Práctica 1: Web scraping

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

Durante el año 2019 empezó la enfermedad coronavirus también conocida como COVID-19, una enfermedad provocada por el virus SARS-COV-2. La rápida expansión de la enfermedad hizo que la Organización Mundial de la Salud la declarara una emergencia sanitaria de preocupación internacional y la comunidad de todas las ramas científicas se volcaron a estudiar la grave enfermedad. En el conjunto de datos pueden verse diferentes artículos que hablan sobre la vacuna del COVID-19 recogidos de la página web arXiv.org.

arXiv.org es un archivo en línea que consta de artículos científicos en los campos de la física, las matemáticas, la astronomía, la ingeniería eléctrica, la informática, la biología cuantitativa, la estadística, las matemáticas financieras y la economía.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

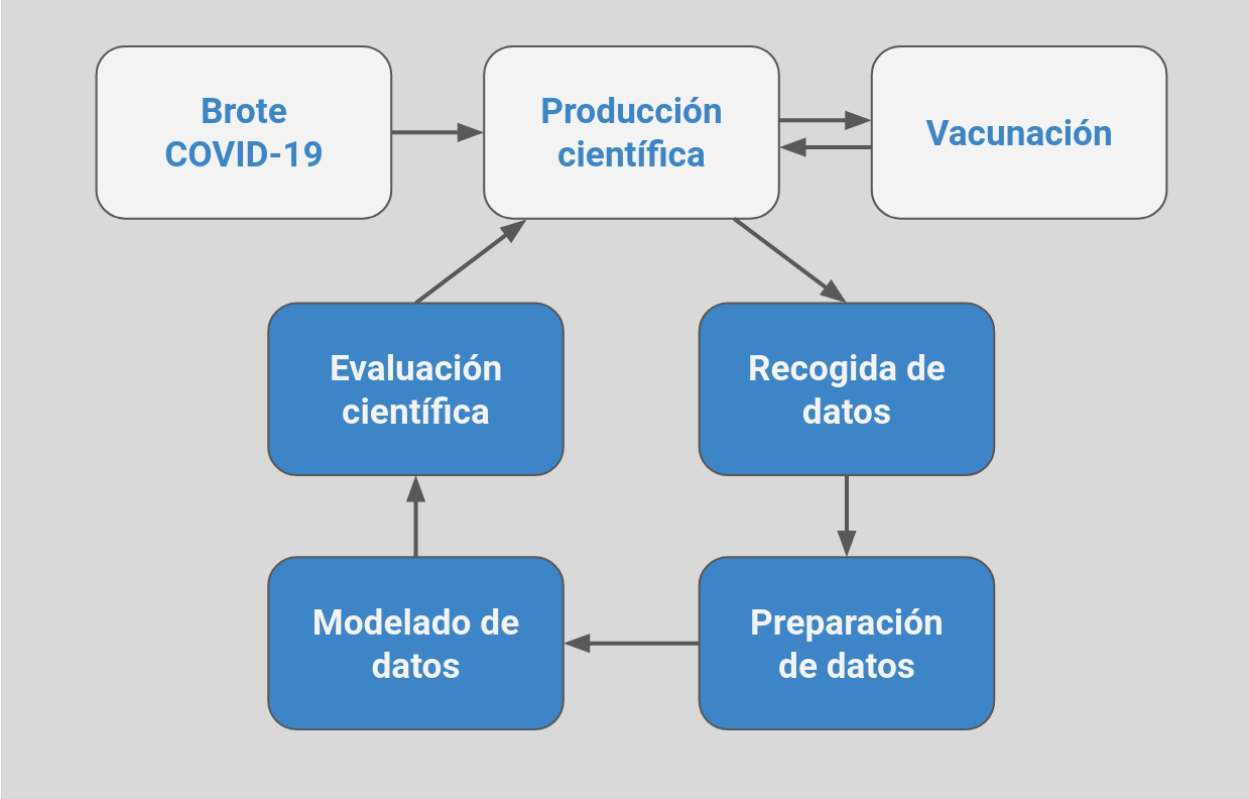
COVID-19 vaccine research on arXiv.org dataset

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).

En el dataset se almacena una colección de datos de publicaciones científicas relacionadas con las vacunas del COVID-19 publicadas en el repositorio arXiv.org. Se trata de una recopilación de todos los artículos disponibles en dicho repositorio, pero no exhaustiva del campo.

Para cada artículo recopilado, se almacenan hasta 12 campos de tipo texto, que permiten identificar el artículo, así como información de este como el campo específico de investigación, autores o fecha de publicación. Se almacena información de los 266 artículos, a 11 de abril de 2021, resultado de la búsqueda ‘vaccines AND (COVID-19 OR SARS-COV-2)’ en el resumen de las publicaciones almacenadas.

