Práctica 1: Web Scraping

0. Integrantes.

Bryan Steven Cortez Chichande

César Alexander Guzmán Vásquez

1. Contexto. La información recolectada tiene relación con los datos bibliográficos de documentación científica (libros, artículos científicos, ensayos) orientados a temas como Data Science, Data Mining, Web Mining, Web Scraping, Text Mining y Data Visualization.

Para este proyecto se seleccionaron tres páginas web a las que quisimos analizar. El primer sitio web seleccionado es **Semantic Scholar**, el mismo que es un motor de búsqueda respaldado por un sistema de inteligencia artificial dedicado a trabajar con publicaciones académicas. A su vez provee resúmenes de documentación científica de varias áreas de interés, proporcionando los datos necesarios para referenciar y citar publicaciones con sus respectivos autores.

El objetivo de este dataset es poder visualizar en que áreas se están desarrollando más publicaciones con rigor científico relacionadas con la ciencia de datos, lo que permite abrir un panorama de posibilidades para futuros estudios, en los cuales su área esté en auge o por el contrario, no hayan sido muy explorados. Adicionalmente se podrá obtener información sobre el ranking de los artículos más citados, su influencia e impacto, el número de referencias que ha tenido y si es de libre acceso o no.

El segundo sitio web analizado es **Dialnet**, que es un portal de difusión de la producción científica hispana especializado en ciencias humanas y sociales. El objetivo con este sitio fue probar los mecanismos de seguridad del web scraping, donde se hizo uso del User-Agent y timers para poder obtener la información sin problemas.

El tercer sitio web analizado es **Google Scholar**, es un motor de búsqueda de Google enfocado y especializado en la búsqueda de contenido y bibliografía científico-académica. El objetivo con este sitio web es similar al primero, obtener los resultados y observar las áreas de mayor desarrollo e investigación de la ciencia de datos. Y a su vez comparar los resultados con los que arroja Semantic Scholar y notar diferencias y coincidencias.

2. Título. Datos bibliográficos de publicaciones científicas de las áreas anexas a la ciencia de datos.

3. Descripción del dataset. Para el primer dataset, el conjunto de datos recopilado contiene una gran cantidad de información bibliográfica de publicaciones científicas relacionadas con la ciencia de datos. Tomando en cuenta la relevancia de estos, sin importar el año de publicación.

El dataset incluye un total de 7278 registros obtenidos del web scraping hechos a la página, por medio de la API que proporciona Semantic Scholar.

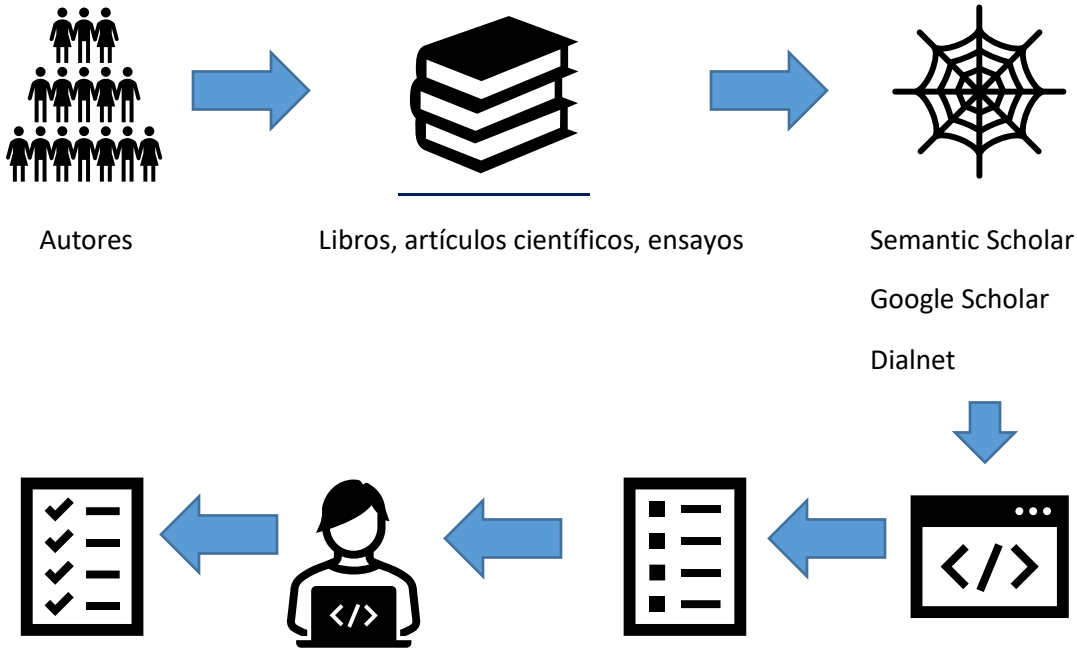
El segundo dataset incluye un total de 982 registros obtenidos del web scraping a la página Dialnet, por medio de diferentes técnicas.

