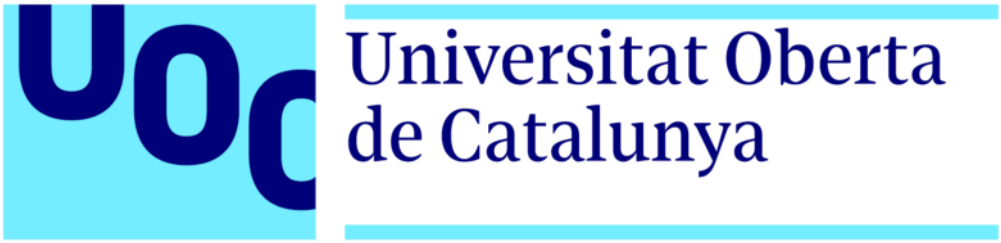

Máster universitario de Ciencia de Datos

Prueba de Evaluación Continua 1 – PEC1

Trabajo sobre los conceptos generales del ciclo de vida de los datos y *Web Scraping*

Autor:

Mario Ubierna San Mamés

Índice de Contenido

Índice de Contenido 3
Índice de ilustraciones 4
1. Introducción 5
1.1. Presentación 5
1.2. Objetivos 5
2. Enunciado 6
2.1. Ejercicio 1 6
2.2. Ejercicio 2 14
3. Bibliografía 16

Índice de ilustraciones

Ilustración 1 – Visualización panorama general. 8
Ilustración 2 - Visualización acercamiento. 9
Ilustración 3 – Filtrado. 9
Ilustración 4 - Visualización filtrado. 10
Ilustración 5 - Visualización relaciones. 10
Ilustración 6 - Visualización historial. 11
Ilustración 7 - Visualización detalles a petición. 12
Ilustración 8 – Extracción. 13
Ilustración 9 - Visualización extracción. 14

1. Introducción

1.1. Presentación

En esta Prueba de Evaluación Continuada se trabajan los conceptos generales de cuál es el ciclo de vida de los datos, y se identifican y conocen sus características. También se trabajan los conceptos esenciales de *Web Scraping*.

1.2. Objetivos

Los objetivos concretos de esta Prueba de Evaluación Continua son:

- Conocer el ciclo de vida de los datos y los principales tipos de datos.
- 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.
- Desarrollar las habilidades de aprendizaje que permitan continuar estudiando de una manera que tendrá que ser en gran medida autodirigida o autónoma.
- Desarrollar la capacidad de busca, gestión y uso de información y recursos en el ámbito de la ciencia de datos.
- Entender la utilidad, la legalidad y algunas características de *Web Scraping*.

2. Enunciado

2.1. Ejercicio 1

Después de leer el recurso "Calvo, M., Pérez, D., Subirats, L. (2019). Introducción al ciclo de vida de los datos." Contesta a las siguientes preguntas con tus propias palabras:

1. *¿Qué perfil profesional relacionado con la ciencia de datos te gustaría ser? ¿Y cuál te gustaría menos ser? Razona tu respuesta (máximo 200 palabras).*

   Dada a mi poca experiencia profesional no sé si lo que me gusta hoy va a ser lo mismo que lo que me gustará mañana, pero a día de hoy el rol que más encaja conmigo es el de *data scientist*. Soy una persona inquita de mente y por lo que he leído tanto en los recursos de esta asignatura como en otras, el *data scientist* es un filósofo (curioso) [1], pero con datos, por lo que trata de buscar respuestas a preguntas que puedan solucionar o mejorar el futuro a medio/largo plazo, además hace uso de las tecnologías, matemáticas, lógica de negocio y comunicación, las cuales son las áreas del conocimiento que más me gustan.

   Por otro lado, lo que sí tengo claro es que no quiero el día de mañana hacer tareas que sean más mecánicas, es decir, que no se haga uso de esa curiosidad o querer saber, por lo que roles como por ejemplo: arquitecto de datos, administrador de base de datos, entre otros no son los roles a los que me quiero dedicar.

2. *Lista los diferentes factores que influyen en la calidad de los datos y pon un ejemplo diferente al que se explica en los materiales (máximo 300 palabras).*

   Los diferentes factores que influyen en la calidad de los datos son [1]:

   - Exactitud: un ejemplo sería cuando tratamos de representar números decimales, es decir, en España por ejemplo usamos la coma para separar la parte entera de la parte decimal, y en el sistema inglés se usa el punto. Por lo que, si tenemos por ejemplo la variable peso medida en Kg dicho valor si viene como en el sistema inglés con un

punto y nos encontramos en España, ambos formatos serían válidos pero no exactos ya en este caso tendría que representarse con una coma.

