UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

TFG del Grado en Ingeniería Informática

Impact Factor Oracle

Presentado por Gadea Lucas Pérez
en Universidad de Burgos
a 2 de febrero de 2023
Tutores: Virginia Ahedo y Álvar Arnaiz

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

Dña. Virginia Ahedo, profesora del departamento de Ingeniería de Organización, área de Organización de Empresas.

Expone:

Que el alumno Dña. Gadea Lucas Pérez, con DNI 71483074V, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado "Impact Factor Oracle".

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 2 de febrero de 2023

Vº. Bº. del Tutor:	Vº. Bº. del co-tutor:
Dña. Virginia Ahedo	D. Álvar Arnaiz

Resumen

El presente proyecto se centra en el desarrollo de una aplicación web que utiliza técnicas de aprendizaje automático para predecir el factor de impacto de las revistas científicas. Esta métrica se utiliza para evaluar la importancia de una revista en un campo científico determinado. Se mide a través de la frecuencia con la que los artículos de la misma han sido citados en un año específico.

El proyecto utilizará los datos históricos disponibles en Google Scholar como inputs para los algoritmos de aprendizaje automático. Estos algoritmos serán supervisados y se utilizarán para estimar el valor del índice de impacto de las revistas indexadas en el JCR (Journal Citation Reports). El objetivo final es desarrollar una aplicación web accesible y de fácil uso para la comunidad científica, que permita predecir la importancia de las revistas científicas en tiempo real.

Este proyecto es relevante para la comunidad científica ya que el factor de impacto es un criterio importante en la evaluación de la calidad del trabajo científico y puede ser de gran ayuda en la selección de la revista adecuada para publicar un nuevo trabajo. Además, la aplicación web será de acceso abierto y se encontrará en un repositorio público para garantizar su disponibilidad y uso por parte de la comunidad científica.

Descriptores

Bibliometría, publicación de artículos, revistas científicas, índice de impacto.

Abstract

This project focuses on the development of a web application that uses machine learning techniques to predict the impact factor of scientific journals. The impact factor is a metric used to assess the importance of a journal in a given scientific field. It is measured through the frequency with which the articles of a journal have been cited in a specific year.

The project will use the historical data available in Google Scholar as inputs for the machine learning algorithms. These algorithms will be supervised and used to estimate the impact index value of the journals indexed in the JCR (Journal Citation Reports). The ultimate goal is to develop an accessible and easy-to-use web application for the scientific community, which allows predicting the importance of scientific journals in real time.

This project is relevant for the scientific community since the impact factor is an important criterion in the evaluation of the quality of scientific work and can be of great help in the selection of the appropriate journal to publish a new work. In addition, the web application will be open access and will be in a public repository to guarantee its availability and use by the scientific community.

Keywords

Bibliometrics, articles publication, scientific journals, impact index.

Índice general

Índice de figuras

Índice de tablas

Introducción

El factor de impacto de una publicación científica mide la frecuencia con la cual ha sido citado el artículo promedio de una revista en un año en particular. Específicamente, sirve para evaluar la importancia de una revista dentro de un determinado campo científico. Existen múltiples metodologías de cálculo y sus correspondientes métricas, siendo el JCR (Journal Citation Reports) y el SJR (Scimago Journal Rank) los dos índices de impacto más utilizados. Además, cabe destacar que el factor de impacto es uno de los principales criterios empleados en los procesos de acreditación y promoción interna para evaluar la calidad del trabajo científico de millones de académicos en todo el mundo. Por su propia naturaleza, el factor de impacto se calcula con carácter retrospectivo, i.e., sobre datos de años anteriores. Así pues, si a la hora de seleccionar la revista a la que mandarán un nuevo trabajo, los académicos quieren tener en cuenta el posible factor de impacto que tendrá la revista en el año de publicación del artículo, lo único que pueden hacer es fijarse en su índice de impacto de los años anteriores y hacer sus propias hipótesis/predicciones de futuro. Dado que a día de hoy la herramienta Google Scholar recoge información extremadamente actualizada sobre la publicación y citación de artículos científicos (podría decirse que se actualiza prácticamente en tiempo real), creemos que puede ser muy útil para la comunidad científica utilizar los datos de Google Scholar como *inputs* de modelos de aprendizaje automático para estimar el índice de impacto que tendrán las distintas revistas científicas en el año en curso. El *output* esperado del proyecto será una aplicación web de tipo open-access, la cual implementará algoritmos de aprendizaje supervisado que utilizarán los datos históricos disponibles en Google Scholar para predecir el valor del índice de impacto de todas las revistas científicas indexadas en el JCR. Dicha aplicación se dejará en un repositorio público, para así garantizar que pueda ser utilizada por toda la comunidad científica.

