Bibliometric analysis of scientific literature on deep learning

Produced by: Walid Benchabekh

Under the supervision of: Professor Caroline Petitjean

Table of contents

Introduction

Since the rise of deep learning a few years ago, the scientific literature in the fields of data science, image processing, signal processing, and text analysis has undergone profound changes.

The objective of this internship is to investigate how the use of APIs enables querying large databases of scientific articles. Specifically, we are interested in exploring APIs provided by renowned publishers such as Elsevier and Scopus, as well as those offered by citation databases like OpenAlex and OpenCitation.

By exploring these APIs, we aim to leverage their advanced functionalities to access and analyze massive sets of scientific data. This will allow us to extract relevant information, conduct bibliometric studies, identify emerging trends, and contribute to the advancement of knowledge in our fields of interest.

By utilizing these APIs, we can automate the process of data collection and analysis, thereby speeding up the research process. This also opens up the possibility of exploring larger and more diverse datasets, which can lead to significant discoveries and a deeper understanding of the subjects under study.

In summary, this internship offers a unique opportunity to explore recent advancements in deep learning within the field of scientific research. By utilizing APIs from renowned publishers and citation databases, we can harness massive datasets, accelerate research, and contribute to the evolution of our fields of study.

Chapter 1

Data Collection

In this chapter, we will discuss the techniques used, specifically in Python, to retrieve data from APIs (specifically Scopus and OpenAlex) and the methods for saving and preserving that data. We will explore the data extraction techniques, including accessing the APIs and handling data formats using Python libraries and tools. Additionally, we will cover the mechanisms and strategies for storing the obtained data effectively using Python programming. By examining these aspects and utilizing Python, we aim to gain a comprehensive understanding of the data retrieval and preservation processes for Scopus, OpenAlex, and similar APIs.

1. Application Programming Interface

An API, or Application Programming Interface, is a set of rules and protocols that allows different software applications to communicate and interact with each other. It serves as a bridge between different systems, enabling them to exchange data, request services, and execute operations. APIs define the methods, data formats, and authentication mechanisms that developers can use to access the functionality of a particular software or platform. By providing a standardized interface, APIs facilitate the integration of different software components, enabling them to work together seamlessly.

APIs are widely used in various domains, such as web development, mobile app development, cloud computing, and IoT (Internet of Things), enabling developers to leverage existing services and build new applications more efficiently.

1.1 Elsevier Scopus APIs

Elsevier is a renowned publisher of scientific, technical, and medical research articles. Scopus is one of their flagship databases, providing comprehensive coverage of scholarly literature across various disciplines. Elsevier offers a range of APIs that allow developers to access Scopus data programmatically. These APIs enable users to retrieve metadata, abstracts, citation information, and other relevant data from the Scopus database.

1.2 OpenAlex APIs

OpenAlex is a project that aims to make large-scale scientific data more accessible to researchers and developers. It provides a collection of APIs that grant access to an extensive corpus of scientific articles and associated metadata.

2. Developing our database

In our study, we will specifically focus on articles published in the Medical Image Analysis journal.

2.1 Medical Image Analysis

Medical Image Analysis (MedIA) is an esteemed scholarly journal that centers around the analysis of medical and biological images. This peer-reviewed publication serves as a platform for disseminating research papers that make significant contributions to the fundamental principles of analyzing and processing biomedical images obtained through various modalities, including magnetic resonance imaging, ultrasound, computed tomography, nuclear medicine, x-ray, optical and confocal microscopy, among others. The journal encompasses a wide range of topics, such as feature extraction, image segmentation, image registration, and various other image processing techniques that find applications in areas such as diagnosis, prognosis, and computer-assisted interventions.

2.2 Retrieving the data

We used the Scopus API to retrieve all the DOIs (Digital Object Identifiers) of the Medical Image Analysis journal. Our search yielded a total of 2578 DOIs. These DOIs cover articles published between 1996 and 2023. Along with the **DOIs**, we also obtained the **titles**, publication **years**, and the number of **citations** for each article. This data will enable us to analyze and gain insights into the field of medical image analysis.