El tercer dataset incluye un total de 510 registros obtenidos del web scraping a la página Google Scholar, por medio de diferentes técnicas en las que se escarbó en el contenido de la página y sus resultados.

Debido a que las tres páginas tienen una estructura demasiado diferente, se optó por mantener 3 datasets separados en lugar de unirlos, ya que a efectos prácticos para este caso no habría utilidad en su unión.

4. Representación gráfica.

Proceso de creación del dataset



5. Contenido. Para el primer dataset de Semantic Scholar tenemos lo siguiente:

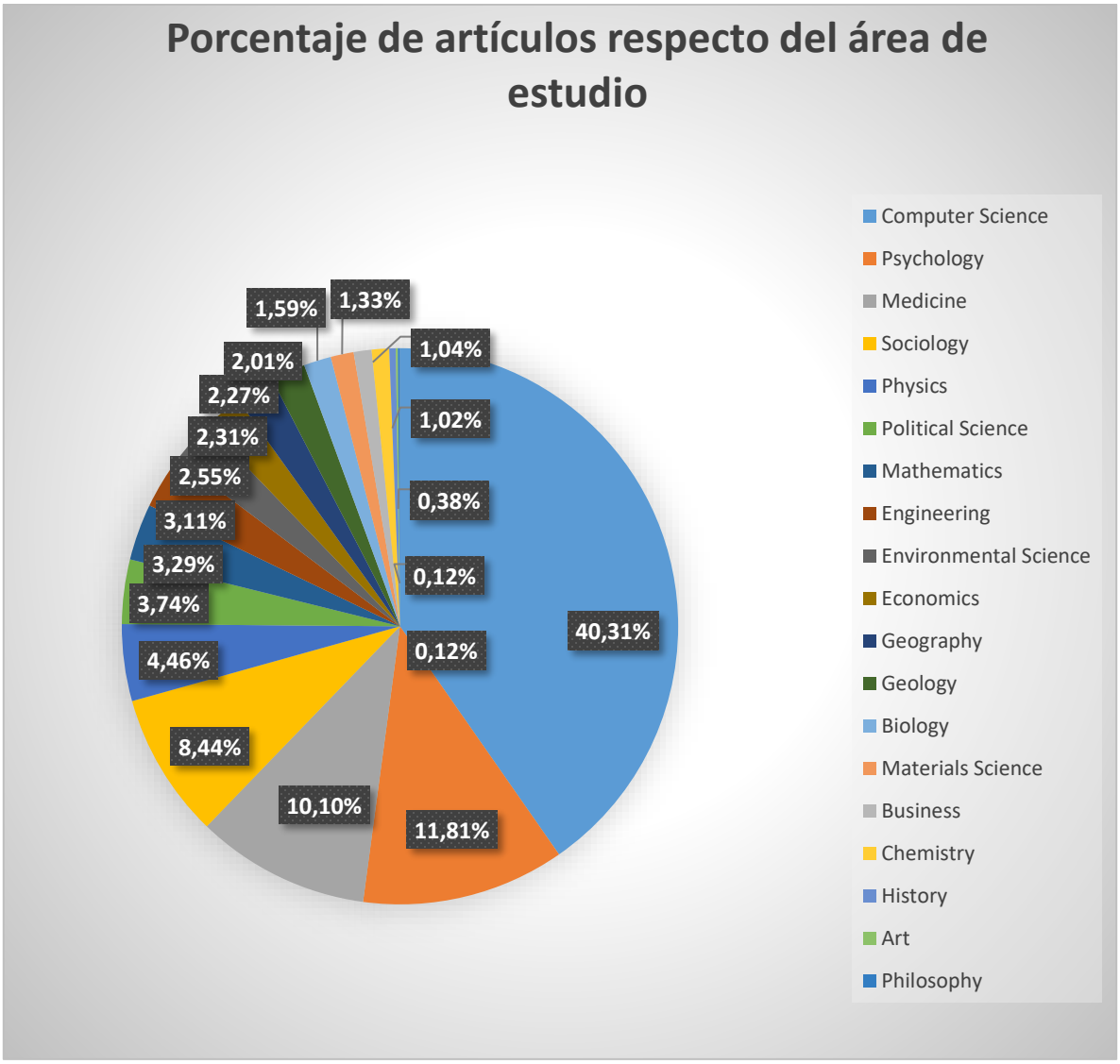
Cada publicación científica contenida en el conjunto de datos recopilado cuenta con los siguientes atributos:

