Tipología y ciclo de vida de los datos

PRA1: WEB SCRAPING

Antes de comenzar con la explicación, se me indicó que dejase claro que estoy realizando esta práctica de manera individual ya que, a pesar de hacer todo lo posible para encontrar una pareja, existió un malentendido. En un primer lugar me comprometí con un compañero a formar equipo, pero por motivos personales ambos estuvimos varios días sin ponernos en contacto. En esos días él recibió una propuesta que le convenció más y yo me quedé solo cuando no quedaba tiempo suficiente para emparejarme.

1. Contexto:

El primer motivo por el que se ha elegido este dataset es por pura motivación personal. Como matmático, cuando aún estaba en la facultad un profesor nos mostró la página web sobre la que se ha llevado a cabo el proyecto (https://www.genealogy.math.ndsu.nodak.edu/) y recuerdo que en su momento mis compañeros y yo nos pásabamos horas recorriéndola y buscando que grado de *parentesco profesional* tenían aquellos profesores que tanto admirábamos con los genios del pasado que estudiábamos en la carrera, ya que al final casi todos los doctores en matemáticas están conectados académicamente en mayor o menor grado.

Más allá de esto la página es de dominio público, lo que ofrece una mayor facilidad a la hora de acceder a los datos ya que su robots.txt es muy permisivo con el web scraping y, como añadido, ofrece una gran variedad de variables de búsqueda tanto seleccionables como escritas lo que también hace que en la práctica se puedan usar más variedad de herramientas.

Aunque dicho esto, el motivo principal es que los datos con los que contamos aquí tienen un valor científico y demográfico. La investigación es un reflejo de lo avanzada que está una sociedad y, en campos donde a priori no se tiene por qué buscar una aplicación práctica, esto se acrecenta aún más. Puesto que descubrimientos que hoy son aparentemente inútiles, con los avances tecnológicos pueden resultar marcando la diferencia; tal y como pasó con los algoritmos destinados a clusterización y machine learning, que no resultaron determinantes hasta la aparición de Internet y posteriormente la aparición de los daros masivos (el Big Data). Por todo esto, analizando el número de doctores y sus especialidades, podemos ver qué países están más desarrollados, y incluso analizando año a año podemos ver en que temas la investigación está en auge y en cuales está decayendo; dando esto un reflejo inequívoco del desarrollo y prosperidad de un país.

2. Título:

Aunque podemos generar una gran variedad de datos en base a los parámetros dados, se he elegido como ejecución por defecto la siguiente:
Título descriptivo: Generación de un histórico de los doctores cuyas tesis doctorales han sido realizadas en España en el ámbito de la ciencia computacional.

Título corto para el CSV: "**historical_spanish_PhD_computer_scientis.csv**"

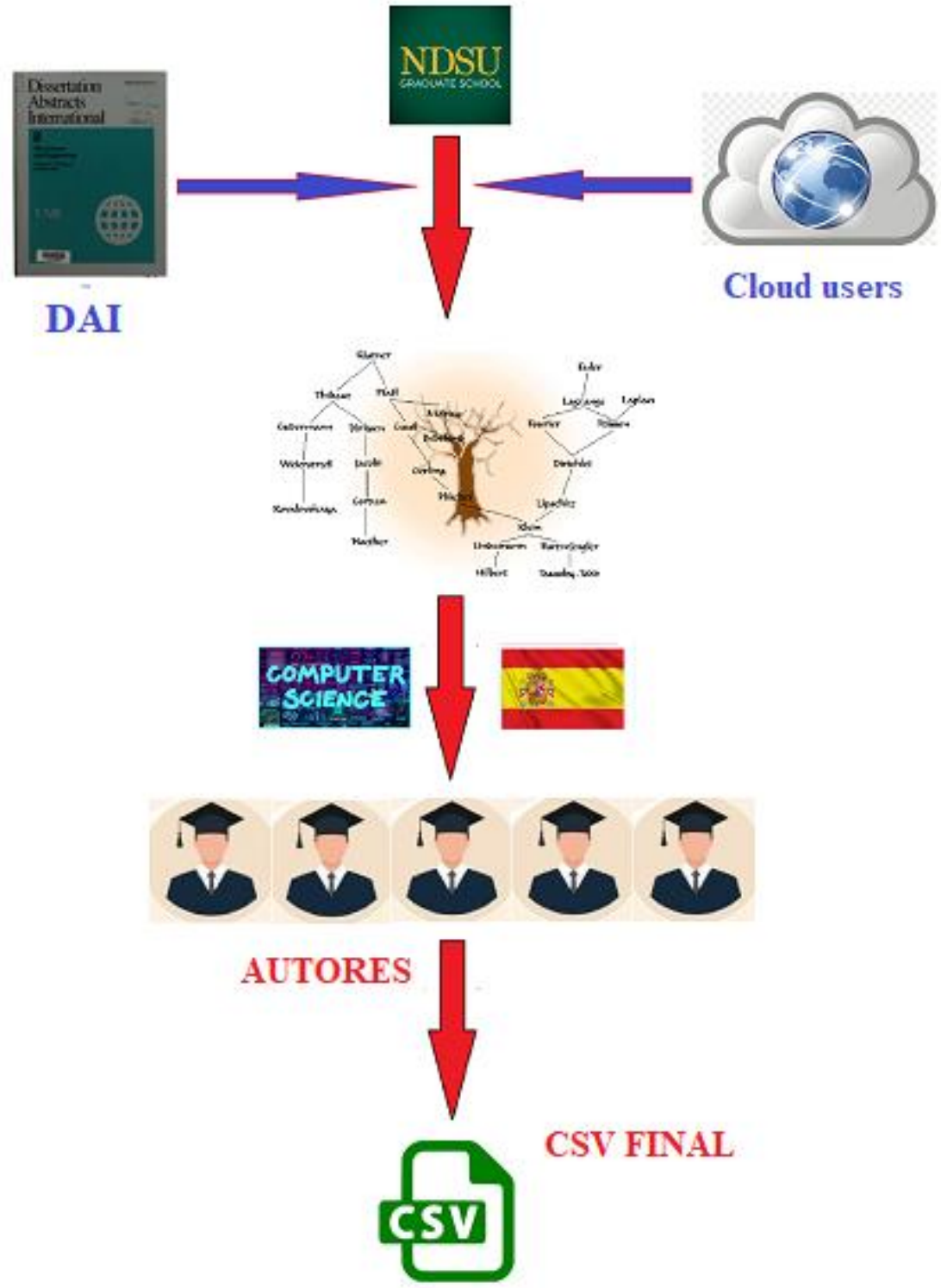
3. Descripción:

Como se comentaba en el apartado anterior, el set de datos contiene información de aquellos doctores que realizaron su tesis en España en el ámbito de la ciencia computacional. Dentro de este marco se ofrece una gran cantidad de información como puede ser saber como ha ido variando el número de doctorados al año desde que apareció este campo hasta hoy en día, así como por los títulos de las tesis podríamos ver una muestra de cómo se ha ido avanzando. También podríamos realizar un análisis por universidades, e incluso ver cuáles son más colaborativas fijándonos en el número de maestros y alumnos por doctor. Como información general el set de datos será un fichero en formato .csv separado por comas, que recogerá los datos desde 1981 (primer doctorado en España en este ámbito) hasta la actualidad; con un total de **334 filas** a fecha de ejecución del 22 de Noviembre de 2022 y **10 columnas**. El proceso de obtención de este set datos comienza cuando desde el **Departamento de genealogía matemática la universidad de Dakota del Norte** generan una base de datos pública a partir de las publicaciones de sus usuarios en internet, la cual complementan con los datos de Dissertion Abstract Internacional. A partir de esta generan su web que nos permite acceder a la base datos, y es aquí donde aplicamos las técnicas de web scraping para obtener un listado de todos los autores de tesis filtrando por país y especialidad. Una vez hecho esto, recorremos los enlaces donde está la información de dichos autores para ir rellenando un dataframe de pandas, que finalmente guardamos como CSV pudiendo así extraer la información.

4. Representación gráfica:

Vamos ahora a visualizar el esquema de obtención del dato ya explicado en la descripción anterior, pero esta vez de forma más ilustrativa. La parte del manejo del dato que se ha realizado en esta práctica sería desde donde aparece el árbol de conexiones (que representa la web) hasta la obtención del CSV final. La parte previa a la web correspondería a la obtención y tratamiento del dato que realizan desde la universidad de Dakota del Norte como propietarios del dato:

Este esquema puede verse gráficamente a continuación:


NDSU
GRADUATE SCHOOL
Dissertation Abstracts International
DAI
Cloud users
COMPUTER SCIENCE
AUTORES
CSV FINAL
CSV

5. Contenido:

Como comentábamos anteriormente, la estructura del set de datos será un fichero en formato .csv separado por comas, que recogerá los datos **desde 1981 (primer doctorado en España en este ámbito) hasta la actualidad**; con un total de **334 filas** a fecha de ejecución del 22 de Noviembre de 2022 y **10 columnas** que serán como sigue:

- **Autor**: Nombre el doctor.
- **PhD_publication_year**: Año de publicación de la tesis.
- **University_name**: Nombre de la universidad.
- **Country_name**: País de la universidad (en este caso España)
- **PdH_tittle**: Título de la tesis.
- **Subjet**: Tema de la tesis (ciencia de la computación en este caso)
- **Number_of_mentors**: Número de mentores / tutores.
- **Mentors**: Nombre de los tutores.
- **Number_of_students**: Número de estudiantes.
- **Students**: Nombre de los estudiantes.

