**PRÁCTICA 1: Web scraping**

**Tipología y ciclo de vida de los datos**

**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 “mathematicians\_PhD\_spain”.

**3. Descripción del dataset**

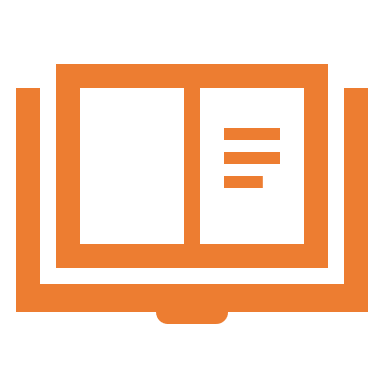
**Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.**

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 ya que mejora la interpretación de pares nombre/valor y *arrays*.

**4. Representación gráfica**

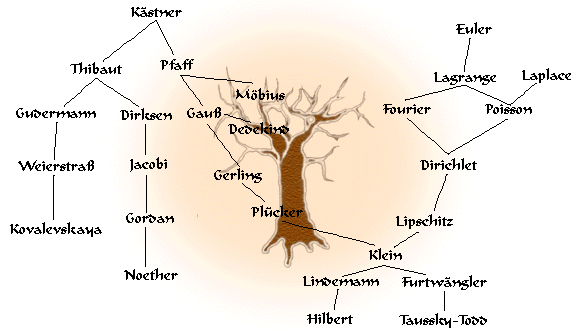
**Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.**





Submit data

*Dissertation Abstracts*



Mathematics Genealogy Project

Doctores matemáticos:

- country

- subject

España - Wikipedia, la enciclopedia libre



name

country

university

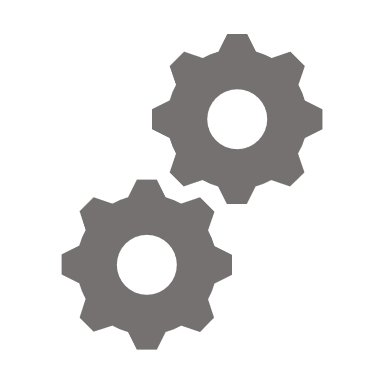
PHD title

year of publication

subject

advisor

students



Autor/a 1

Autor/a 2

Autor/a 3

…

Dataset (formato JSON)

**5. Contenido**

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

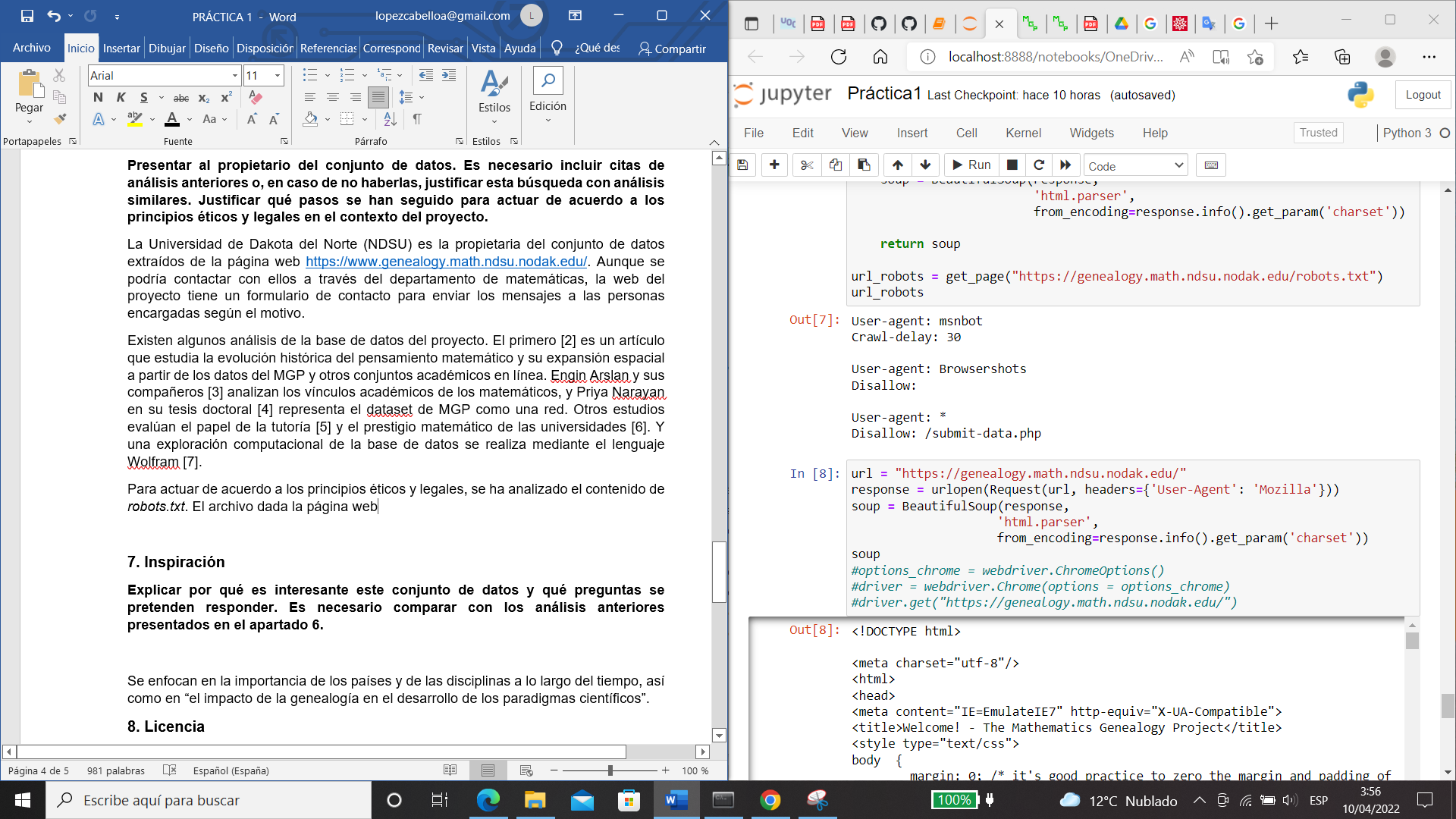
**6. Agradecimientos**

**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. 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. Otros estudios evalúan el papel de la tutoría [5] y el prestigio matemático de las universidades [6]. Y una exploración computacional de la base de datos se realiza mediante el lenguaje Wolfram [7].

Para actuar de acuerdo a 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).



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.**

**8. Licencia**

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

**9. Código**

**10. Dataset**

**11. Video**

**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] Myers, S., Mucha, P., & Porter, M. (2011). *Mathematical genealogy and department prestige.* Chaos 21, 041104. [En línea]. Disponible en: <https://doi.org/10.1063/1.3668043>

[7] 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/>