	DOI	Title	Year	Citations
0	10.1016/j.media.2023.102828	Dynamic weighted hypergraph convolutional netw...	2023	0
1	10.1016/j.media.2023.102832	DLGNet: A dual-branch lesion-aware network wit...	2023	0
2	10.1016/j.media.2023.102829	DeepSTI: Towards tensor reconstruction using f...	2023	0
3	10.1016/j.media.2023.102826	Calibrating segmentation networks with margin-...	2023	0
4	10.1016/j.media.2023.102807	Low-field magnetic resonance image enhancement...	2023	0
...	...	...	...	...
2573	10.1016/S1361-8415(01)80006-2	Development and preliminary evaluation of VISL...	1996	68
2574	10.1016/S1361-8415(01)80002-5	A representation for mammographic image proces...	1996	63
2575	10.1016/S1361-8415(01)80004-9	Multi-modal volume registration by maximizatio...	1996	1574
2576	10.1016/S1361-8415(01)80005-0	Analysis of left ventricular wall motion based...	1996	174
2577	10.1016/S1361-8415(01)80003-7	Segmentation of 2-D and 3-D objects from MRI v...	1996	188

2578 rows × 4 columns

Figure 1 Pandas Dataframe of MedIA DOIs

We wanted to retrieve the number of citations by year for each article of media using the Scopus API. However, this is only possible if your institution has a subscription to the Elsevier API and if you have access to the network while performing the task. Since these conditions are not met, we will instead utilize the OpenAlex API, which offers free usage and provides the necessary information.

We utilize the OpenAlex API to retrieve the count of references and citations for every year:

DOI	Title	Year	Citations	Reference	1996	1997	1998	1999	2000	...	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
6/j.media.2023.102828	Dynamic weighted hypergraph convolutional netw...	2023	0	58	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6/j.media.2023.102832	DLGNet: A dual-branch lesion-aware network wit...	2023	0	38	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6/j.media.2023.102829	DeepSTI: Towards tensor reconstruction using f...	2023	0	80	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6/j.media.2023.102826	Calibrating segmentation networks with margin-...	2023	0	11	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6/j.media.2023.102807	Low-field magnetic resonance image enhancement...	2023	0	60	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10.1016/S1361-8415(01)80006-2	Development and preliminary evaluation of VISL...	1996	68	9	1	12	11	5	7	...	1	0	0	0	2	0	1	1	1	0
10.1016/S1361-8415(01)80002-5	A representation for mammographic image proces...	1996	63	9	1	2	4	9	4	...	6	0	1	0	1	0	0	0	1	0
10.1016/S1361-8415(01)80004-9	Multi-modal volume registration by maximizatio...	1996	1574	9	3	14	48	40	52	...	86	73	59	58	52	49	46	48	43	10

Figure 2 Complete MedIA Dataframe

Chapter 2

Data Analysis

In this chapter, we analyze the article data obtained from the Scopus and OpenAlex APIs. By leveraging these APIs, we access a rich dataset that includes citation numbers and references. Using statistical analysis and visualizations, we uncover trends and patterns within the scholarly literature. Our analysis provides insights into citation impact, influential authors, knowledge flow, and emerging research areas. The use of multiple plots enhances our understanding of the data, allowing us to make significant conclusions. This analysis contributes to the overall understanding of the research landscape and informs future studies.