Atributo	Tipo	Descripción
Paper ID	Cadena	Código identificador de la publicación científica
Título	Cadena	Nombre completo del documento
Año	Numérico	Año de publicación
URL	Cadena	Enlace a la página donde se encuentra el documento
Número de referencias	Numérico	Cantidad de publicaciones que han sido citadas en el documento.
Número de citaciones	Numérico	Cantidad de menciones a textos de autores ajenos a la publicación.
Número de citaciones influyentes	Numérico	Cantidad de menciones a textos relevantes de autores ajenos a la publicación.
Acceso abierto	Binario (0-1)	Descripción del tipo de licencia: • True: Libre acceso, False: Sin libre acceso
Campos de Estudio	Arreglo de Strings	El o las áreas de estudio dentro de la que se clasifica la publicación.

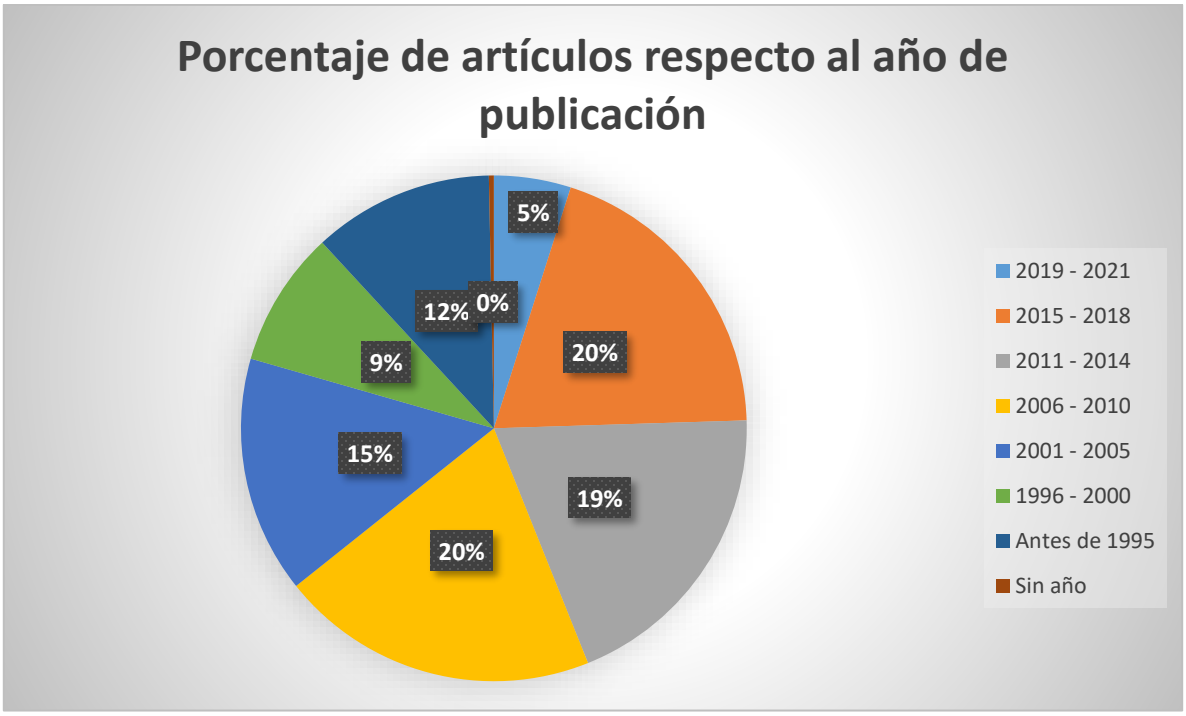
El período de tiempo de los datos se encuentra entre los años 1934 hasta el año actual, tomando como referencia la fecha de publicación de los documentos. Las publicaciones científicas que aparecen en la página web de Semantic Scholar son recopiladas de diferentes editoriales tales como Springer, IEEE, Scielo, entre otros.