Objetivos del proyecto

A continuación, se enumeran los principales objetivos de este proyecto:

1. Lectura de literatura científica sobre bibliometría para comprender bien el marco conceptual en el que se encuadra este proyecto.
2. Estudio de la metodología de cálculo de los distintos índices de impacto en general, y del JCR en particular. Así mismo, se estudiarán las sucesivas modificaciones/excepciones que se han ido introduciendo en el cálculo del JCR a lo largo de los años.
3. Estudio de la API de Google Scholar y estructuración de una base de datos en la que se almacenará la información descargada de Google Scholar.
4. Implementación de las funciones de cálculo del índice JCR para aplicarlas sobre los datos extraídos de Google Scholar.
5. Implementación de distintos modelos de regresión (aprendizaje supervisado) para predecir el índice de impacto JCR a partir de las series temporales históricas disponibles. Selección del mejor algoritmo.
6. Inclusión del mejor modelo de regresión en una aplicación web para ponerlo a disposición de la comunidad científica.

Conceptos teóricos

Cuando un investigador termina una etapa de su investigación, genera un artículo científico en el que plasma el resultado de su trabajo. Esta documentación sirve de precedente para aquellos que posteriormente investiguen sobre temáticas relacionadas. Así pues, el artículo científico es el elemento principal en torno al que giran los estudios bibliométricos.

Antes de adentrarnos en más detalles, se ilustrarán algunos conceptos esenciales sobre este proyecto, para dotarlo de mayor comprensión y claridad. En esta sección, por tanto, se abarcarán los conceptos de bibliometría, cienciometría e índice de impacto, como conceptos relativos al avance de la ciencia y la producción de conocimientos a partir de la actividad de la investigación.

3.1. Bibliometría

Palabra que proviene del griego *biblio* (libro) y *-metría* (medición)

Ciencia que aplica métodos matemáticos para encontrar comportamientos estadísticos en la literatura científica. Estos estudios y análisis pretenden cuantificar toda la actividad científica escrita con el objetivo final de orientar sobre el impacto de una investigación [6].

Antiguamente, se evaluaba la producción científica por pares, que es un proceso en el que expertos en un campo revisan y evalúan la calidad y el mérito de una investigación antes de su publicación. Al final de la década de 1950, en un momento en que la producción científica aumentaba progresivamente a un ritmo sostenido, nació la idea de una evaluación basada en cantidades, ciertamente menos costosa que la evaluación de pares: desde

ese momento, comenzó a hablarse de bibliometría [9]. Así, los estudios bibliométricos tienen su origen en la década de 1960 con la creación del Science Citation Index por Eugene Garfield y el análisis de redes de citas realizado por Derek John de Solla Price. Estos trabajos establecieron las bases sólidas de la bibliometría como disciplina.

El objetivo de los estudios bibliométricos pueden limitase solamente al análisis de la envergadura, el crecimiento y la distribución de la literatura científica, pero también son útiles a la hora de encontrar las revistas donde es más conveniente publicar un artículo o descubrir a los autores más importantes en cada ámbito, así como las nuevas tendencias.

Sin embargo, la información organizada de esta forma también tiene sus **limitaciones**. Ejemplo de ello es el patrón de citas usado en cada área de investigación; es mucho más común incluir citas en investigaciones tecnológicas y científicas. Es decir, cada temática deberá ser tratada de forma distinta. Para ello se normalizará el índice de impacto. Por otro lado, también influye la base de datos utilizada, ya que cada una tiene un método de indexación distinta. El idioma es otro ejemplo de limitación. Ya que el inglés es el idioma predominante, será más complicado encontrar citas de documentos escritos en otros idiomas. También es preciso tener en cuenta problemas de dispersión debidos a perfiles duplicados o nombres similares. Para poder solucionar este tipo de confusiones, es recomendable crear identificadores (por ejemplo, el ORCID) para cada autor. De esta forma, todos los comportamientos irregulares (como las citas a uno mismo) pueden tenerse en consideración. Finalmente, el momento en que se realizan los estudios es también fundamental, ya que el nivel de conocimientos en torno a una materia determinada varía con el tiempo.