- Completitud: siguiendo con el ejemplo anterior sería el porcentaje de pesos que sí que tienen valor, es decir, que tienen datos sin valores en blanco.
- Consistencia: por ejemplo, a la hora de sacarse el carnet de conducir una persona es apta si y solo si ha pasado el test psicotécnico, ha aprobado el examen teórico y el examen práctico. Por lo tanto, hay que comprobar que estos tres campos sean consistentes, ya que de lo contrario la persona no sería apta.
- Atemporalidad: siguiendo con el ejemplo anterior, podríamos tener un atributo temporal que indicase cuánto ha tardado una persona en sacarse el carnet desde que hizo el examen psicotécnico hasta que aprobó el examen práctico.
- Unicidad: se correspondería con el porcentaje de valores duplicados respecto a todos los datos que tenemos en tráfico sobre las personas que quieren sacarse el carnet de conducir.
- Validez: por ejemplo, en el test psicotécnico, en el examen teórico y en el examen práctico solo es válido una cadena que indique "A" si es apto o "NA" si es no apto, de lo contrario el resto de los valores no estarían permitidos.

3. *¿En qué tipos de bases de datos no es necesario conocer a priori los datos que se quieren almacenar? Pon tres ejemplos de tecnologías que utilicen estas bases de datos (máximo 100 palabras).*

   Las bases de datos que a priori no necesitan conocer los datos que se van a almacenar son las bases de datos no relacionales. Las bases de datos no relacionales más utilizadas son *Cassandra*, *Redis* y *MongoDB* [2].

   - *Cassandra*: es una base de datos creada por *Apache* del tipo clave-valor (cada elemento está definido por una clave única, dentro del mundo de la programación se puede corresponder a la estructura de datos de un diccionario, y su información se suele guardar como un objeto binario).
   - *Redis*: esta base de datos está apoyada por *VMWare* y también es del tipo clave-valor.
   - *MongoDB*: creada por *10gen* y ésta es del tipo orientada a documentos (se caracteriza porque almacena la información como un documento JSON o XML, y cada registro también está identificado con una clave única).

4. *Pon ejemplos visuales con imágenes de las 7 visualizaciones que permiten un nivel más alto de abstracción (adjuntar las 7 imágenes)*

   Antes que nada, cabe mencionar que este punto le debo dar las gracias a mi compañero de universidad Jorge Navarro González, el cual desarrolló como Trabajo de Fin de Grado *NetExtractor* [3].

   Las visualizaciones que se van a ver a continuación vienen dadas a partir de los diálogos de la película *Joker*, en el que hay un vínculo entre dos personajes si salen en una misma escena, para que quedara más legible los ejemplos he seleccionado que los personajes al menos tienen que aparecer 5 veces o más en toda la película.

   - Panorama general (*overview*):