Como breve análisis de los datos se puede observar cuanto menos curioso, y es el hecho de que las 5 primeras categorías más investigadas, no resultan tan obvias como se creería en la percepción común del público en general. Primero tenemos como principal categoría a las ciencias de la computación, esta área de estudio si puede resultar un tanto obvia porque esta área es la que se encarga de desarrollar más la ciencia de datos; sin embargo, si analizamos las siguientes categorías más estudiadas, vemos que tenemos a la psicología como segunda área más estudiada, lo que nos hace pensar que se debe al estudio de la mente del ser humano para poder mejorar los comportamientos de las IA. En tercer lugar, podemos ver el área de medicina, lo cual, si nos resulta más familiar, por todos los artículos que se escuchan donde se une la ciencia de datos con las diferentes áreas médicas, desde tratamientos, detecciones tempranas de enfermedades, etc. En el cuarto lugar vemos a la sociología, la cual también es un área poco común en el público en general, pero se entiende porque se trata de detectar patrones en el comportamiento de la sociedad humana, lo que es perfecto para la ciencia de datos. Y por último, tenemos en quinto lugar a la física, lo que muestra la gran importancia que está teniendo la ciencia de datos y la IA en el entendimiento del universo, sobre todo cuando se habla de física cuántica.

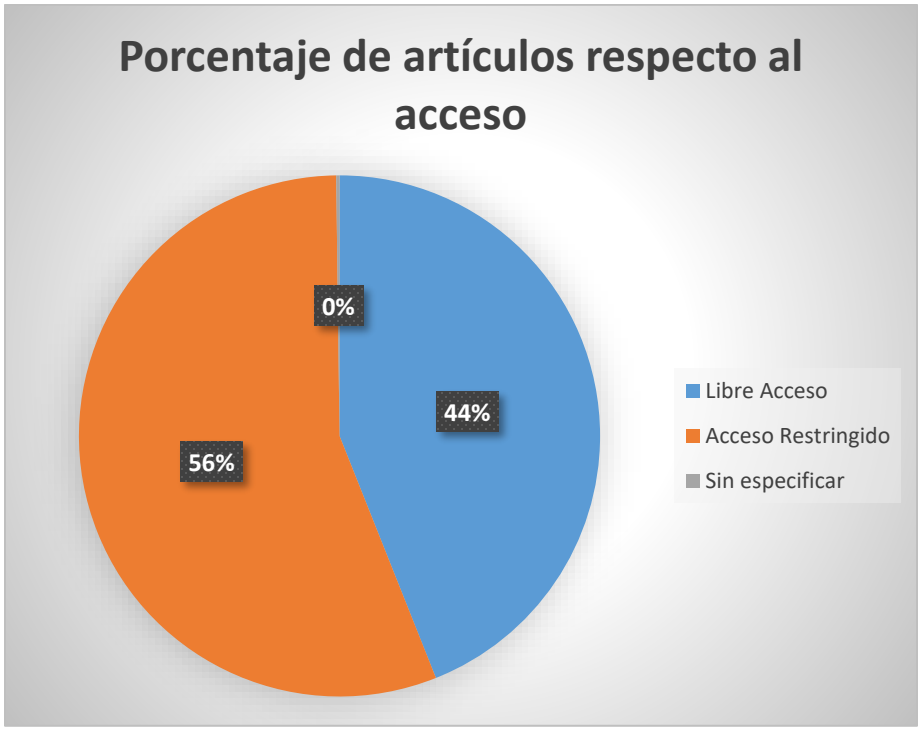
El resto de áreas que son minoritarias, son las que nos resultan más comunes al oído son justamente las menos estudiadas, como política, medio ambiente, economía, ingeniería, química, biología, entre otros.