Ejemplo

Un ejemplo de análisis bibliométrico (tomado del libro "Documentación científica y nuevas tecnologías de la información"[6]) podría involucrar el determinar los autores que abordan un tema específico en una revista en particular. Para ello, deberían registrarse los siguientes datos:

1. Lista de identificadores de los autores (ordenados alfabéticamente) aparecidos en la revista en cuestión.

2. Cantidad de artículos de cada uno de estos autores en la revista.

3. Revistas en las que aparecen los artículos citados.

4. Los artículos (cantidad) que aparecen en cada una de las revistas listadas.

3.2. Cienciometría

Palabra que proviene del latín *scientia* (ciencia) y del griego *-metría* (medición)

Se trata del estudio de los metadatos e indicadores sobre la bibliografía científica con el fin de medir y analizar toda la producción científica [6].

La cienciometría tiene como objetivo clasificar y organizar el conocimiento científico a través de la creación de sistemas de conceptos y facilitar su transferencia a través de la educación y formación. Además, promueve la comunicación de conocimientos de un idioma a otro mediante la utilización de símbolos lingüísticos, permite la síntesis y resumen de la información científica, y proporciona un medio para recuperar y almacenar información a través de la indexación y lenguajes documentales.

La cienciometría, puesta en práctica, trabaja relacionada con otras ciencias y disciplinas como las que se muestran en la siguiente imagen:

Figura 3.1: Ambos campos se superponen

Es decir, se trata de un concepto más amplio que engloba al anterior (bibliometría). Así, se podría decir que la bibliometría surge como resultado del contacto interdisciplinar de entre el conjunto de disciplinas que integran lo que se conoce como «ciencia de la ciencia», cuya fuente es la propia cienciometría [6].

Sin embargo, hoy en día la frontera entre cienciometría y bibliometría ha desaparecido casi por completo y ambos términos se usan prácticamente de forma sinónima [10].

3.3. Webmetría

Palabra que proviene del inglés web y del griego -metría (medición)

Aparte de las citaciones tradicionales, existen también referencias generadas por los lectores en la web. Se trata de la «métrica de la web» (*webmetrics* o *cybermetrics*). Se define como el estudio de los aspectos cuantitativos de la construcción y uso de recursos de información, estructuras y tecnologías en la web, basándose en enfoques bibliométricos [2].

El objetivo es obtener información sobre el número y el tipo de hiperenlaces, la estructura del World Wide Web y los modelos de uso de los recursos. Así pues, se contabiliza el número de veces que un sitio web o un documento publicado en internet es accedido y se divide entre el número de páginas del sitio mismo [9]. Esto nos permite calcular la frecuencia con la que una página web promedio ha sido enlazada en un momento dado; un alto factor de impacto de la web indica la popularidad y, probablemente, el prestigio de una página web. Debido al creciente número de documentos publicados y disponibles en la red, en especial los *e-journals* de acceso abierto, se han desarrollado nuevas herramientas de medición, lo que ha llevado a algunos investigadores a considerar un posible paralelismo entre las citaciones tradicionales y las referencias en la web. [9].

3.4. Índice de impacto

Existe un gran abanico de índices distintos englobados dentro de la bibliometría. Sin embargo, en este caso, nos centraremos especialmente en los índices de citas o, en inglés, “Citation Index”.

Se trata de índices de autor con características especiales, pues no solamente citan junto a cada autor la lista de los documentos por él publicados, sino que añade, en cada referencia, la lista de los documentos que ha citado esta referencia en su propia bibliografía. Permiten localizar otros autores que han tratado las mismas materias y buscar también documentos más recientes que éste ya conocido. [6]. Contienen literatura científica de medicina, psicología, agricultura, tecnología y documentación científica en general.

A continuación, se enumeran algunos de los índices de impacto más comunes.

SJR

SJR (SCImago Journal Rank) es un indicador de calidad de revistas científicas basado en la idea de que *cuanto más citada es una revista, mayor importancia tiene.*

Además, se entiende que *no todas las citas son iguales*[1]. Es decir, el indicador se calcula utilizando un algoritmo que también tiene en cuenta la importancia de la revista que hace la cita, de tal manera que una cita hecha por una revista de alto impacto tendrá más peso que una cita de una revista de bajo impacto [1].