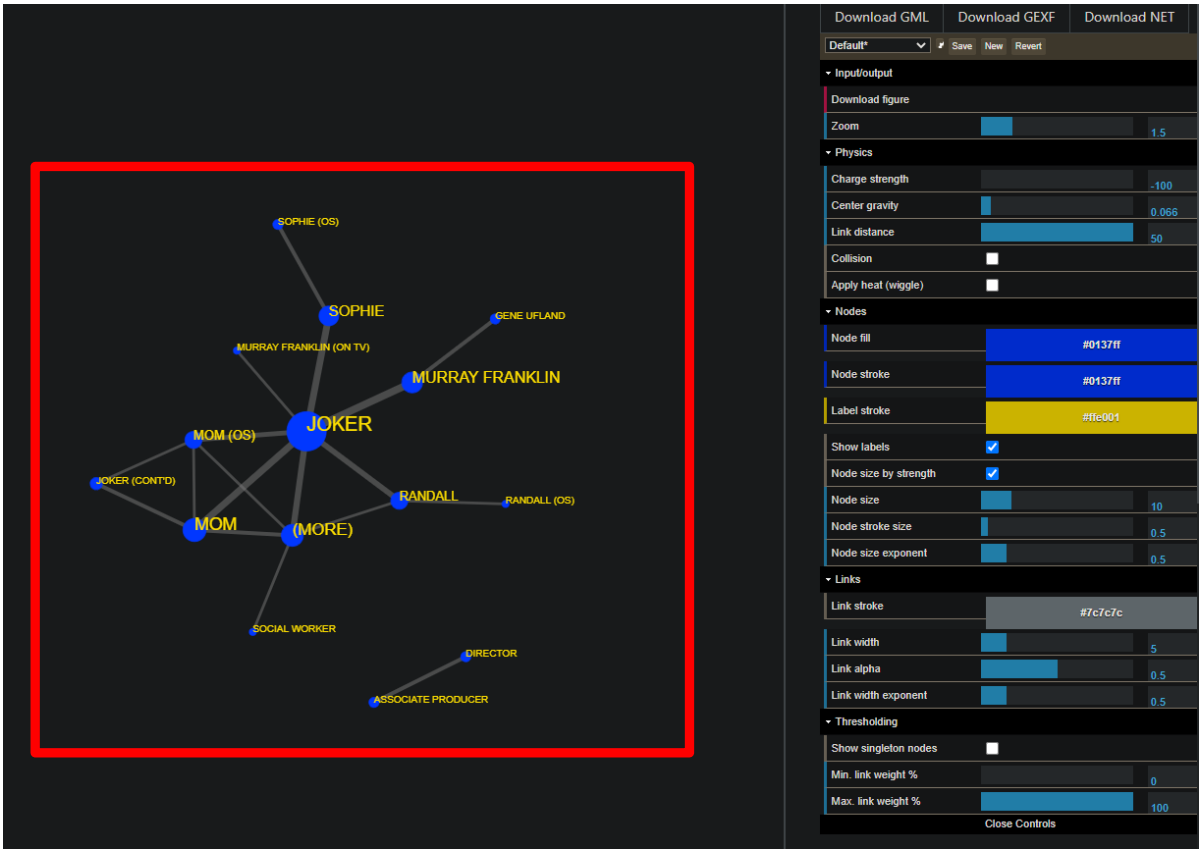

Ilustración 1 – Visualización panorama general.

   - Acercamiento (*zoom*): en la siguiente captura podemos ver el cómo hacer el zoom, para ello nos tenemos que fijar en el cuadrado rojo.

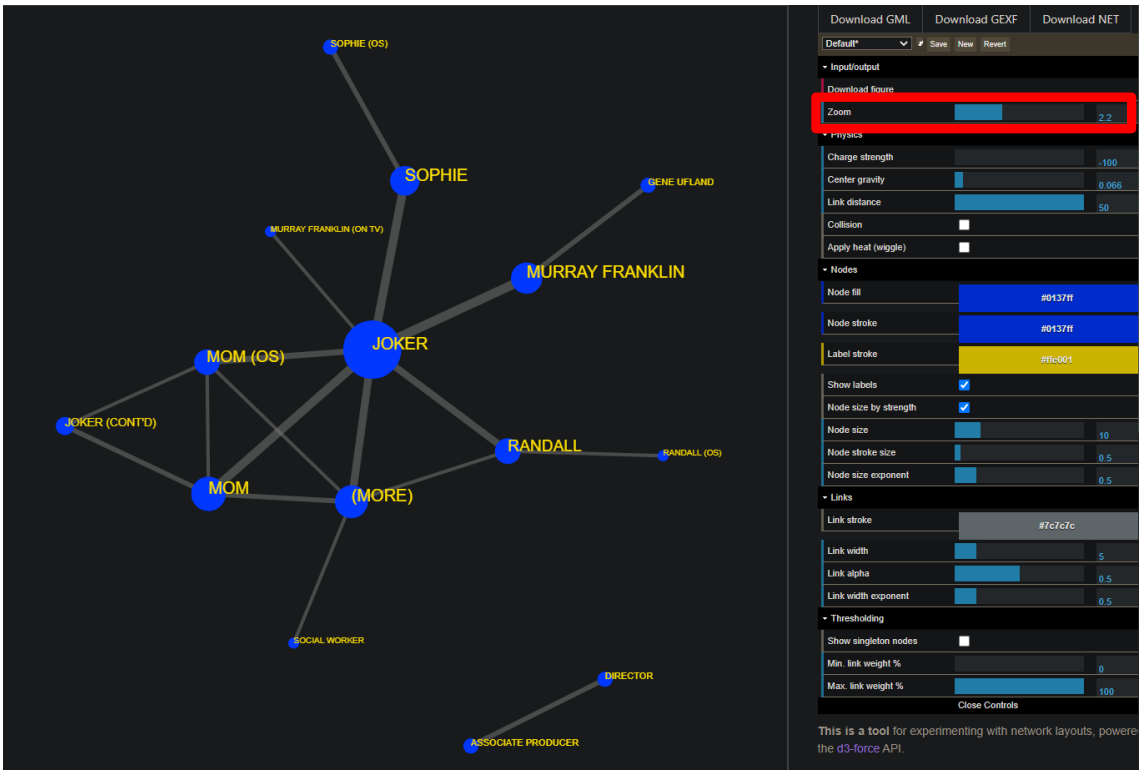

Ilustración 2 - Visualización acercamiento.