Si se continua con el análisis rápido de los datos, podemos observar también que en los últimos 3 años, se han publicado a penas el 25% de los artículos relacionados con la ciencia de datos, si los comparamos con los periodos anteriores desde el 2010, que fue cuando hubo un verdadero auge en la materia. Si vamos un poco más atrás vemos que desde el año 2000 hasta el año 2010 se han hecho una quinta parte de todas las publicaciones relacionadas a la ciencia de datos, que era el periodo donde se empezaba a explorar sus bondades. Y por último, antes del año 95, apenas era explorado el tema, en comparación con las dos últimas décadas.



Continuando con nuestro breve análisis del dataset, podemos ver que tristemente la mayoría de la información o conocimiento sigue siendo de acceso restringido o sujeto a algún tipo de licencia. Sería interesante observar para futuros análisis si las publicaciones de la última década han disminuido o aumentado el libre acceso a la información, lo que nos daría una idea de que tanto ha avanzado la idea de una sociedad open-source respecto al conocimiento.



Para el segundo dataset de Dialnet tenemos lo siguiente:

Cada publicación científica contenida en el conjunto de datos recopilado cuenta con los siguientes atributos:

Atributo	Tipo	Descripción
Autor	Cadena	Nombre del autor de la publicación
Editorial	Cadena	Editorial de la publicación
ISSN-e	Cadena	Código ISSN-e de la publicación
ISSN	Cadena	Código ISSN de la publicación
ISBN	Cadena	Código ISBN de la publicación
Volumen	Cadena	Número del volumen de la publicación
Número	Cadena	Número del segmento dentro del volumen de la publicación
Páginas	Cadena	Rango de las páginas donde se encuentra la publicación

Lamentablemente de este dataset, no se pudo obtener mayor información útil que ayude a nuestro análisis, pero sirvió para probar las diferentes seguridades de la página.

Para el tercer dataset de Google Scholar tenemos lo siguiente:

Cada publicación científica contenida en el conjunto de datos recopilado cuenta con los siguientes atributos:

Atributo	Tipo	Descripción
Autor	Cadena	Nombre del autor de la publicación
Especializacion	Cadena	La especialización del autor de la publicación
Total de citas	Número	Recoge el número total de publicaciones citadas del autor
Citas Ultimos 5 años	Número	Recoge el número total de publicaciones citadas del autor de los últimos 5 años
Total Indice H	Número	Es el mayor número h, de forma que h publicaciones se han citado al menos h veces
Indice H Ultimos 5 años	Número	Es el mayor número h, de forma que h publicaciones se han citado al menos h veces de los últimos 5 años
Indice i10	Número	Recoge las publicaciones que se han citado al menos diez veces
Indice i10 Ultimos 5 años	Número	Recoge las publicaciones que se han citado al menos diez veces de los últimos 5 años

Al principio para poder realizar el web scraping de Google Scholar se presentó como principal obstáculo el hecho de que cada cierto tiempo Google bloqueaba el acceso a su página por medio del script, por lo que se optaba por cambiar las IPs y configuraciones de DNS hasta que Google desbloqueara el acceso pasadas unas horas.