1. Visual representations

1.1 Number of published articles by year

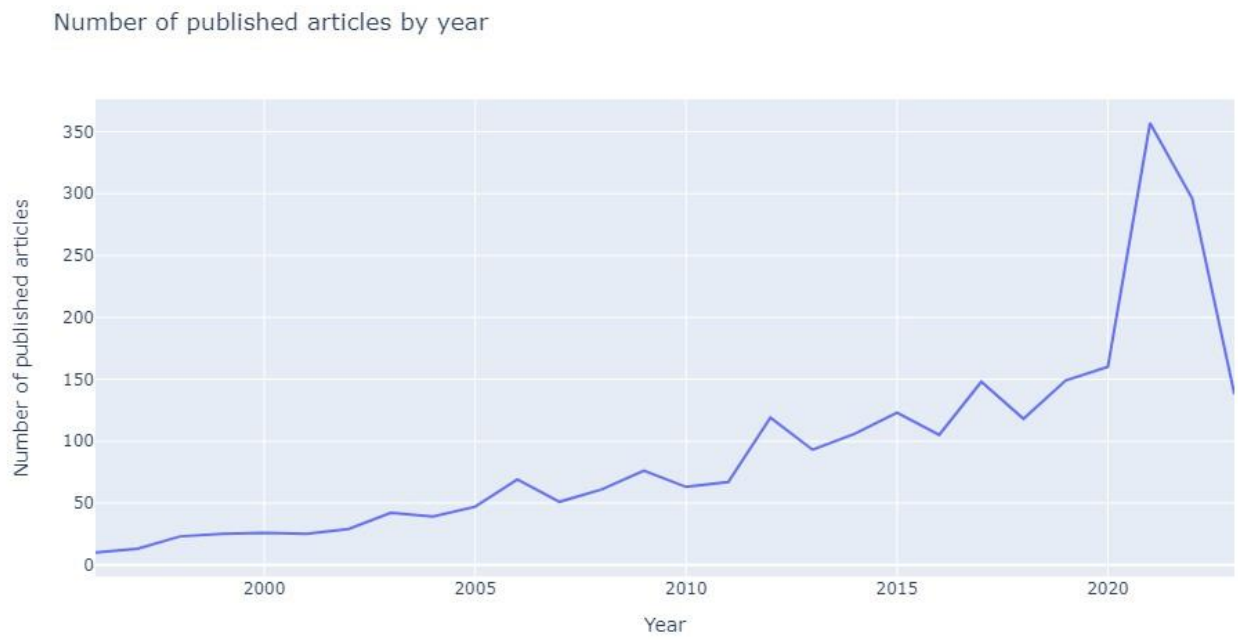

Figure 3 Number of published articles by year

The number of published articles by year has been steadily increasing since 1996. However, in 2021, there was a significant surge in publications, surpassing 350 articles.

This growth can be attributed to the COVID-19 pandemic and the development of deep learning and AI, which greatly impacted the field of medical image analysis.

These factors created a perfect storm, leading to an explosive increase in the publication count and opening new possibilities for advancements in healthcare and scientific knowledge.

We have observed a decline in the number of published articles recently, primarily because we do not possess the complete set of articles published in 2023.

