



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 18 de enero 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 Organización, área de (preguntar).

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, 18 de enero de 2023

Vº. Bº. del Tutor:

Vº. Bº. del co-tutor:

Dña. Virginia Ahedo

D. Álvar Arnaiz





## Resumen

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.

## Descriptores

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

## **Abstract**

The impact factor of a scientific publication measures the frequency with which the average article in a journal has been cited in a particular year. Specifically, it serves to evaluate the importance of a journal within a certain scientific field. There are multiple calculation methodologies and their corresponding metrics, with the JCR (Journal Citation Reports) and the SJR (Scimago Journal Rank) being the two most widely used impact index. In addition, it should be noted that the impact factor is one of the main criteria used in the processes of accreditation and internal promotion to assess the quality of the scientific work of millions of academics around the world. By its very nature, the impact factor is calculated retrospectively, i.e., based on data from previous years. Thus, if academics want to take into account the possible impact factor that the journal will have in the year of publication of the article when selecting the journal to which they will send a new article, all they can do is look at your impact index from previous years and make your own assumptions or predictions for the future. Nowadays the Google Scholar tool collects extremely up-to-date information on the publication and citation of scientific articles (it could be said that it is updated practically in real time), so we believe that it can be very useful for the scientific community to use Google Scholar data as inputs of machine learning models to estimate the impact index that the different scientific journals will have in the current year. The expected output of the project will be an open-access type web application, which will implement supervised learning algorithms that will use the historical data available in Google Scholar to predict the value of the impact index of all the scientific journals indexed in the JCR. This application will be left in a public repository, in order to guarantee its use for the entire scientific community.

## **Keywords**

Bibliometrics, articles publication, scientific journals, impact index.



---

# Índice general

---

Índice general	iii
Índice de figuras	v
Índice de tablas	vi
Introducción	1
Objetivos del proyecto	3
Conceptos teóricos	5
3.1. Bibliometría . . . . .	5
3.2. Cienciometría . . . . .	6
3.3. Índice de impacto . . . . .	7
3.4. JCR . . . . .	9
Técnicas y herramientas	11
4.1. Lenguaje de programación . . . . .	11
4.2. Bibliotecas . . . . .	12
4.3. Bases de datos . . . . .	13
4.4. APIs . . . . .	15
Aspectos relevantes del desarrollo del proyecto	17
Trabajos relacionados	19
Conclusiones y Líneas de trabajo futuras	21

<b>Prototipo inicial</b>	<b>23</b>
<b>Bibliografía</b>	<b>27</b>

---

## Índice de figuras

---

4.1. Icono de MariaDB . . . . .	14
4.2. Icono de Postgres . . . . .	14
4.3. Icono de Google Scholar . . . . .	15
4.4. Icono de Crossref . . . . .	16
8.1. Captura de reCAPTCHA de Google . . . . .	24

---

# Índice de tablas

---

4.1. Comparativa entre MariaDB y PostgreSQL . . . . .	14
---	----

---

# Introducción

---

Descripción del contenido del trabajo y del estructura de la memoria y del resto de materiales entregados.



---

## Objetivos del proyecto

---

Este apartado explica de forma precisa y concisa cuales son los objetivos que se persiguen con la realización del proyecto. Se puede distinguir entre los objetivos marcados por los requisitos del software a construir y los objetivos de carácter técnico que plantea a la hora de llevar a la práctica el proyecto.





---

## Conceptos teóricos

---

Cuando un investigador termina su trabajo, genera un artículo científico en el que plasma el resultado de su investigación. 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. Así pues, en esta sección se abarcarán los conceptos de bibliometría, cienciometría e índice de impacto; conceptos relativos al avance de la ciencia y la producción de conocimientos a partir de la actividad de la investigación.

### 3.1. Bibliometría

Biblio (libros) - 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.

Los estudios bibliométricos tienen origen en 1960 con la aparición del Science Citation Index (Eugene Garfield) y el análisis de redes de citas (Derek John de Solla Price), que sentaron las bases fundamentales sobre bibliometría.