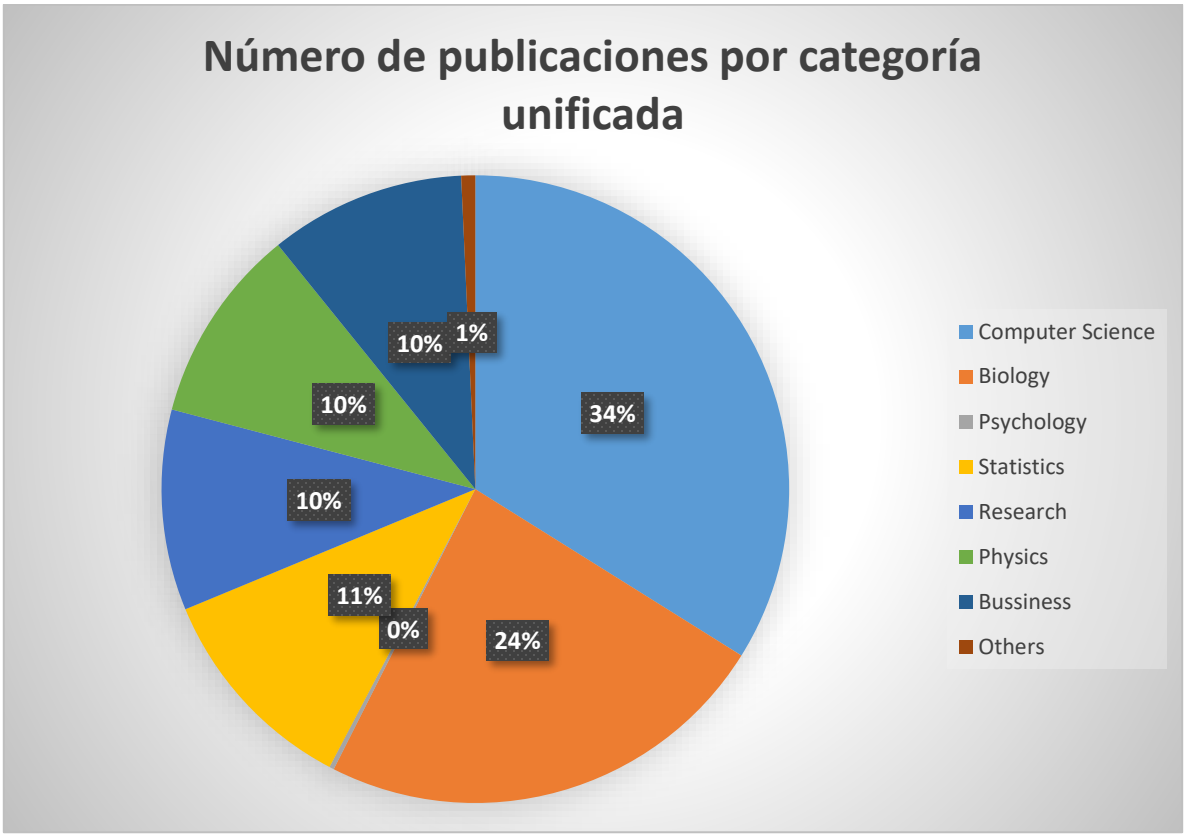
A diferencia de los dos casos anteriores, este script se lo realizó directamente en Python, codificándolo directamente en Spyder es una estructura de clases, y no como un Jupiter Notebook como los dos casos anteriores.

Superado ese obstáculo del bloqueo temporal de Google, podemos observar lo siguiente de los datos obtenidos en esta página. Si lo comparamos con el resultado de Semantic Scholar, podemos ver que aquí tenemos una distribución más uniforme en las categorías, sin embargo, en este dataset las categorías varían con las del dataset de Semantic Scholar, ya que son más específicas.

Para poder comparar de mejor forma con el gráfico de categorías de Semantic Scholar, se optó por unificar las categorías relacionadas en una sola, obteniendo categorías similares. En este caso como podemos apreciar vemos que la rama más estudiada sigue siendo las ciencias de la computación con un 34%. Seguido por la biología (que este caso abarca también las ramas de la medicina) en un 24%. Lo curioso es que en este caso la rama de psicología que en el otro dataset representaba alrededor de un 10%, aquí podemos ver que

es menos del 1%. Por otro lado, podemos observar que la física, los negocios, y la investigación tienen casi el mismo porcentaje, alrededor del 10%, La estadística se vuelve más importante y ocupa un 11%, mientras que todo el resto de ramas ocupa cerca del 1%.