1.2 Number of citations of the most cited articles per year

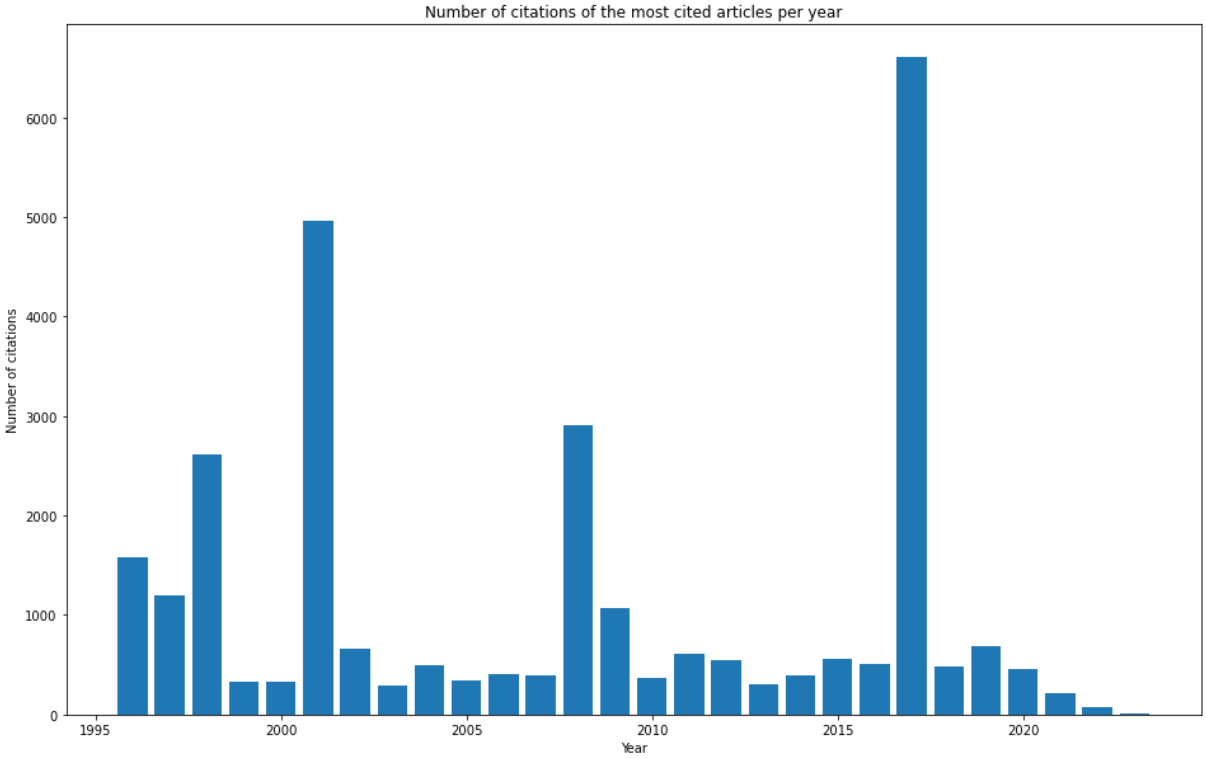

Figure 4 Number of citations of the most cited articles per year

The three most cited articles on MedIA:

1. A survey on deep learning in medical image analysis (2017)
2. A global optimisation method for robust affine registration of brain images (2001)
3. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain (2008)

1.3 Evolution of citations over time for each article from 1996 to 2023

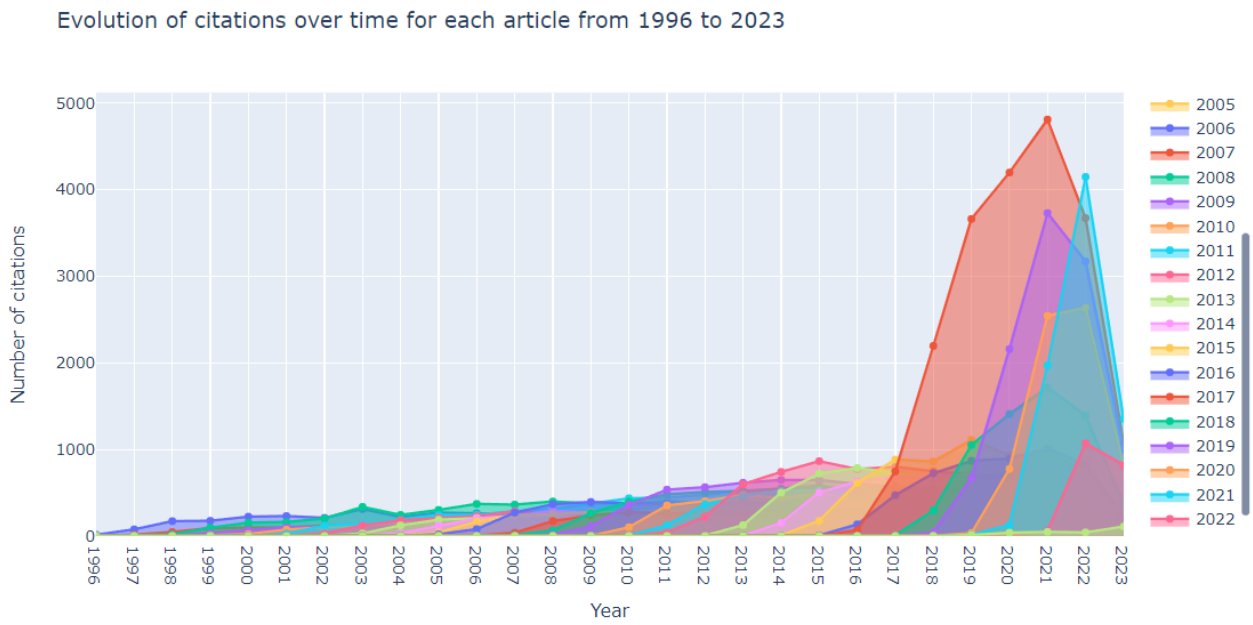

Figure 5 Evolution of citations over time for each article from 1996 to 2023

we notice that we are citing very young articles, which illustrates the acceleration of research and the pace of publications.

For example, we observe that the most cited articles within the past 5 years are those published in 2017.