- Filtrado (*filter*): justo antes de ver el gráfico final podemos filtrar en este caso por el número de apariciones de un personaje, una vez realizado eso podemos ver la red final con dicho filtro.

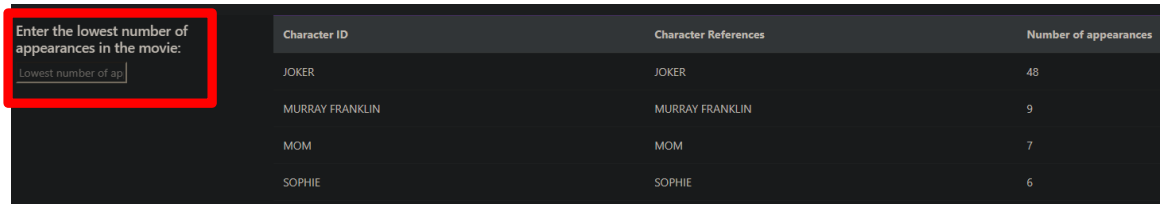

Ilustración 3 – Filtrado.

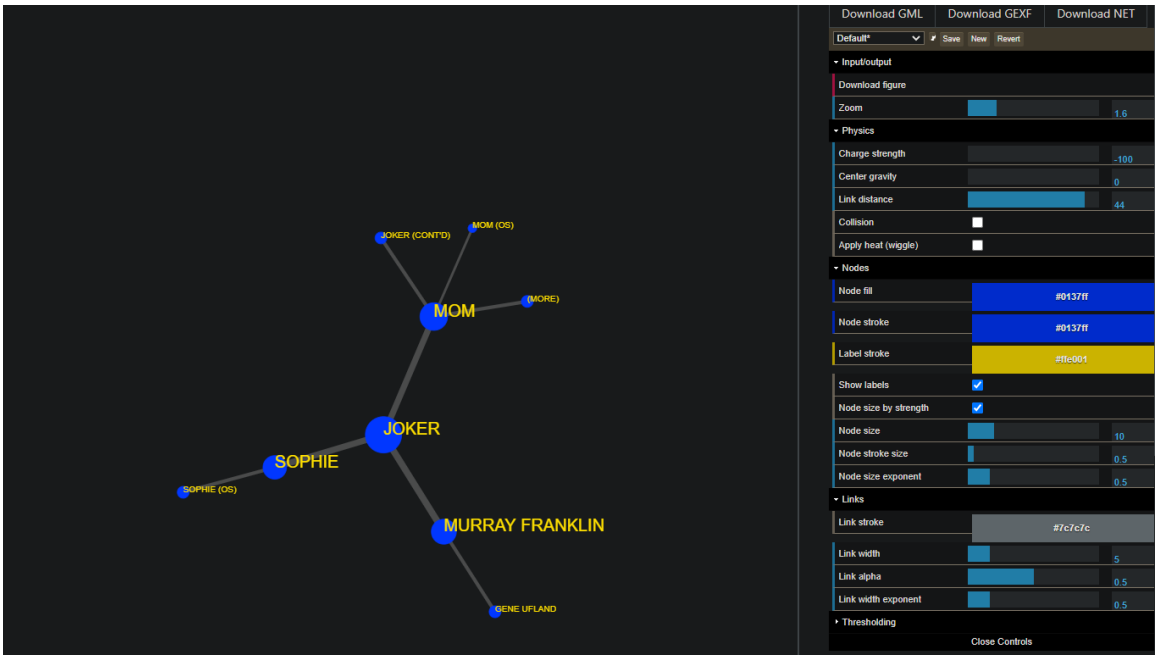

Ilustración 4 - Visualización filtrado.