Las principales diferencias que vemos en ambos dataset es que tanto la psicología como la sociología, son áreas de estudio que varían bastante dependiendo de la página de estudio. Probablemente depende de los países o de los años en lo que se estudiaron esos temas. Por lo que con esto se abren las puertas a más interrogantes para futuras investigaciones.



6. Agradecimientos.

Para el primer dataset, el propietario de los datos es la página **Semantic Scholar,** agradecemos la facilidad que da para acceder a sus datos por medio de su API. Para este caso de web scraping hemos actuado de acuerdo a las indicaciones del API, donde se especifica los límites de uso de esta.

Para el segundo y tercer dataset, los propietarios son **Dialnet** y **Google Scholar**, que agradecemos el proveer la información, sin embargo, en estas páginas, el acceso es más restringido y no cuentan con APIs que faciliten el acceso a la información.

En cuanto al tema ético, en el caso de Semantic Scholar se ha respetado completamente, ya que la información que provee la API es de acceso público, e incluso el dataset no contiene datos personales de ninguna persona, ya que tampoco incluimos la información

de los autores; si alguna persona quisiera saber más sobre el artículo tendría que ir al URL que indica el dataset, de forma que acceda a la información de forma directa desde la página oficial.

En los otros dos casos de Dialnet y Google Scholar, accedemos igualmente a información pública, no guardamos datos personales a excepción de los nombres de los autores y respetamos los tiempos de carga de los servidores para evitar bloqueos.

Gracias a páginas como Semantic Scholar, Dialnet y Google Scholar hemos logrado nuestro objetivo, que básicamente era poder visualizar de forma clara, que áreas de estudio estaban siendo más desarrolladas haciendo uso de la ciencia de datos o campos similares. Sería muy interesante a futuro, explorar más sitios o buscadores similares para ampliar la cantidad de datos y observar si las proporciones son las mismas o si por el contrario cambian dependiendo del buscador, como ya hemos podido ver en este caso.

7. Inspiración.

Para este proyecto nos inspiramos en el hecho de todas las infinitas posibilidades que ofrece la ciencia de datos y nuestra curiosidad por encontrar cuál resulta más atractiva para los investigadores desde su área, el poder observar como la tendencia de investigación ha cambiado según los años. Así también como el acceso a la información, el poder ver si el sueño de los movimientos open-source se va cumpliendo con el acceso universal y libre de la información, o si por el contrario, cada vez este acceso se vuelve más privativo. Por esa razón, estas preguntas y a otras que podrían surgir en el camino nos planteamos este proyecto, seleccionado a Semantic Scholar, Dialnet y Google Scholar como páginas de estudio.

8. Licencia.

La licencia escogida para los dataset resultantes es la “Released Under CC BY-NC-SA 4.0 License”, ya que se permite compartir y transformar libremente el dataset, dando el crédito a los creadores, haciéndola de uso no comercial y sus derivados manejan la misma licencia. Ya que los dataset se hicieron con fines meramente académicos, lo más adecuado a nuestra consideración es hacer uso de esta licencia.

9. Código. A continuación, se indica el link al repositorio en GitHub:

https://github.com/BryanSTV25/PRA1_Web_Scraping

10. Dataset. A continuación, se indica el link al repositorio en Zenodo:

https://zenodo.org/record/5636786#.YYAXA56ZOUk

11. Contribuciones.

Contribuciones	Firma
Investigación previa	BC, CG
Redacción de las respuestas	BC, CG
Desarrollo del código	BC, CG

Bibliografía

Rodó, D. M. (2020). El lenguaje Pyhton. Barcelona.

Subirats, L. M., & Calvo, M. G. (2019). Web scraping. In L. M. Subirats, & M. G. Calvo. Barcelona: 2019.