1.4 Evolution of countries' contributions over the years

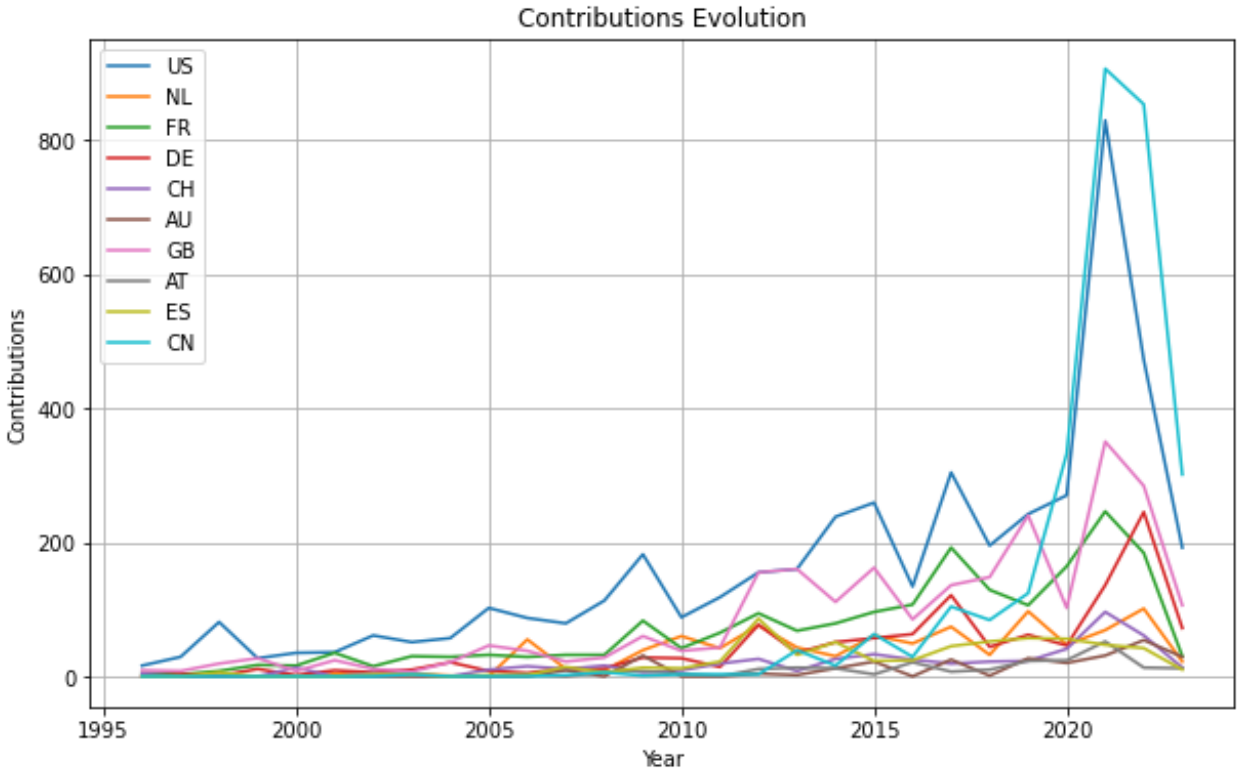

Figure 6 Evolution of countries' contributions over the years

The United States, Great Britain, and France have been the countries that contribute the most consistently. However, in recent years, we have noticed a significant increase not only in China's contribution but also in the contribution of the United States.