Así pues, se calculan las citas **ponderadas** de los últimos tres años para determinar el SJR de cada revista.

Hindex

Hindex es un indicador de productividad y visibilidad de un investigador. El Hindex se calcula a partir de la cantidad de artículos publicados por un autor y la cantidad de veces que estos artículos han sido citados.

El Hindex se determina buscando el número de publicaciones con al menos ese número de citas. Por ejemplo, un investigador con un Hindex de 10 tiene al menos 10 publicaciones que han sido citadas al menos 10 veces cada una [1].

Por lo tanto, se podría decir que este indicador es una métrica a nivel de autor se utiliza para medir la importancia, la productividad y el impacto de un investigador en la comunidad científica [1]. Este índice también puede ser aplicado a un grupo de investigadores (por ejemplo, un departamento, universidad o país).

Es importante mencionar que Google Scholar utiliza métricas basadas en el Hindex.

CiteScore

Es un indicador desarrollado por Scopus que tiene en cuenta el número de citas recibidas por una revista en un período de tres años, el número de documentos en la base de datos de Scopus y el número de documentos publicados en esa revista en ese período.

De esta forma, el CiteScore se calcula dividiendo el número de citas que recibe una revista en un año por los documentos publicados en los tres años previos, y dividiendo este número por el número de documentos indexados en Scopus publicados en los mismos tres años [1].

JCR

El JCR (*Journal Citation Reports*) es una base de datos desarrollada por Clarivate Analytics que proporciona información sobre revistas científicas.

Consiste en el análisis biométrico de las revistas del banco de datos ISI (*Intellect Scientific Information*), con número de citas al año de cada una, revistas que citan a otras revistas, listado de abreviaturas de títulos y su desarrollo, etc. JCR proporciona información estadística sobre la frecuencia de citas de las revistas, incluyendo el *Impact Factor* (IF) y el *Impact Factor of a Year* (IFY) entre otros indicadores [7].

De entre los datos estadísticos que se obtienen, nos importa especialmente el ya mencionado **Factor de Impacto**, que permiten determinar de una manera sistemática y objetiva la importancia relativa de las principales revistas de investigación internacionales dentro de sus categorías temáticas.

JCR cubre más de 12 000 revistas de más de 80 disciplinas diferentes, lo que permite comparar el impacto y la calidad de las revistas en diferentes campos. Los datos de JCR se utilizan a menudo como un indicador de la calidad y el impacto de una revista en la comunidad científica [7].

La última actualización del JCR ofrece los datos del Factor de Impacto del 2022. Este índice se calcula con un cierto retraso respecto al final del año: suele aparecer alrededor de los meses de mayo o junio del año siguiente. Esto limita el acceso a la información actualizada y justifica la importancia del trabajo que está siendo realizado, de forma que se pueda brindar información actualizada y accesible sobre el impacto de los artículos y revistas a la comunidad científica, lo cual es esencial para la toma de decisiones y la evaluación del desempeño científico.

Técnicas y herramientas

En esta sección se presentarán las distintas herramientas y recursos que se han utilizado para la realización del proyecto.

4.1. Técnicas

En el presente proyecto, se han utilizado diversas técnicas para llevar a cabo el análisis y cálculo del índice de impacto de publicaciones científicas. Estas técnicas han sido seleccionadas con el objetivo de obtener resultados precisos y confiables, y de cubrir las necesidades específicas de este estudio.

Web-Scraping

Es la práctica de recopilar datos a través de un programa que interactúe con una API [5]. Más concretamente, un programa automatizado compuesto por *queries* que realizan solicitudes HTTP para adquirir recursos de un sitio web específico. Esta solicitud se puede formatear en una URL que contenga una consulta GET o en un mensaje HTTP que contenga una consulta POST [11]. Una vez que la petición es exitosamente recibida y manejada por el sitio web seleccionado, el recurso requerido será extraído y luego devuelto al programa de *web-scraping* específico.

El uso *web-scraping* resulta ser una técnica efectiva en este proyecto debido a que permite obtener datos de Google Scholar de manera automatizada y eficiente. La información obtenida a través de esta técnica es esencial para calcular el índice de impacto (JCR), ya que Google Scholar es una de las principales fuentes de datos para su cálculo. Además, el *web-scraping* permite obtener grandes cantidades de información en un corto período de

tiempo, lo que resulta muy útil en proyectos en el que se requiere una gran cantidad de datos.