El objetivo de los estudios bibliométricos pueden limitarse 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 (ORCID) para cada autor. De esta forma, todos los comportamientos irregulares (como las citas a uno mismo) deben tenerse en consideración. Por otro lado, el momento en que se realizan los estudios es también fundamental, ya que los datos pueden cambiar de un momento a otro.

## Ejemplo

Un ejemplo de análisis bibliométrico podría consistir en averiguar los autores que, en cierta revista, traten un tema en concreto. Para ello se deberían anotar 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

Ciencio (ciencia) - metría (medición)

Estudio de los metadatos e indicadores sobre la bibliografía científica con el fin de medir y analizar toda la producción científica. La cienciometría es útil para ordenar conocimientos científicos mediante la creación de sistemas

de conceptos. También para transferir conocimientos a través de la enseñanza y la formación. A su vez, también busca comunicar conocimientos de un lenguaje a otro a través de símbolos lingüísticos. Logra resumir y sintetizar la información científica y, por último, recupera y almacena información mediante la indexación y lenguajes documentales.

La cienciometría, puesta en práctica, trabaja relacionada con otras ciencias y disciplinas como las que se muestran en la figura 1 (insertar imagen en la plantilla del TFG).

(Idea aproximada de la imagen) Figura 1: Ambos campos se superponen en gran medida

Es decir, se trata de un concepto más amplio que engloba al anterior (bibliometría). 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.

### 3.3. 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. Contienen literatura científica de medicina, psicología, agricultura, tecnología y documentación científica en general.

#### Ejemplo

En un índice de citas, aparecen los artículos de un mismo autor agrupados por orden cronológico. Justo debajo, aparece la lista de autores que han citado este artículo y, a su lado, la referencia del documento (ver figura 2).

(Insertar imagen en plantilla del TFG)

Es decir, bastará con saber el nombre de un solo autor que trabaje en el campo en el que se esté interesado, para obtener de forma inmediata

(consultando simplemente el índice) el conjunto de documentos que traten del mismo tema.

Generalmente se divide en dos subíndices: uno de autores (con referencias de documentos citados), y otro de fuentes (acompañadas de los títulos de los documentos).

Otros intereses que se deducen de este índice de autores consiste en el valor que pueden tener los documentos citados debido a que el autor de un documento, especialista en un campo, a tenido interés en dar a conocer a dichos autores.

## **SJR**

SJR (SCImago Journal Rank) es un indicador de calidad de revistas científicas basado en la idea de que "más citadas son mejores..<sup>EI</sup> 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 de una revista de alto impacto tendrá más peso que una cita de una revista de bajo impacto. Este indicador se utiliza para medir la calidad y el impacto de una revista en la comunidad científica.

## **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 investigador 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. Este indicador se utiliza para medir la importancia y el impacto de un investigador en la comunidad científica.

## **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.

### g-index

Es un indicador que se utiliza para medir la productividad y la impacto de un investigador. El g-index se calcula ordenando los artículos de un investigador por el número de citas y luego identificando el número más grande "g" tal que al menos "g" de los artículos han sido citados al menos "g" veces.

## 3.4. 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, 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.

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.

La última actualización del JCR ofrece los datos del Factor de Impacto del 2022.



---

# Técnicas y herramientas

---

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

## 4.1. Lenguaje de programación

En primer lugar, nos planteamos la cuestión del lenguaje o lenguajes de programación más eficaces para nuestro objetivo. Las listas de popularidad actuales nos muestran dos entornos ganadores para proyectos de *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 *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*.

## 4.2. 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 librería de Python que permite acceder a los datos de Google Scholar de manera fácil y rápida. La librería 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 está 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 scrapping*, 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 librería muy fácil de usar y esta 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.