- Relaciones (*relate*): como podemos apreciar en la red, se establece una relación entre dos nodos siempre y cuando aparezcan en la misma escena. Este es uno de los significados que entiendo por relación según la teoría, el otro es por ejemplo el ver si dos variables están correlacionadas, decidí este significado ya que es más visual.

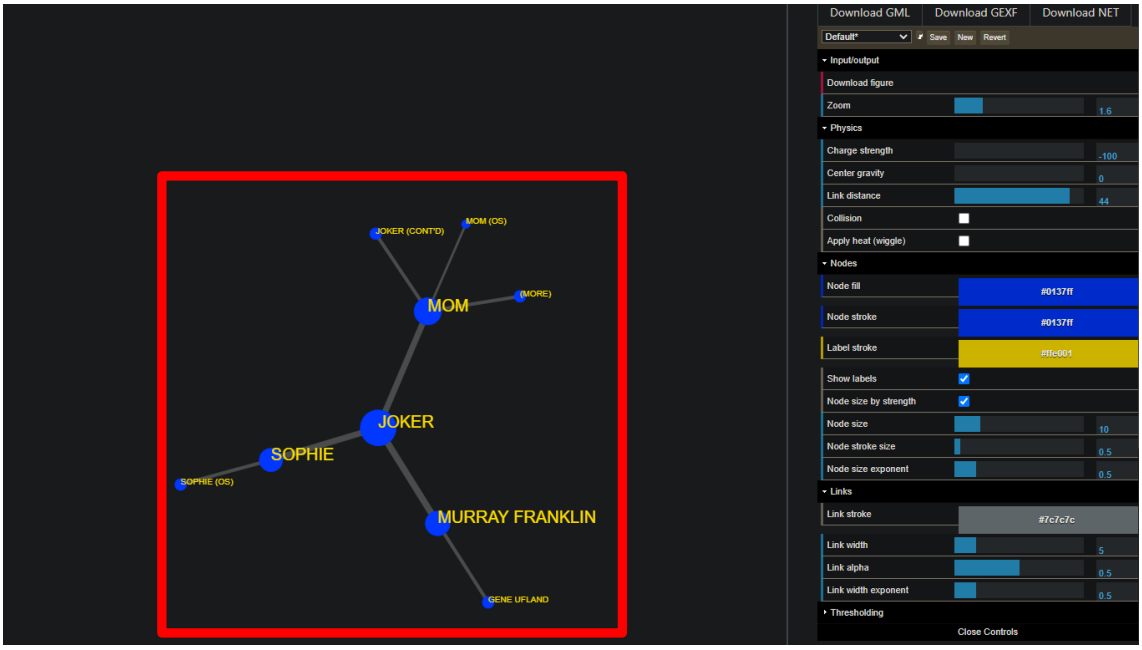

Ilustración 5 - Visualización relaciones.

- Historial (*history*): en la parte superior derecha podemos revertir los cambios.