4.2. Lenguaje de programación

En primer lugar, nos planteamos la cuestión del lenguaje o lenguajes de programación más adecuados para nuestro objetivo. Las listas de popularidad actuales nos muestran dos entornos ganadores para proyectos de *web-scraping*: Python y Javascript.

Aunque ambos lenguajes son altamente capaces para nuestro proyecto, la enorme base de conocimientos y la diversidad de herramientas creadas en el universo **Python** decanta la balanza hacia ese lado. Quizás JavaScript permita mejores resultados usando la gestión de memoria en *requests* simultáneas, pero a costa de un código más oscuro y difícil de mantener. Aunque JavaScript cuenta con un gran repertorio de paquetes Node.JS como utilidades de *web-scraping*, en el entorno Python es difícil imaginar una tarea para la que no se haya escrito una (o más) herramientas que resuelvan eficazmente nuestro problema.

Por otro lado, la comunidad de programadores de Python es inmensa y su creciente popularidad facilita el hallazgo de soluciones rápidamente, tanto en los foros como en la extensa documentación con la que cuenta. Aunque encontramos un rendimiento ligeramente inferior a otros lenguajes en ciertas búsquedas, es el precio a pagar por el tipado dinámico.

Como valor añadido, Python es fácil de mantener cuando necesitamos adaptar nuestro código a las cambiantes estructuras de las páginas web. Además, sus reconocidas herramientas de análisis de datos nos permiten continuar en el mismo entorno, sin necesidad de buscar alternativas para afrontar tareas relacionadas con la *data science*.

4.3. Bibliotecas

A lo largo del proyecto se ha recurrido a diversas bibliotecas de **Python**. A continuación, se presenta brevemente cada una de ellas.

Scholarly

Se trata de una biblioteca de Python que permite acceder a los datos de Google Scholar de manera fácil y rápida. La biblioteca proporciona una

interfaz sencilla para buscar y recuperar información sobre artículos, autores y revistas en Google Scholar, incluyendo metadatos, citas y otra información relacionada.

Sin embargo, esta biblioteca ha terminado siendo descartada para este proyecto. A diferencia de otras técnicas de *web-scraping*, Scholarly no está diseñada para extraer grandes cantidades de datos de Google Scholar. La biblioteca tiene una serie de limitaciones en cuanto a la cantidad de datos que se pueden recolectar, ya que está diseñada para ser utilizada en investigaciones científicas y no para la extracción masiva de datos. Además, Scholarly está diseñada para respetar los términos de servicio de Google Scholar y no violar la política de uso de la plataforma, por lo que no se recomienda su uso para recolectar grandes cantidades de datos.

En resumen Scholarly es una herramienta útil para acceder a información científica y académica de manera rápida y sencilla, pero no esta diseñada para extraer grandes cantidades de datos.

Beautiful Soup

Beautiful Soup es una biblioteca de Python extremadamente útil para la extracción de datos de páginas HTML o XML. Actualmente se encuentra en su versión 4.8.1. En el siguiente enlace se puede acceder a la documentación de su página oficial.

Esta biblioteca nos proporciona numerosos módulos para navegar a través de las páginas web y para extraer fácilmente su contenido. Puesto que gran parte del proyecto se basa en el uso de *web-scraping*, se ha hecho uso intensivo de la misma para extraer los principales datos de los artículos científicos que posteriormente conformarán la BBDD del proyecto.

De Beautiful Soup hay que destacar su facilidad de uso y la amplia documentación que aporta. En su página web nos aconseja utilizar el analizador *lxml*, que proporciona al entorno Python la disponibilidad de las bibliotecas *libxml2* y *libxslt*. Allí mismo, se anima incluso a utilizar aisladamente este parser cuando el tiempo de respuesta sea una cuestión crítica. En nuestro caso, las facilidades que proporciona Beautiful Soup justifican ampliamente su uso, aunque la rapidez de resultados no iguale la de la utilización aislada de los analizadores sobre los que trabaja.

Habanero

Para poder hacer uso de la **API de Crossref**, se han explorado diversas alternativas. De entre todas las bibliotecas de Python se ha seleccionado Habanero, ya que es una biblioteca muy fácil de usar y está en constante actualización.