4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido



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

Nombre del campo	Tipo	Explicación	Ejemplo
id	Texto	Código de identificación del documento.	arXiv:2007.16063
category_code	Texto	Código de la categoría	[cs.CY]
category	Texto	Campo y categoría al que pertenece el documento	Computer Science > Computers and Society
title	Texto	Título del documento	A Review on the State of the Art in Non Contact Sensing for COVID-19
date	Texto	Fecha de publicación	[Submitted on 28 Jul 2020]

author_1	Texto	Nombre del primer autor	William Taylor
author_2	Texto	Nombre del segundo autor	Qammer H. Abbasi
author_3	Texto	Nombre del tercer autor	Kia Dashtipour
author_4	Texto	Nombre del cuarto autor	Shuja Ansari
author_5	Texto	Nombre del quinto autor	Aziz Shah
summary	Texto	Resumen del contenido	COVID-19 disease, caused by SARS-CoV-2, has resulted in a global pandemic recently. With no approved vaccination or treatment, governments around the world have issued guidance to their citizens to remain at home in efforts to control the spread of the disease.
link	Texto	link del documento	https://arxiv.org/abs/2007.16063

Los datos son relativamente recientes, ya que la enfermedad del COVID-19 apareció en 2019 y se empezó a buscar una vacuna después de que apareciese. Los datos se han recogido de la web arXiv.org la cual lleva recopilando artículos desde 1991, los datos que se incluyen en el dataset son la información básica de los artículos almacenados.

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.

Los datos de los artículos publicados se han recogido del repositorio archivado de arXiv.org, un servicio de distribución de publicaciones de diferentes disciplinas respaldado por la Universidad de Cornell, en abierto y activa desde 1993.

En concreto, la recolección de información se ha realizado en el dominio export.arxiv.org, preparado para la extracción automatizada de información. De esta forma no se interfiere con la navegación de los usuarios de la página principal.

Hasta la fecha, la producción científica de la COVID-19 y el coronavirus que lo provoca, el SARS-COV-2, ha sido objeto de varios análisis. Un análisis reciente es el que propone el equipo de Erika Morgana Neves de Oliveira, de la Universidad Federal de Piauí, en Brasil [1]. En él se realiza un análisis cuantitativo de mayor alcance, basado en el portal de referencia Web of Science (WOS). Este grupo realiza un análisis de las publicaciones por autores, la

nacionalidad de estos, así como de las revistas científicas en las que se publicaron dichos artículos.

Análisis similares al aquí presentado permitirían analizar la cobertura de medios de comunicación o conocer cuál es la forma de actuar de un cierto grupo de personas a partir de lo que publican en sus redes sociales. A partir de las fechas y comentarios poder saber en qué etapa, época u hora están más activos y sus intereses.

Referencias:
[1] OLIVEIRA, Erika Morganna Neves de et al. Analysis of scientific production on the new coronavirus (COVID-19): a bibliometric analysis. *Sao Paulo Med. J.* [en línea]. 2021, vol.139, n.1 [consultado 2021-04-11], pp.3-9. Disponible en: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-31802021000100003>. Epub 15 de enero, 2021. ISSN 1806-9460. https://doi.org/10.1590/1516-3180.2020.0449.r1.01102020.

7. Inspiración. Explique 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.

Este conjunto de datos es interesante porque podemos obtener información sobre el avance y uso de las diferentes vacunas contra el COVID-19. Podríamos responder preguntas como:

- ¿A qué velocidad se han desarrollado las vacunas y la producción científica sobre estas?
- ¿Qué campo de la ciencia se interesa más por la vacuna?
- ¿Qué categoría de cada campo de la ciencia se interesa más por la vacuna?
- ¿Qué autor/a ha escrito más sobre la vacuna?

Se puede comparar con el análisis anterior, ya que nuestro grupo de personas serían las diferentes comunidades científicas y los autores de cada artículo y poder conocer el avance de la vacuna del COVID-19 o cuál es la distribución temporal del debate científico sobre el tema. También poder conocer las aportaciones de las comunidades científicas a la investigación y preocupaciones por acabar con la enfermedad.

8. Licencia.

Se ha publicado el conjunto de datos bajo la licencia Creative Commons Attribution 4.0 International Public License, o CC BY 4.0, en el repositorio Zenodo. Esta licencia, detallada en el enlace a continuación, otorga las siguientes características al conjunto de datos licenciado:

- Atribución (BY): se permite la reutilización, copia, derivación y distribución del dataset bajo el único requerimiento de la cita a los autores. De esta forma se reconoce la autoría del conjunto de datos cuando este ha sido útil para análisis posteriores.

- Comercial: las acciones previamente definidas están permitidas, también con fines comerciales. Esto habilita un mayor impacto potencial del trabajo realizado, evitando la duplicación de tareas en torno a datos de interés público.
- Licencia: la licencia adoptada no impone sobre las obras derivadas restricciones adicionales sobre su licencia, pudiendo ser igualmente o más restrictivas según los autores estimen oportuno. Esta licencia se adapta por tanto a todo tipo de iniciativas, públicas o privadas.

https://creativecommons.org/licenses/by/4.0/legalcode.es

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El código con el que se ha generado el dataset se encuentra escrito en Python en el documento de Jupyter Notebooks que se encuentra a continuación.

https://github.com/sfunesolaria/Pra1WebScraping/blob/main/src/WebScraping.ipynb

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

El conjunto de datos se encuentra alojado en Zenodo, bajo el DOI a continuación.

https://doi.org/10.5281/zenodo.4679528

Contribuciones	Firma
Investigación previa	DO, SF
Redacción de las respuestas	DO, SF
Desarrollo código	DO, SF