Jorge Sainero Valle
Álvaro López Cabello

PRÁCTICA 1: Web scraping

Tipología y ciclo de vida de los datos

Jorge Sainero Valle

Álvaro López Cabello

Jorge Sainero Valle
Álvaro López Cabello

1. Contexto

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

Como toda familia, las profesiones tienen una historia y un origen. Sus árboles genealógicos ayudan a entender quiénes son nuestros antecesores y de quiénes descendemos. Reflejan el parentesco intelectual entre los cualificados en un área educativa y laboral. Y tener referencias de profesionales con diferentes logros nos puede ser útil para direccionar nuestras ideas.

En el *Mathematics Genealogy Project* de la Universidad de Dakota del Norte y creado por Harry Coonce podemos adentrarnos en la genealogía de las matemáticas. Tratan de recopilar información de todos los matemáticos doctorados en el mundo, utilizando la palabra "matemático" en el sentido más inclusivo, aceptando campos alineados como la estadística o las ciencias de la computación.

Se ha escogido el sitio web [1] porque proporciona información de los matemáticos, que proviene de diferentes fuentes, y permite conocer qué doctores publicaron su PhD sobre un área o un país en concreto, sus tutores y, sus estudiantes si tienen. Con la labor de los encargados es más fácil investigar los trabajos publicados según los elementos que interesen.

2. Título

Definir un título que sea descriptivo para el dataset.

El título descriptivo del dataset es: Información relevante de los doctores matemáticos y su tesis sobre estadística en España.

El archivo generado JSON tiene por nombre "statisticians_PhD_spain".

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.

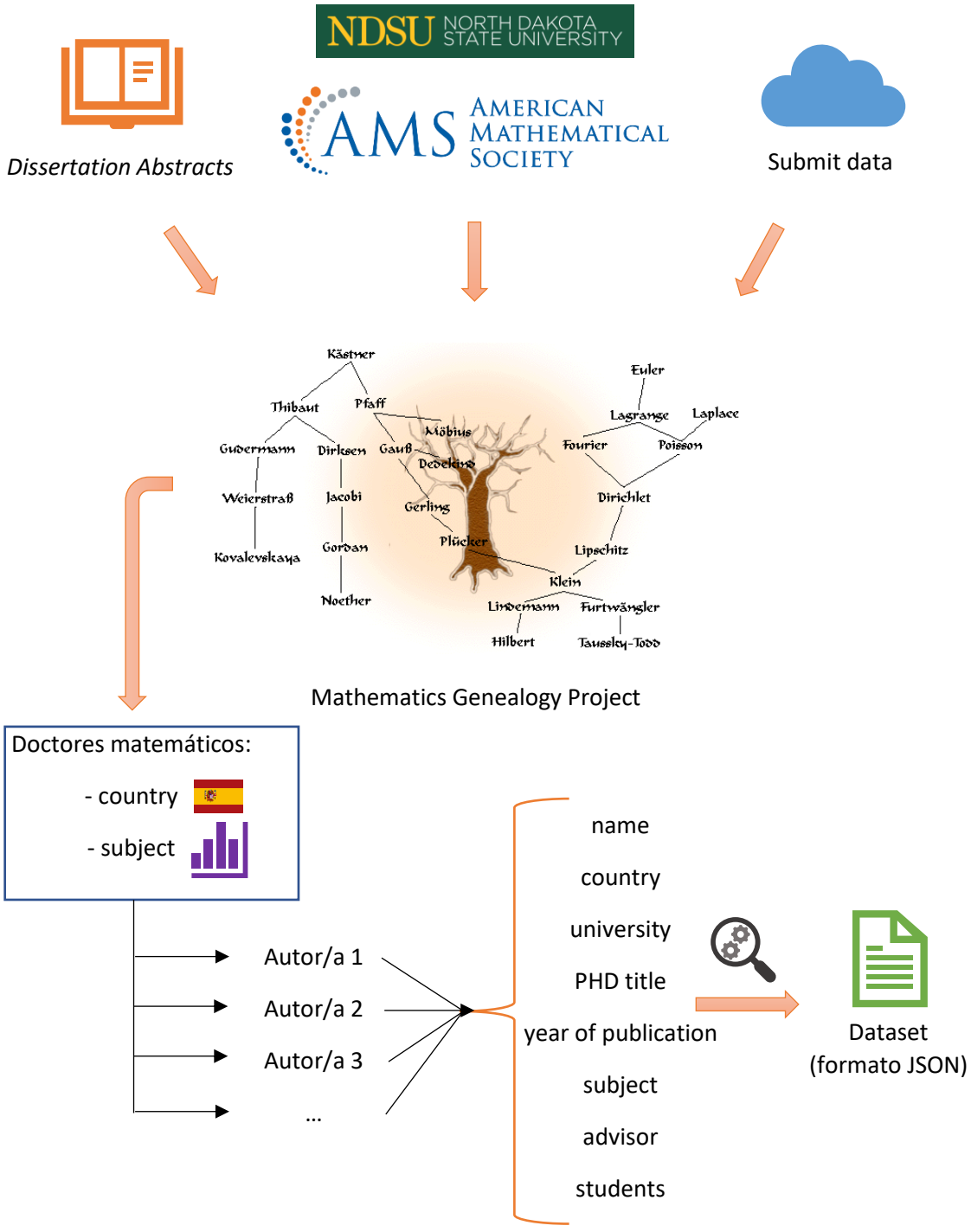
Como dice el título, el conjunto de datos está formado por la información relevante que la página web contiene de los matemáticos doctorados en España en el área de la estadística (nombre, universidad, año, país, título, asignatura, tutores y estudiantes). La extracción se realizó por última vez el 9 de abril, así que los autores incluidos son los que hasta ese día disponía el sitio. Siempre y cuando la estructura del sitio web no se modificase, cualquier cambio o adición de información de cualquier autor del dataset, como la introducción de un nuevo autor doctorado en España, no supondría ningún problema. Se ejecutaría de nuevo el código y el dataset se actualizaría.