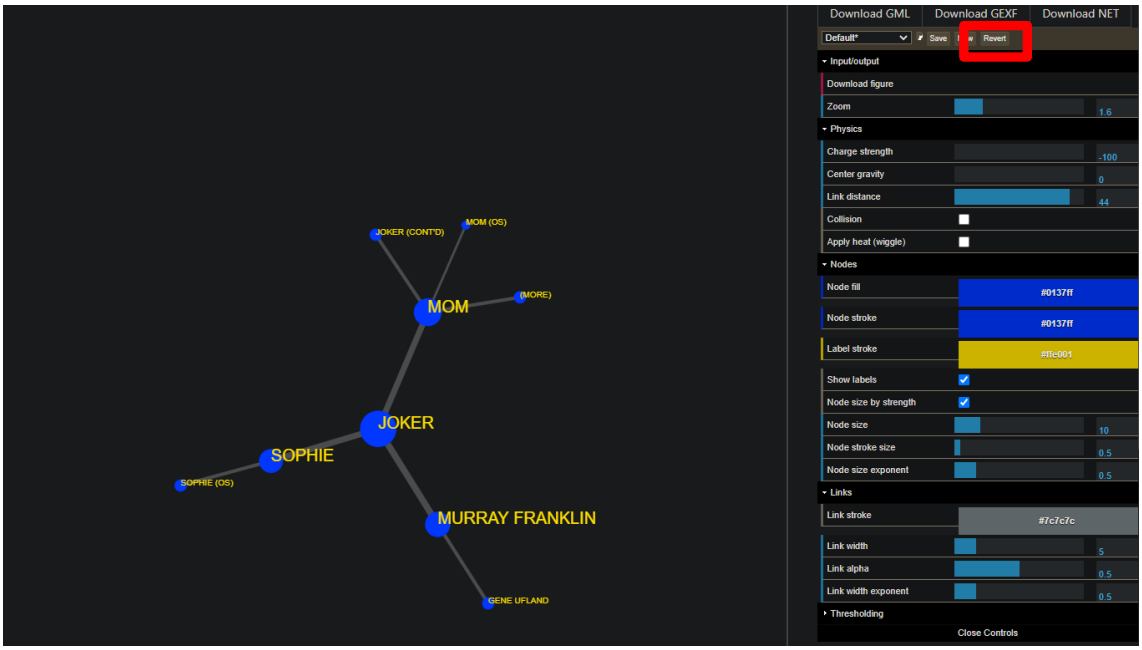

Ilustración 6 - Visualización historial.

- Detalles a petición (*details on demand*): tanto este punto como el siguiente, no me sirvieron la página de mi compañero por lo que proporciono otras imágenes. Respecto a los detalles a petición usé un gráfico de Paula López Casado, el cual fue mostrado en un charla sobre una visualización de datos [4]. Como podemos apreciar a pasar el cursor sobre cada una de las barras podemos obtener información del artista.

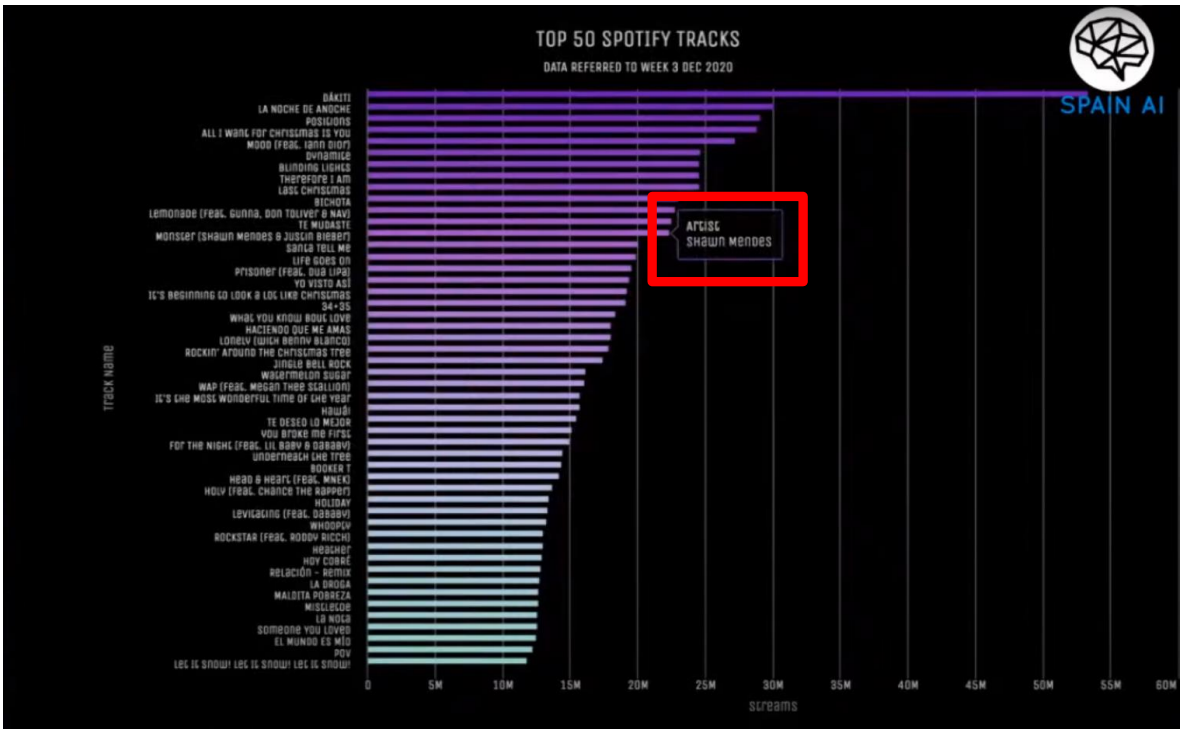

Ilustración 7 - Visualización detalles a petición.