1.5 Analyzing the frequency of keywords in titles

1.5.1 Segmentation

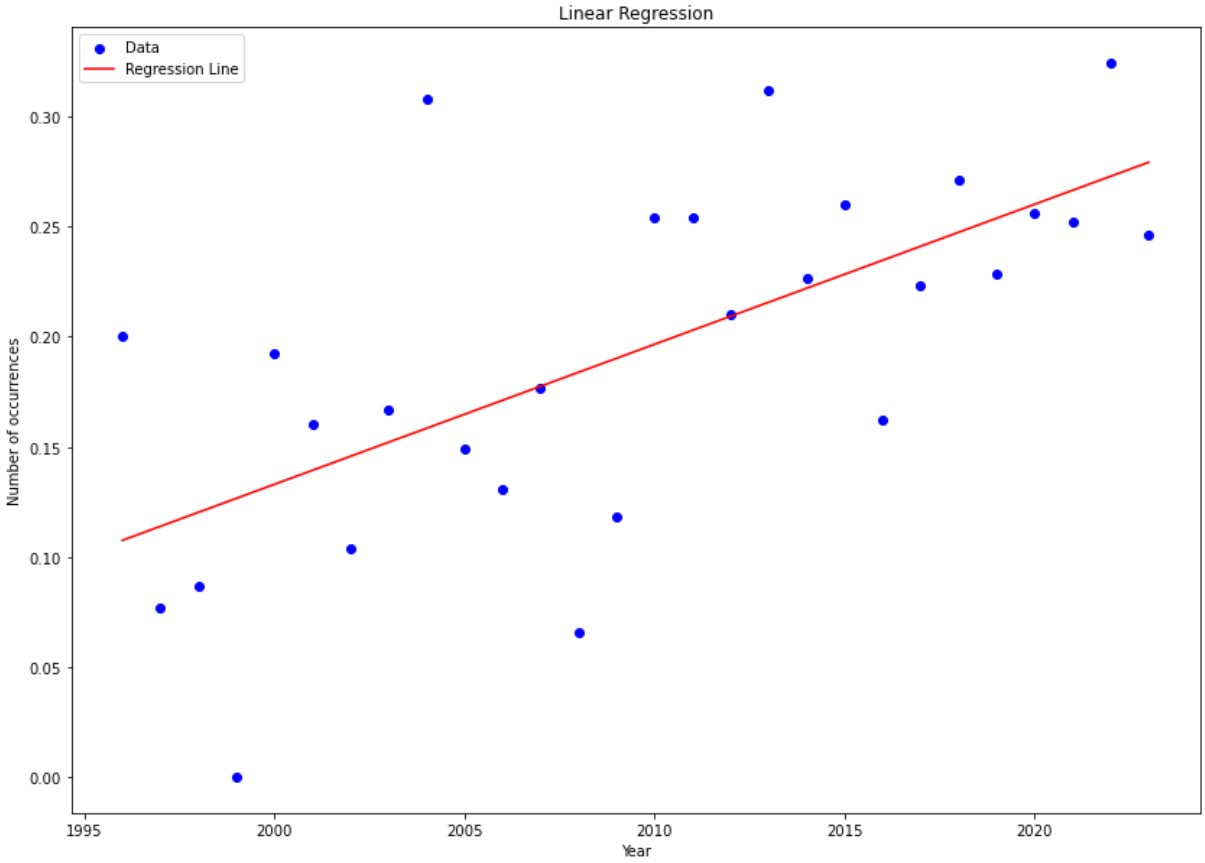

Figure 7 Analyzing the frequency of the keyword Segmentation

We notice an increase in the use of the keyword "segmentation".

