-NC-SA 4.0 License
- ○Released Under CC BY-SA 4.0 License
- ○Database released under Open Database License, individual contents under Database Contents License
- ○Other (specified above)
- ○Unknown License

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

Recursos

- Los siguientes recursos son de utilidad para la realización de la PEC:
- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. El lenguaje Python. Editorial UOC.

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

- Los apartados 1, 2, 3 y 4 valen 0,25 puntos cada uno.
- Los apartados 5 y 8 valen 1 punto cada uno.
- Los apartados 6 y 7 valen 1,5 puntos cadauno.
- Los apartados 9 y 10 valen 2 puntos cada uno.Otros criterios que se tomarán en cuenta para la evaluación son:
- Idoneidad de las respuestas (deberán ser claras y completas).
- Complejidad del sitio web elegido para la extracción.
- Síntesis y claridad, através del uso de comentarios, del código resultante.
- Presentación adecuada de los datos.
- Organización y claridad de los documentos de entrega final.
- Completitud de los documentos requeridos para la entrega final.
- Seguimiento de recomendaciones para el buen uso del web scraping.

Criterios de valoración

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

   La información que se puede obtener es de las principales compañías las mismas que son obligadas a presentar sus estados financieros una vez termine el período fiscal, el ente de control en quien recae dicha responsabilidad para Ecuador es La Superintendencia de Compañías, Valores y Seguros.

   Con esa información se puede generar análisis para ofrecer productos y servicios financieros especializados a estas compañías, así mismo entender

por provincia la concentración de dichas compañías y también analizar clientes potenciales en base al número estimado dependiendo del tamaño de la compañía, cuyos rangos se encuentran a continuación:

- Microempresas: Entre 1 a 9 trabajadores ó Ingresos menores a $100.000,00
- Pequeña empresa: Entre 10 a 49 trabajadores ó Ingresos entre $100.001,00 y $1'000.000,00
- Mediana empresa: Entre 50 a 199 trabajadores ó Ingresos entre $1'000.001,00 y $5'000.000,00
- Empresa grande: Más de 200 trabajadores ó Ingresos superiores a los $5'000.001,00

El link para obtener esta valiosa información es: https://appscvs.supercias.gob.ec/rankingCias/principal.zul

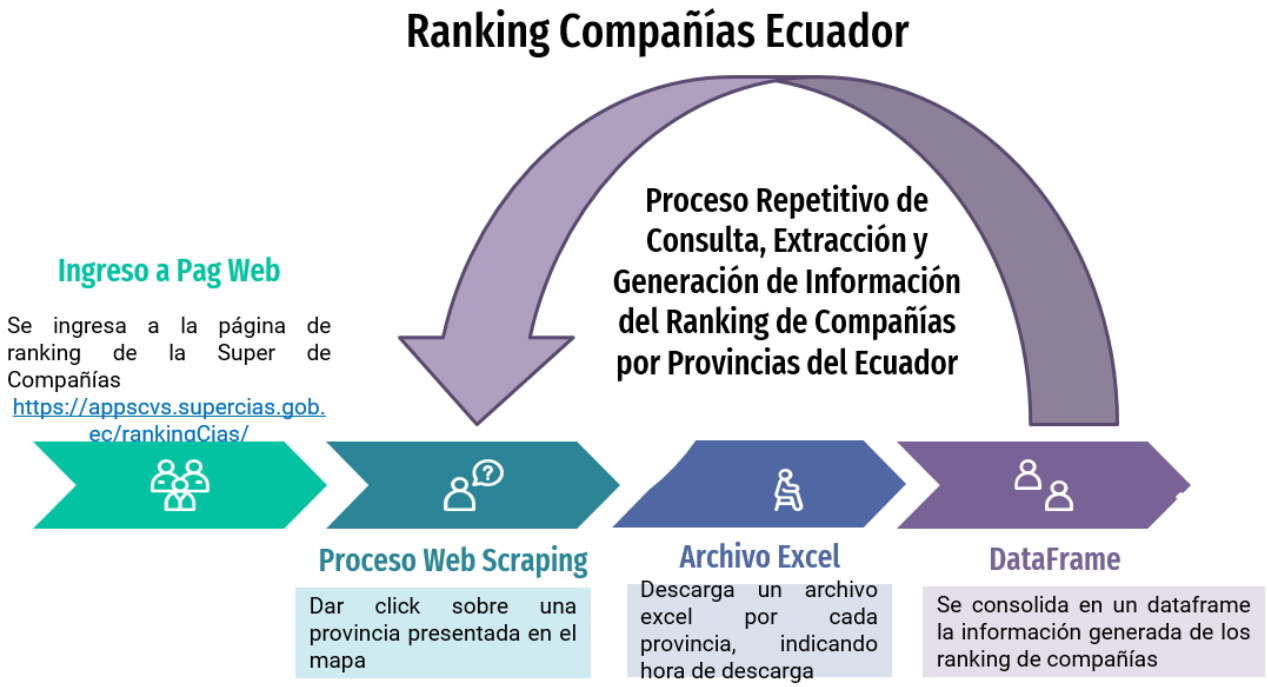
2. **Definir un título para el dataset.** Elegir un título que sea descriptivo.