- Extracción (*extract*): para este apartado he usado el resultado de las elecciones de Reino Unido en 2017 [5], al hacer click sobre el mapa interactivo en un condado nos da la opción de ver más información, en la cual se ve de forma detallada todos los aspectos a tener en cuenta en las elecciones. A la hora de guarda esta información detallada podemos hacer uso del *copy-paste* o de *Web Scraping* ya que dicha página no da la opción de guardar directamente los detalles.

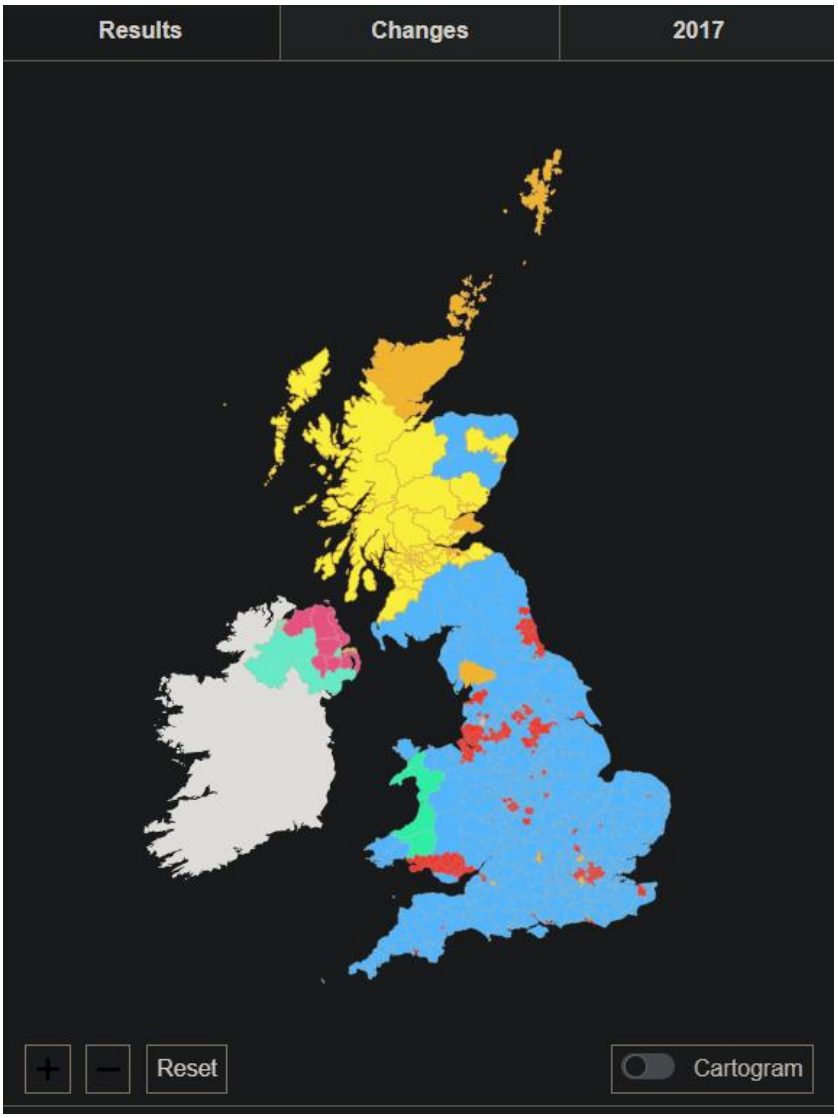

Ilustración 8 – Extracción.