La forma de obtención de dichos datos será mediante el uso de la librería Selenium sobre la web que nos permite conectar con la base de datos anteriormente mencionada. Una vez tenemos identificado el dato que nos interesa lo tratamos con BeautifulSoup para poder extraer su contenido y darle, finalmente, el formato deseado. De hecho, las principales dificultades del código residen en primer lugar en el manejo de Selenium para ser capaces de identificar los objetos sobre los que nos movemos inspeccionando los elementos de Chrome y, sobre todo, lo más difícil será interpretar el dato con BeautifulSoup, lo cual ha requerido de mucho tiempo analizando el formato en que esta librería nos proporciona el dato para poder adaptarlo posteriormente al formato que se nos requería.

6. Propietario:

Como ya se ha especificado, el propietario de dicha base de datos será el **Departamento de genealogía matemática la universidad de Dakota del Norte**, en Estados Unidos. Aunque esta base de datos será de dominio público y cualquiera podrá acceder a ella desde https://www.genealogy.math.ndsu.nodak.edu , pudiendo incluso escribir sobre esta siempre que se pasen los pertinentes controles de calidad del dato.

Respecto a la existencia de estudios sobre esta fuentes de datos, podemos encontrar varios, e incluso algunos en la propia página. Un caso curioso de esto es el de [1] Priya Narayan, que realiza su tesis basándose en la genealogía representando el dataset de MGP como una red. Otro muy interesante es el que estudia la evolución histórica del pensamiento matemático y su expansión utilizando el MGP [2]. Finalmente tenemos un enfoque diferente que evalúa las relaciones para crear relaciones entre el éxito en el doctorado y la tutoría [3].

Finalmente, con respecto a respetar los márgenes éticos y legales, se ha analizado el fichero robots.txt de la página para ver su política frente al web scraping y el resultado ha sido el siguiente:

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

User-agent: Browsershots
Disallow:

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

Como podemos ver, todo está permitido salvo el envío de datos, asegurándose así que este pasa por ellos y cumple los controles de calidad. Sin embargo, respecto al web scraping no hay limitaciones de uso por lo que podemos acceder a todo para obtener la información necesaria.

7. Inspiración:

Tal y como se ha comentado en la motivación inicial, las posibilidades que nos ofrece este dataset son muy amplias, ya que ser capaces de analizar la tendencia en la que se mueve la investigación, se traduce en poder predecir cuáles serán los mercados emergentes y poder sacar una ventaja real de esto. Tecnologías como los NFTs o las criptomonedas, nacieron como resultado de la prosperación en la investigación del BlockChain, y como esto podemos encontrar muchos otros ejemplos que nos muestran el valor, incluso comercial, que se puede obtener de estos datos. Para nuestro caso concreto en el que nos centramos en dato de España para los doctorados en ciencia de la computación, podríamos plantearnos cuestiones como el estudiar el posible crecimiento que se ha producido en esta área respecto a años anteriores, analizar en que universidades de España se están dedicando más a esta especialidad, e incluso basándonos en los títulos analizar cómo han ido avanzando las tecnologías. Por último, podríamos valorar la importancia del asesoramiento en estos campos.

Respecto a los análisis anteriores, el caso [3] se centra bastante en este último punto, desde una analítica bastante avanzada llega a obtener conclusiones muy interesantes. Por otro lado, para los casos [1] y [2] también podríamos sacar valor sobre este dataset, puesto que al pretender evaluar el grado de interconexiones entre matemáticos, así como la evolución del pensamiento matemático, que podríamos basar en títulos en lugar de en temas, podríamos aplicar un modelado de redes MGP para obtener patrones relacionales y obtener conclusiones de calidad.

8. Licencia:

Finalmente se ha escogido como licencia la **Attribution 4.0 Internacional**, ya que basándonos en la fuente del dato parece ético mantener una política continuista en cuanto a que el dato que trata sobre la investigación y está pensado para la investigación debe ser internacional y de libre acceso. Aún así se reserva el uso al propietario en este caso de otorgar o no el derecho a usarlo para fines comerciales.

9. Código:

Podremos encontrar el código en el siguiente enlace de github:
https://github.com/jmguerreroca/WebScraping

10. Dataset:

Podremos encontrar el DOI del dataset obtenido en el en enlace:
https://doi.org/10.5281/zenodo.7346828

11. Tabla de contribuciones:

Contribuciones	Firma
Investigación previa	Juan María Guerrero Carrasco
Redacción de propuestas	Juan María Guerrero Carrasco
Desarrollo del código	Juan María Guerrero Carrasco

Video explicativo:
https://drive.google.com/drive/u/3/folders/1qFPfsQQCRtDm3mBLa586vbOggp3E23i_

Bibliografía:

[1] 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

[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] 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-mathematicsgenealogy-project-in-the-wolfram-language/