El nombre del dataset es **RankingEmpresas,**

3. **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).

El dataset RankingEmpresas, contiene el ranking de las Compañías que conserva la posición del ranking general pero que se ordena en base al tamaño de la Compañía (Activos), como lo define el Código Orgánico de la Producción, Comercio e Inversiones:
En este dataset contiene 16 columnas de datos y 60 registros de 3 provincias seleccionadas, las mismas que podrían llegar a extraer de las 24 provincias que tiene Ecuador, se evidencia los siguientes campos:

4. **Representación gráfica.** Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido



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

- **Ranking_2020:** Es un campo de formato numérico, el mismo que indica el ranking que la compañía tenía en el año 2020.
- **Ranking_2019:** Es un campo de formato numérico, el mismo que indica el ranking que la compañía tenía en el año 2019.
- **Posicion_Provincia:** Es un campo de formato numérico, el mismo que indica el ranking que la compañía tiene durante el año 2021.
- **Actividad:** Es un campo de formato carácter, el mismo que viene con valores en blanco.

- **Nombre_Empresa:** Es un campo de formato carácter, el mismo que indica el nombre de la Compañía rankeada.
- **Provincia:** Es un campo de formato carácter, el mismo que indica la provincia en donde la Compañía está registrada.
- **Ciudad:** Es un campo de formato carácter, el mismo que indica la ciudad en donde la Compañía está registrada.
- **Tamanio:** Es un campo de formato carácter, el mismo que indica el tamaño de la Compañía, (Microempresa, Pequeña, Mediana y Empresa Grande).
- **Activos:** Es un campo de formato numérico, el mismo que indica el valor de los activos que tiene la Compañía al cierre del período 2021.
- **Patrimonio:** Es un campo de formato numérico, el mismo que indica el valor del patrimonio que tiene la Compañía al cierre del período 2021.
- **Venta:** Es un campo de formato numérico, el mismo que indica el valor de las ventas que tiene la Compañía ha realizado al cierre del período 2021.
- **Ut_Ant_Imp:** Es un campo de formato numérico, el mismo que indica el valor de la utilidad antes de impuestos que tiene la Compañía al cierre del período 2021.
- **Ut_Ejercicio:** Es un campo de formato numérico, el mismo que indica el valor de la utilidad del ejercicio que tiene la Compañía al cierre del período 2021.
- **Ut_Neta:** Es un campo de formato numérico, el mismo que indica el valor de la utilidad neta que tiene la Compañía al cierre del período 2021.
- **Imp_Rent_Causado:** Es un campo de formato numérico, el mismo que indica el valor impuesto a la renta causado de la Compañía al cierre del período 2021.
- **Ingreso_Total:** Es un campo de formato numérico, el mismo que indica el valor del ingreso que tiene la Compañía al cierre del período 2021.

5. **Agradecimientos.** Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

El propietario de este conjunto de datos es la entidad de control conocida como La Superintendencia de Compañías, Valores y Seguros es el organismo técnico, con autonomía administrativa y económica, que vigila y controla la organización, actividades, funcionamiento, disolución y liquidación de las compañías y otras entidades en las circunstancias y condiciones establecidas por la Ley.

6. **Inspiración.** Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Bueno este conjunto de datos tiene dos inspiraciones la primera es la rica explotación de datos para análisis y la otra debido a la estructura de ingreso y descarga de información permitió aprovechar los conocimientos de Web Scraping.

Este conjunto de datos es muy interesante ya que la misma va a generar gran valor para poder segmentar nuevamente a las empresas, ofrecer productos especializados, generar ofertas de valor y generar alianzas estratégicas.

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

- ○Database released under Open Database License

La Licencia Abierta de Bases de Datos fue creada con el fin de permitir a los usuarios compartir sus datos con libertad y sin temor a los derechos de autor o cuestiones de propiedad. Este sistema permite a los usuarios hacer uso libre de los datos contenidos en el repositorio sin temor a la infracción de derechos de autor, y a partir de los datos que han recogido añadirlas a las bases de datos, calculadas o elaboradas por ellos mismos. La licencia establece los derechos de los usuarios que poseen cuando hacen uso de los datos contenidos dentro de la base de datos, así como el procedimiento correcto para la atribución de crédito para la base de datos, y para la presentación de modificación o alteración de datos. Con esto, se consigue comparar y compartir información y

datos de forma más fácil. Los usuarios ya no necesitan vivir con el temor de las repercusiones de la infracción de derechos de autor o información robada cuando se trabaja bajo la Licencia Abierta de Bases de Datos.
Fuente: https://es.wikipedia.org/wiki/Licencia_Abierta_de_Bases_de_Datos

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El código que se generó el dataset **RankingEmpresas,** se lo desarrollo en Phyton, y se encuentra alojado en la cuenta de github.com

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

10.5281/zenodo.4679917
https://zenodo.org/search?page=1&size=20&q=4679917

Contribuciones	Firma
Investigación previa	Rossy Miranda
Redacción de las respuestas	Rossy Miranda
Desarrollo código	Rossy Miranda