El conjunto de datos extraído está en formato JSON porque el conjunto de datos es estructurado y la hora de representar variables con formato array es más adecuado.

Jorge Sainero Valle
Álvaro López Cabello

4. Representación gráfica

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



Jorge Sainero Valle
Álvaro López Cabello

5. Contenido

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

El dataset está formado por los campos: *country*, *subject*, *number_of_records* y *mathmaticians* donde este último es el más importante.

El campo *country* contiene el nombre del país que se ha empleado en la búsqueda, *subject* contiene el código del área empleado en la búsqueda, *number_of_record* que indica el número de matemáticos que hay en el campo *mathmaticians*.

El campo *mathmaticians* es el que aporta información al dataset y consiste en un array con la información de los distintos matemáticos de la búsqueda. Los campos que identifican a un matemático son: *name*, *country*, *university*, *PHD title*, *year of publication*, *subject*, *advisor* y *students*.

- *Name*: nombre del matemático.
- *Country*: País donde está la universidad donde realizó la tesis.
- *University*: Universidad donde realizó la tesis.
- *PHD title*: Título de la tesis.
- *Year of publication*: Año de publicación de la tesis.
- *Subject*: Asignatura o área al que pertenece la tesis.
- *Advisor*: Tutores de la tesis.
- *Students*: Alumno a los que el autor dirigió la tesis.

Los datos datan del siglo XII hasta la actualidad ya que es la fecha en la que se empezó a recolectar información sobre la investigación en el campo de las matemáticas.

Los datos se introducen en la web rellenando un formulario que después se valida por parte de la organización creadora de la web y nosotros hemos utilizado web scraping para extraerlo.

Primero utilizamos *Selenium* para navegar por la web y realizar la búsqueda con los datos que nos interesan y después utilizamos *BeautifulSoup* para extraer del contenido los campos que nos interesan. Después de la primera búsqueda obtenemos un identificador de cada autor que aparece en esta y para extraer la información del autor hacemos una petición a la web con el id de este como parámetro.

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. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

La Universidad de Dakota del Norte (NDSU) es la propietaria del conjunto de datos extraídos de la página web https://www.genealogy.math.ndsu.nodak.edu/, y Mitchell T. Keller el director gerente. Aunque se podría contactar con ellos a través del departamento de matemáticas, la web del proyecto tiene un formulario de contacto para enviar los mensajes a las personas encargadas según el motivo.

Existen algunos análisis de la base de datos del proyecto. El primero [2] es un artículo que estudia la evolución histórica del pensamiento matemático y su expansión espacial a partir de los datos del MGP y otros conjuntos académicos en línea. Engin Arslan y sus compañeros [3] analizan los vínculos académicos de los matemáticos, y Priya Narayan en su tesis doctoral [4] representa el dataset de MGP como una red. Otro estudio evalúa el papel de la tutoría [5]. Y una exploración computacional de la base de datos se realiza mediante el lenguaje Wolfram [6].

Para actuar de acuerdo con los principios éticos y legales, se ha analizado el contenido de *robots.txt*. Según el archivo, podemos rastrear todas las páginas del sitio web excepto la de envío de datos (/submit-data.php).

```
User-agent: msnbot
Crawl-delay: 30

User-agent: Browsershots
Disallow:

User-agent: *
Disallow: /submit-data.php
```

La información rastreada para la obtención del dataset es pública, y todos tenemos acceso a ella. Por otro lado, no disponen de API, y tampoco de condiciones de uso, salvo si se necesitaran sus datos para investigación. En tal caso se rellenaría un acuerdo de permiso para que analizaran la solicitud.

7. Inspiración

Explicar 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.