Esta biblioteca está diseñada para facilitar el acceso a las bases de datos de revistas científicas y a otras relacionadas con el ámbito académico. Ofrece una interfaz simple para recuperar información utilizando los protocolos y las API de diferentes bases de datos, incluyendo JSTOR, Unpaywall, Crossref, DataCite, etc. Además, Habanero es compatible con las normas de Open Access, lo que permite a los usuarios acceder a contenido científico gratuito y de libre acceso.

4.4. Bases de datos

Antes de su compra por Oracle Corporation (2010), MySQL era la aplicación de base de datos más popular de código abierto para la programación web. En el momento actual, el software libre nos ofrece dos soluciones ampliamente contrastadas en gestores de bases de datos relacionales: MariaDB y PostgreSQL. Necesitaremos una herramienta de este tipo para organizar los datos recolectados. La estructura tabular de la información nos permite aplicar sobre ella el lenguaje de interrogación SQL.

En realidad, MariaDB es una bifurcación de MySQL nacida para garantizar la supervivencia del proyecto como código abierto. Hoy en día, MariaDB es altamente compatible con MySQL e incluso superior en sus últimas versiones (10.1.1), ya que la comunidad ha ido añadiendo nuevas características al proyecto original. Por otro lado, aunque no está tan extendido como MySQL, PostgreSQL es posiblemente el gestor de bases de datos de código abierto más sólido y potente a día de hoy.

PostgreSQL

Finalmente se opta por PostgreSQL (también denominado Postgres). Pese a que ambas opciones son muy similares, se ha elegido esta última debido a que ya se ha trabajado con ella anteriormente. Además, PostgreSQL posee una sólida reputación por su arquitectura comprobada, confiabilidad, integridad de datos, conjunto sólido de características, extensibilidad y la dedicación de la comunidad de código abierto detrás del software para ofrecer soluciones innovadoras y de rendimiento constante [3].

MariaDB	PostgreSQL
No totalmente compatible con SQL	Compatible con SQL estándar
Soporte para tipos de datos estándar SQL	Soporta además tipos avanzados
Tipos de datos flexibles	Tipos de datos estrictos
Tamaño pequeño de base de datos	Tamaño grande de base de datos
Sin soporte directo para JSON	Soporte directo de JSON
No índices parciales	Índices parciales
No soporte para web dinámica	Soporta sitios web dinámicos

Tabla 4.1: Comparativa entre MariaDB y PostgreSQL

Al igual que MariaDB, es un sistema de gestión de bases de datos relacional de código abierto. Así pues, está dirigido y desarrollado por una comunidad altruista de desarrolladores (PostgreSQL Global Development Group). PostgreSQL utiliza y amplía el lenguaje SQL combinado con muchas características que almacenan y escalan de forma segura las cargas de trabajo de datos más complicadas [3]. Su última versión y la usada para el proyecto es la versión 15.1 lanzada el 10 de noviembre de este año.

4.5. APIs

A lo largo del proyecto se ha recurrido a varias APIs diferentes, todas ellas de acceso gratuito.

Google Scholar

Se trata de una de las principales herramientas que ofrece Google a los investigadores. Google Scholar es, fundamentalmente, un buscador de contenido y bibliografía científica que permite localizar artículos de revistas especializadas ordenados por relevancia en función de las palabras clave introducidas en el buscador. También se puede filtrar la información en función de su fecha de publicación, idioma o número de citas.

Crossref

Crossref es una herramienta que facilita el acceso a la información de los artículos científicos a partir de su DOI[1].

Crossref es una organización sin fines de lucro que pretende facilitar las conexiones académicas.

Como ya se ha mencionado previamente, en base a la información de los DOIs la agencia de registro CrossRef es capaz de proporcionarnos aplicaciones útiles para hacer nuestro flujo de investigación más sencillo. Se trata de una asociación sin ánimo de lucro de editoriales científicas que no solo facilita el registro de DOIs a las editoriales, sino que también ofrece servicios y aplicaciones para el personal investigador que tienen como base estos códigos.

Los DOIs que CrossRef almacena van acompañados de información que refleja las cualidades básicas de una publicación científica. Me estoy refiriendo a datos como títulos, abstracts, palabras clave, autores...

CrossRef Metadata Search hace posible obtener toda esta información al instante con tan solo proporcionar el DOI asociado a una publicación, o al contrario, obtener el DOI de la publicación con tan solo introducir algunos de estos datos en su buscador.