## 4.3. 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 y 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.

	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



Figura 4.1: Icono de MariaDB

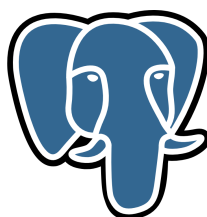


Figura 4.2: Icono de Postgres

## PostgreSQL

Finalmente se opta por **PostgreSQL**. 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 [1].

También se denomina Postgres. 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 [1]. Su última versión y la usada para el proyecto es la versión 15.1 lanzada el 10 de noviembre de este año.

## 4.4. 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.

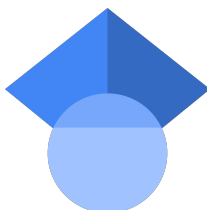


Figura 4.3: Icono de Google Scholar

### Crossref

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

#### NOTA

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.

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



Figura 4.4: Icono de Crossref

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.

---

## Aspectos relevantes del desarrollo del proyecto

---

Este apartado pretende recoger los aspectos más interesantes del desarrollo del proyecto, comentados por los autores del mismo. Debe incluir desde la exposición del ciclo de vida utilizado, hasta los detalles de mayor relevancia de las fases de análisis, diseño e implementación. Se busca que no sea una mera operación de copiar y pegar diagramas y extractos del código fuente, sino que realmente se justifiquen los caminos de solución que se han tomado, especialmente aquellos que no sean triviales. Puede ser el lugar más adecuado para documentar los aspectos más interesantes del diseño y de la implementación, con un mayor hincapié en aspectos tales como el tipo de arquitectura elegido, los índices de las tablas de la base de datos, normalización y desnormalización, distribución en ficheros<sup>3</sup>, reglas de negocio dentro de las bases de datos (EDVHV GH GDWRV DFWLYDV), aspectos de desarrollo relacionados con el WWW... Este apartado, debe convertirse en el resumen de la experiencia práctica del proyecto, y por sí mismo justifica que la memoria se convierta en un documento útil, fuente de referencia para los autores, los tutores y futuros alumnos.



---

## 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: concreto. Para ello se deberían anotar los siguientes datos:

1. **The Journal Impact Factor: A Valid and Reliable Indicator of Journal Quality?:** Este estudio realizado por Bo-Christer Björk y Torgny Roxå en el año 2008, analiza la validez y la fiabilidad del factor de impacto como indicador de la calidad de una revista. El estudio concluye que el factor de impacto es un indicador válido y fiable, pero que debe ser utilizado con precaución y en combinación con otros indicadores.
2. **A New Metric for Assessing the Impact of Scientific Research: The h-index:** 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.
3. **Modeling and Prediction of the Impact Factor of Journals Using Open-Access Databases:** Este estudio de 2016 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.

En el proyecto en curso, se busca continuar en esta línea de investigación con el objetivo de desarrollar una aplicación web de tipo open-access que implemente algoritmos de aprendizaje supervisado para predecir el valor del índice de impacto de todas las revistas científicas indexadas en el JCR. La aplicación se basará en los datos históricos disponibles en Google Scholar para entrenar los algoritmos y proporcionar una herramienta útil para la comunidad científica.



---

## **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

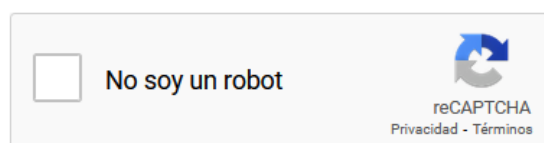
---

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 desalentadora, 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 Scholar nos redirige a una página (figura 8.1) donde se solicita al usuario resolver un *captcha*.



#### About this page

Our systems have detected unusual traffic from your computer network. This page checks to see if it's really you sending the requests, and not a robot. [Why did this happen?](#)

IP address: 193.146.172.152

Time: 2022-10-13T10:19:33Z

URL: [https://scholar.google.com/scholar?q=pera&hl=en&as\\_sdt=0%2C5](https://scholar.google.com/scholar?q=pera&hl=en&as_sdt=0%2C5)

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.: [proxyscraper.com](https://proxyscraper.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**.



---

## Bibliografía

---

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