El conjunto de datos es interesante porque permite conocer los doctores y el título de sus tesis sobre un área publicados en una universidad del país, concretamente, sobre estadística en España, aunque el código está escrito para ejecutar el script según los dos parámetros, al menos uno obligatorio. Es un dataset descriptivo acotado y de fácil búsqueda, interesante para quienes deseen investigar sobre estadística en España. Según el título de la tesis o el número de estudiantes, por ejemplo, puede llamar la atención un estadístico u otro o un trabajo en particular y querer buscar más información acerca de ellos.

Las preguntas que se pretenden responder son:

- ¿Quiénes son los/as autores/as con un doctorado en España en el área de estadística?
- ¿En qué universidades españolas hay mayor cantidad de graduados?
- ¿Existe un aumento en el interés de la estadística en España?
- ¿Cuál es la media de estudiantes que asesoran?

Los análisis anteriores pretenden responder a otras preguntas. El primero de ellos [2] se enfoca en narrar la historia de las matemáticas y la transformación de conocimiento a partir de los datos mediante herramientas para analizar bases de datos genealógicas que podrían extrapolarse, por ejemplo, a AstroGen, el Proyecto de Genealogía Astronómica.

El objetivo principal de los estudios [3] y [4] consiste en analizar las relaciones entre asesores y estudiantes mediante la teoría de redes, el primero concretamente utilizando las teorías de redes sociales. En él también investigan las correlaciones entre universidades y países y su evolución a lo largo de los años. En la tesis, gracias al modelado del MGP mediante tres redes diferentes, se pretenden comprender los patrones en las relaciones y la influencia de los asesores en sus estudiantes.

El estudio [5] también desea evaluar el papel de los mentores en el desempeño de sus estudiantes a partir del número de estudiantes que cada uno capacita (graduados entre 1900 y 1960). Construyen redes aleatorias a partir de la red de genealogía matemática.

Finalmente, en el análisis [6] usan el lenguaje de programación Wolfram para explorar los datos del proyecto genealógico, realizar cálculos y visualizaciones que aportan información sobre los matemáticos y sus trabajos.

Los análisis presentados solo muestran los resultados obtenidos y no los datasets, a excepción del último en el que puede verse parte del código.

Ninguno trabaja exactamente con los datos obtenidos para el dataset. Además, al ser un proyecto en constante crecimiento, el número de registros va aumentando, y se requieren de análisis actualizados.

8. Licencia

Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección.

La licencia escogida para este conjunto de datos ha sido CC BY-SA 4.0 license. Esta licencia pública podrá ser utilizada siempre y cuando se mencione a los autores (BY) y toda modificación que se haga sobre el conjunto de datos deberá ser publicada con la licencia original (SA).

No hemos considerado añadir restricciones comerciales por lo que se podría sacar rédito económico del dataset y a lo mejor esto incentiva a una organización a utilizar el conjunto de datos y así dar publicidad a este.

9. Código

El código de la práctica se encuentra en el siguiente repositorio de GitHub: https://github.com/jsainero/WebScraping.

10. Dataset

El DOI del dataset es el siguiente: https://doi.org/10.5281/zenodo.6436865.

Jorge Sainero Valle
Álvaro López Cabello

11. Tabla de contribuciones

Contribuciones	Firma
Investigación previa	Álvaro LC, Jorge SV
Redacción de las respuestas	Álvaro LC, Jorge SV
Desarrollo del código	Álvaro LC, Jorge SV

Referencias

[1] North Dakota State University. *Mathematics Genealogy Project.* [En línea]. Disponible en: https://genealogy.math.ndsu.nodak.edu/index.php

[2] Gargiulo, F., Caen, A., Lambiotte, R., & Carletti, T. (2016). *The classical origin of modern mathematics*. EPJ Data Sci. 5, 26. [En línea]. Disponible en: https://doi.org/10.1140/epjds/s13688-016-0088-y

[3] Arslan, E., Hadi Gunes, M., & Yuksel, M. (2011). *Analysis of academic ties: A case study of Mathematics Genealogy.* IEEE GLOBECOM Workshops (GC Wkshps). [En línea]. Disponible en: https://doi.org/10.1109/GLOCOMW.2011.6162384

[4] Narayan, P. (2011). *Mathematics Genealogy Networks*. University of Oxford. [En línea]. Disponible en: http://people.maths.ox.ac.uk/porterm/research/priya_thesis_final.pdf

[5] Malmgren, R., Ottino, J., & Nunes Amaral, L. (2010). *The role of mentorship in protégé performance.* Nature 465, 622-626. [En línea]. Disponible en: https://doi.org/10.1038/nature09040

[6] Enright, A., & Weisstein, E. (2018). *Mathematics Genealogy Project: Computational Exploration in the Wolfram Language.* Wolfram. [En línea]. Disponible en: https://blog.wolfram.com/2018/08/02/computational-exploration-of-the-mathematics-genealogy-project-in-the-wolfram-language/