[1] El DOI (*Digital Object Identifier*) es el acrónimo con el que se conoce al identificador inequívoco de un artículo científico. Se trata de un enlace permanente al contenido electrónico de dicho artículo. Por lo general, un DOI tiene forma de código alfanumérico.

Aspectos relevantes del desarrollo del proyecto

Trabajos relacionados

En el campo de la evaluación de revistas científicas, existen varios trabajos y proyectos previos que han tratado de extraer datos y calcular el factor de impacto. A continuación se presenta una breve descripción de algunos de estos trabajos relacionados:

1. **An index to quantify an individual's scientific research output**[4]: Este estudio realizado por Jorge E. Hirsch en el año 2005, presenta el h-index como una nueva métrica para evaluar el impacto de la investigación científica. El estudio argumenta que el h-index es más preciso que el factor de impacto y es menos sujeto a manipulación.

2. **Modeling and Prediction of the Impact Factor of Journals Using Open-Access Databases**[8]: Este estudio de 2020 se enfoca en el uso de bases de datos de acceso abierto para modelar y predecir el factor de impacto de las revistas científicas. El estudio sugiere que es posible utilizar bases de datos como Google Scholar, ResearchGate y Scopus para estimar el factor de impacto de revistas nuevas, pequeñas o independientes que no están incluidas en el Science Citation Index (SCI) y que no reciben un factor de impacto de la base de datos Web of Science (WoS). El estudio también señala que los resultados obtenidos con el modelo desarrollado sugieren que es posible predecir el factor de impacto WoS utilizando bases de datos de acceso abierto alternativas.

Conclusiones y Líneas de trabajo futuras

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

Prototipo inicial

Al revisar la viabilidad del proyecto se plantearon posibles dificultades que podían surgir. Tras comprender que las complicaciones eran numerosas, se decidió crear un **prototipo sencillo**, consistente en un *script* en Python, que trataría de lanzar mil peticiones de búsqueda. Esto nos permitiría establecer los límites de realizar "web scrapping" sobre Google Scholar .

Así pues, manos a la obra, se desarrolló un *script* sencillo, que solicita acceso a la página principal de Google Scholar y, mediante métodos HTTP, realiza una búsqueda (a partir de parámetros solicitados por pantalla). Finalmente, extrae el título de la página resultante tras hacer la búsqueda.

Todo esto se logra haciendo uso de la biblioteca de Python **BeautifulSoup** (ver documentación en el siguiente enlace). Esta biblioteca contiene métodos centrados en la extracción de datos de archivos HTML y XML para su posterior análisis. Es fácil advertir cuán idónea resulta esta biblioteca para nuestro propósito.

Tras diseñar y programar el código mencionado, se procede a su prueba. La primera ejecución del *script* resultó desaletadora, ya que, a partir de la solicitud número **726**, Google Scholar detecta que un ***bot*** está realizando búsquedas. A partir de ese momento, nuestra ip queda bloqueada y las solicitudes fallan sin excepción. Google Schoolar nos redirecciona a una página (figura 8.1) donde se solicita al usuario resolver un *captcha*.

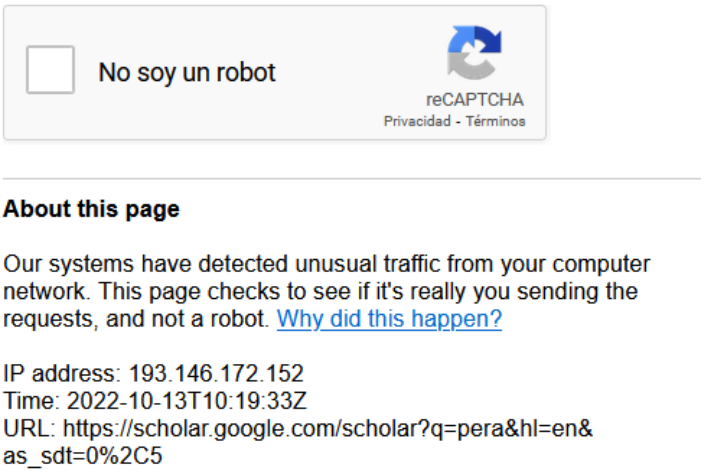

Figura 8.1: Captura de reCAPTCHA de Google

El número de solicitudes exitosas es demasiado bajo para cumplir su función en nuestro proyecto, por lo que se procede a buscar una solución alternativa.