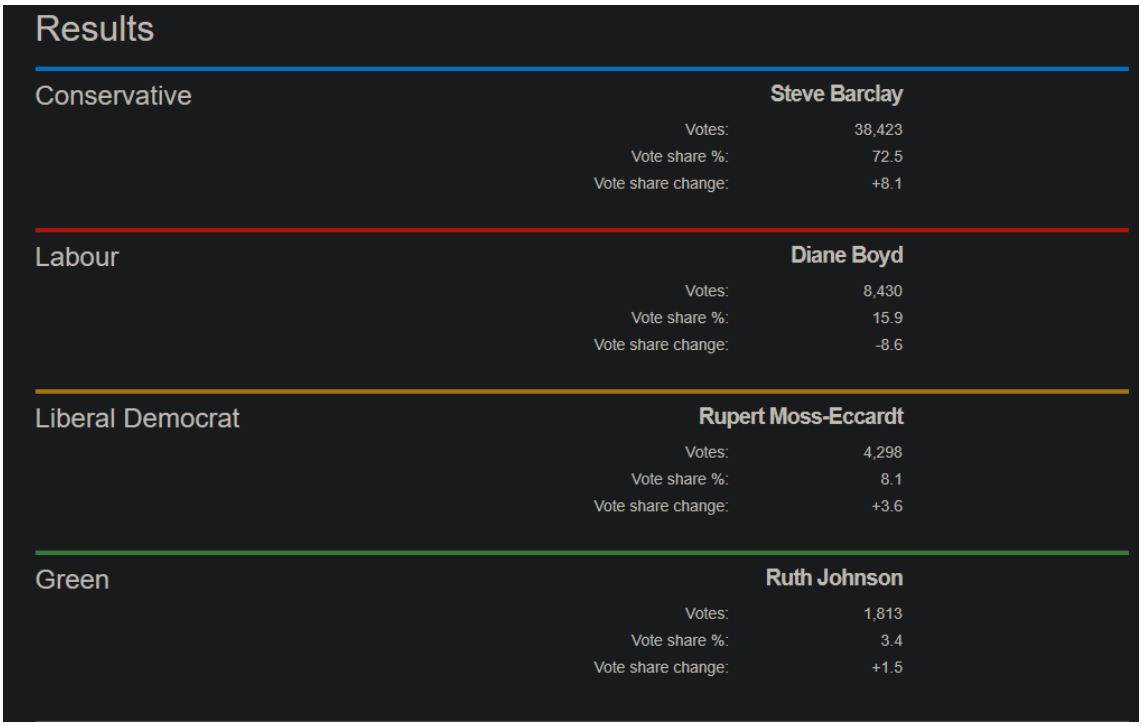

Ilustración 9 - Visualización extracción.

2.2. Ejercicio 2

Después de leer el recurso "Subirats, L., Calvo, M. (2019). Web Scraping.", capítulos 1 y 7. Contesta las siguientes preguntas con tus propias palabras:

1. *¿Por qué crees que es necesario hacer web scraping? (máximo 100 palabras).*

   No considero que sea necesario hacer *Web Scraping*, básicamente porque podemos obtener los datos de un repositorio digital o de un API (*Application Programming Interfaces*), esto es lo ideal siempre y cuando obtengamos los datos tal cual los queremos.

   En caso contrario, sí que veo el *Web Scraping* como la mejor herramienta para, o bien obtener datos, o bien complementar/completar los datos obtenidos. Hay veces que no tenemos acceso a un API, y recopilar la información de forma manual no suele ser la mejor opción, por lo que la única solución que nos queda es hacer *Web Scraping*.

2. *¿Por qué es importante analizar el contenido del archivo robots.txt? ¿Qué riesgo corremos si no lo hacemos? (máximo 100 palabras).*

Es importante analizar dicho fichero, ya que en él se encuentra toda la información relativa a las restricciones que tenemos al hacer uso del *Web Scraping* en una determinada web, como por ejemplo: exclusión de directorios, de determinadas páginas…

Por otro lado, el principal riesgo que corremos si no hacemos caso al contenido de este fichero es que nos bloqueen el acceso a la página que estamos intentado acceder o si el caso va más allá nos pueden denunciar.

3. *Explica como evitarías saturar el servidor con peticiones web (máximo 100 palabras).*

   Para evitar saturar el servidor con peticiones web lo que haría sería tratar de humanizar el *script*, de tal forma que añadiría un *delay* a cada una de las peticiones que se realicen.

   Otra forma de evitar saturar el servidor sería con las mima idea que en el párrafo anterior, pero en este caso optimizarla, es decir, que el *delay* que vamos a aplicar en nuestro *script* se calcule, de tal forma que si tarda el servidor mucho en responder el *delay* sea mayor, por el contrario si tarda menos entonces recalculamos ese *delay* para que sea menor.

3. Bibliografía

[1] «PID_00265705.pdf». Accedido: feb. 25, 2021. [En línea]. Disponible en: http://materials.cv.uoc.edu/daisy/Materials/PID_00265705/pdf/PID_00265705.pdf.

[2] «bbdd-nosql-wp-acens.pdf». Accedido: mar. 01, 2021. [En línea]. Disponible en: https://www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf.

[3] «NetExtractor». http://netextractor.herokuapp.com/ (accedido feb. 25, 2021).

[4] *Webinar (AI Tech Talk): Las 10 claves para una buena visualización de datos. Con Paula L. Casado*. 2021.

[5] «Election results 2019: Analysis in maps and charts», *BBC News*, dic. 13, 2019.