1.5.2 Classification

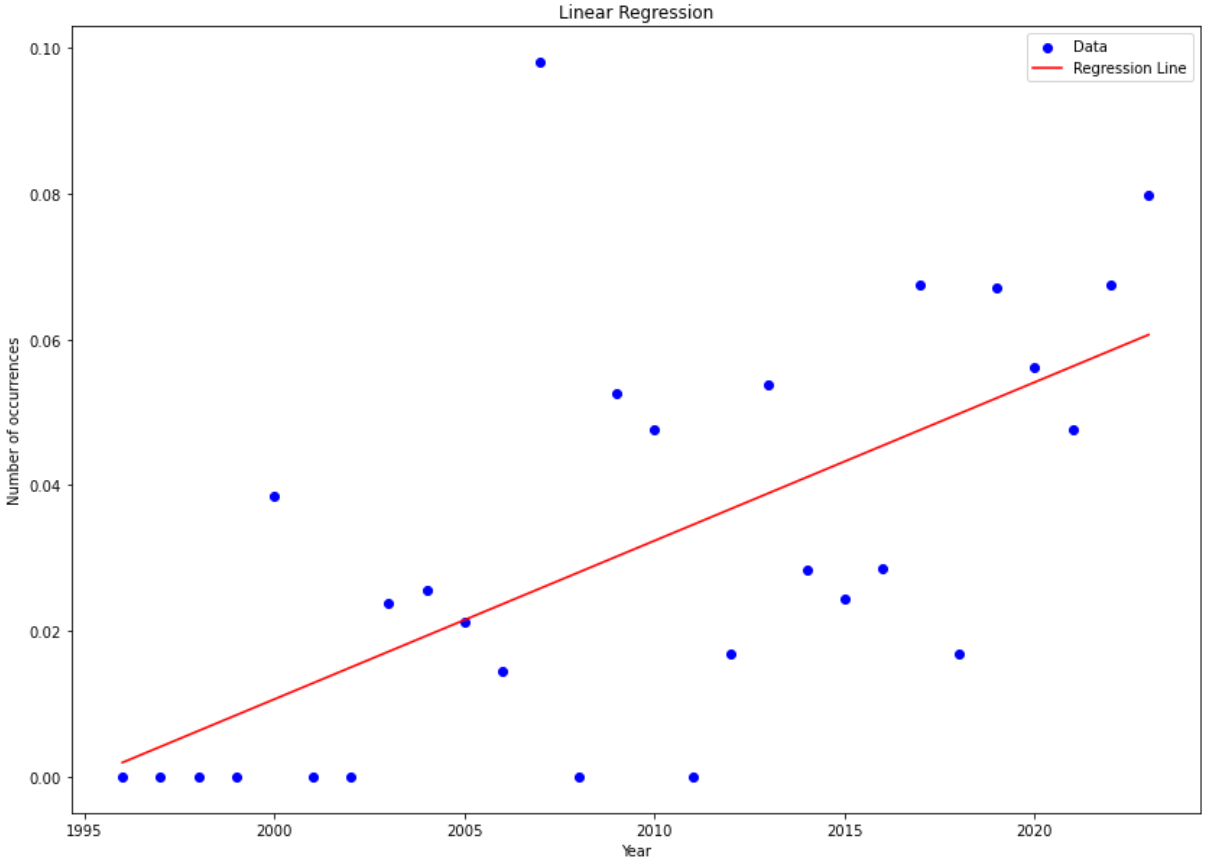

Figure 8 Analyzing the frequency of the keyword Classification

We notice an increase in the use of the keyword " Classification".