Tras distintas pruebas, se encontró una forma de superar la barrera del *captcha*. A saber: añadiendo a la *url* una sección de text extra que permite suprimir esta excepción durante un periodo concreto de tiempo. Así pues, logramos ejecutar con éxito el *script* tantas veces como fuese necesario. Sin embargo, esta solución tampoco es válida a largo plazo. Si se intenta ejecutar de nuevo el programa, en otra sesión, la dirección vuelve a ser inválida. Además, es un remedio poco práctico, ya que se debe concatenar distintos parámetros y cadenas de texto que, dependiendo del momento y el contexto en que se ejecute, pueden no servir.

Como la propuesta anterior no fue satisfactoria, se siguió buscando opciones. La siguiente propuesta consistía en usar un agente de usuario distinto para cada solicitud

agente de usuario es cualquier software, que actúa en nombre de un usuario , que recupera, presenta y facilita la interacción del usuario final con el contenido web". Por lo tanto, un agente de usuario es un tipo especial de agente de software . Algunos ejemplos destacados de agentes de usuario son los navegadores web La cadena User-Agent es uno de los criterios por los cuales los rastreadores web pueden ser excluidos del acceso a ciertas

partes de un sitio web utilizando el Estándar de exclusión de robots (archivo robots.txt).

usar un *proxy* para "enmascarar" nuestra ip. La biblioteca **Request** de Python ofrece métodos para lograrlo. Dicho esto, se implementó un método que extrae *proxies* de listas públicas y gratuitas de Internet (v.g.: proxyscrape.com). Se prueba la ejecución del prototipo una vez más y, finalmente, funciona sin inconvenientes. Ahora ya se puede decir que el proyecto es **viable**.

...

- Incluir doi de crossref y resultados (obteniendo unos 170 artículos en 15 minutos)

Estos resultados no son eficientes, por lo que se tratará de optimizar el prototipo siguiendo dos pautas. Primero, se tratará de extraer las llamadas a Crossref de forma que solo se tenga que realizar una única llamada una vez que se han extraído el resto de detalles sobre los artículos. La segunda pauta es emplear programación concurrente para ejecutar varios hilos al mismo tiempo.

Para poder realizar ambas cosas, se separarán las peticiones a Google Schoolar y las peticiones a Crossref en dos hilos distintos. Para ello, será necesario recurrir a la librería de Python **multiprocessing**.

Tras actualizar estos cambios, el prototipo obtiene resultados mucho más óptimos: artículos en 15 minutos.

Bibliografía

[1] Svetla Baykoucheva. *Driving science information discovery in the digital age*. Chandos information professional series. Chandos Publishing, Cambridge, Massachusetts, 2022.

[2] Lennart Björneborn et al. *Small-world link structures across an academic web space: a library and information science approach*. Citeseer, 2004.

[3] The PostgreSQL Global Development Group. Postgresql: The world's most advanced open source database, 2019. [Internet; acceso 05-diciembre-2022].

[4] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.

[5] Ryan Mitchell. *Web scraping with Python: Collecting more data from the modern web*. .º'Reilly Media, Inc.", 2018.

[6] Nuria Amat Noguera. *Documentación Científica y Nuevas Tecnologías de la Información*. DICIONES PIRÁMIDE S.A., Josefa Valcárcel, 27. 28027 Madrid, 1989.

[7] Web of Science Group. Journal citation reports, 2019. [Internet; acceso 02-enero-2023].

[8] Matthias Templ. Modeling and prediction of the impact factor of journals using open-access databases: With an application to the austrian journal of statistics. *Austrian Journal of Statistics*, 49(5):35–58, 2020.

[9] Simona Turbanti. *Bibliometria e scienze del libro : internazionalizzazione e vitalità degli studi Italiani*. Studi e saggi ; 170. Firenze University Press, Italy, 2017.

[10] Nikolay K. Vitanov. *Science Dynamics and Research Production Indicators, Indexes, Statistical Laws and Mathematical Models*. Qualitative and Quantitative Analysis of Scientific and Scholarly Communication. Springer International Publishing, Cham, 1st ed. 2016. edition, 2016.

[11] Bo Zhao. Web scraping. *Encyclopedia of big data*, pages 1–3, 2017.