1.5.3 Registration

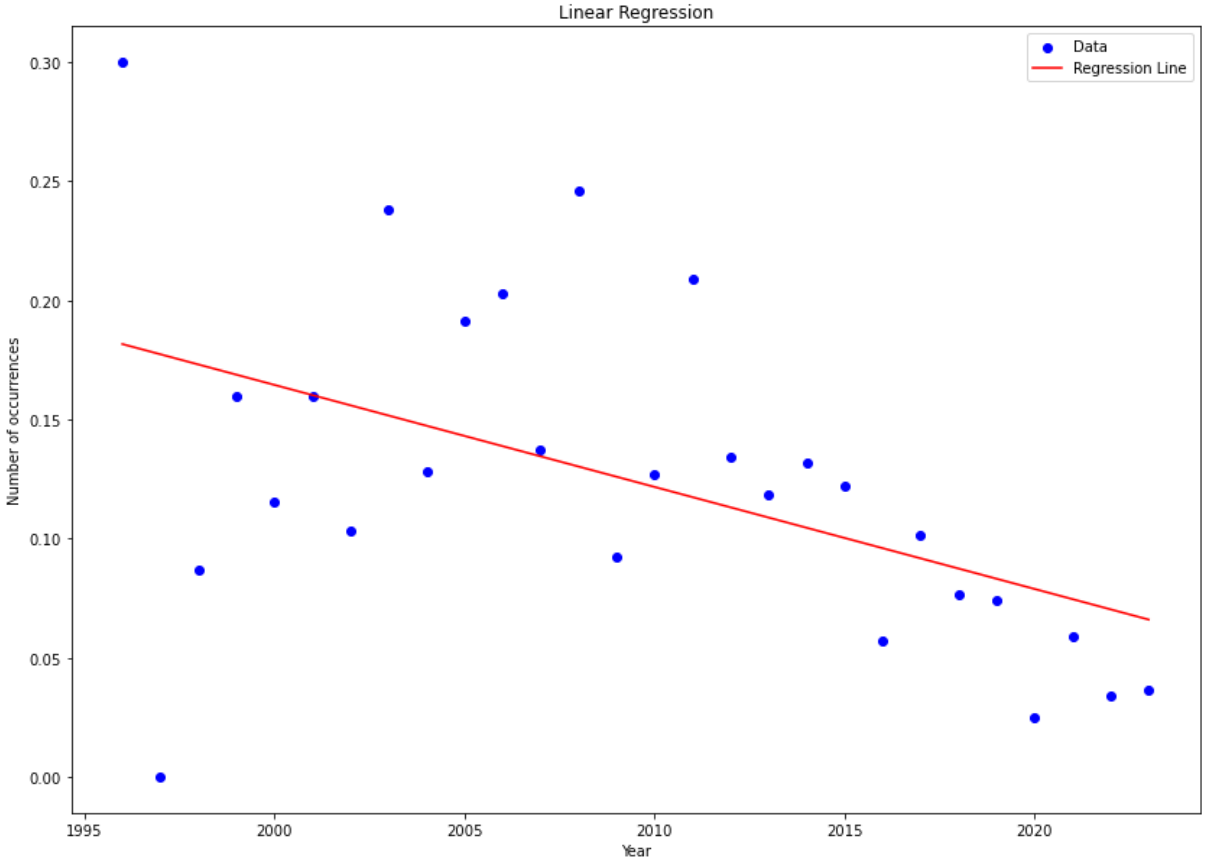

Figure 9 Analyzing the frequency of the keyword Registration

We notice a decline in the use of the keyword " Registration"..

1.6 Analyzing the number of citations that cites an article (references)

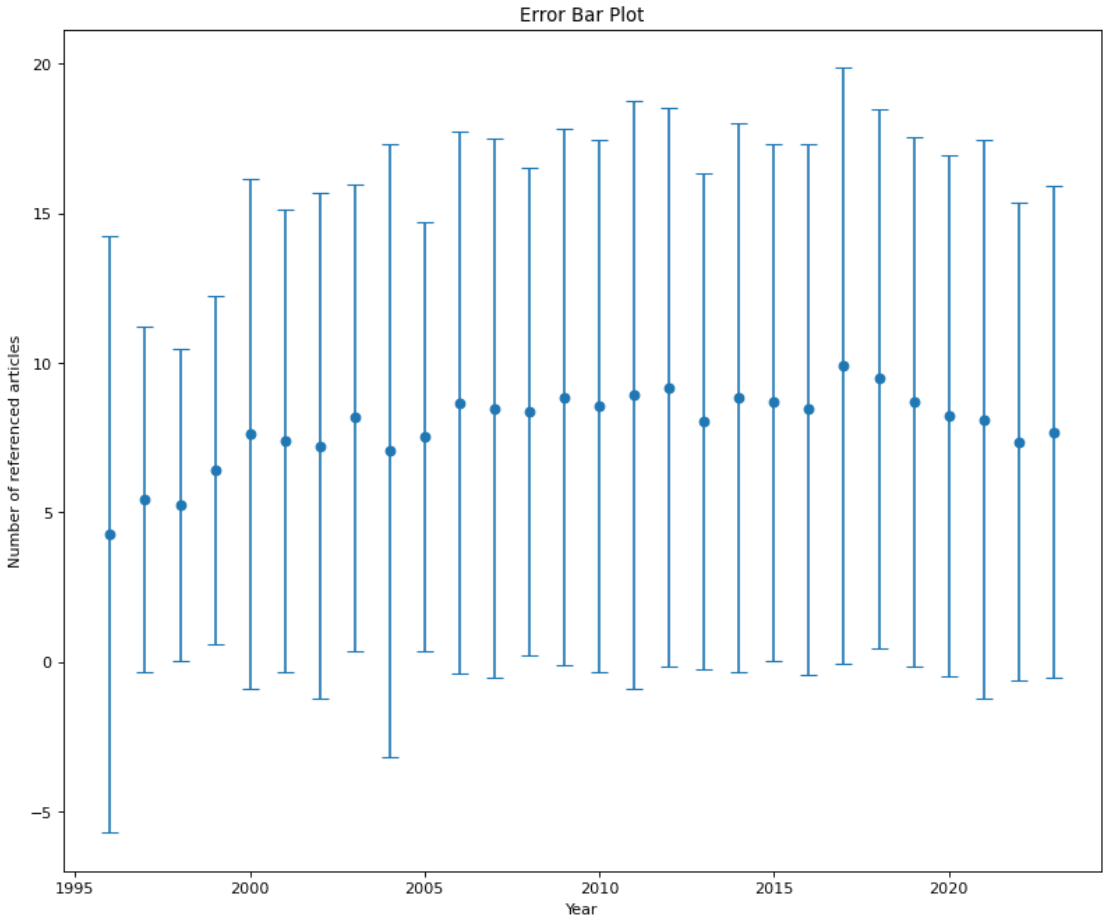

Figure 9 Error bar plot for the number of references per year

Chapter 3

Development tools

In this chapter, we present an extensive overview of the tools and programming languages utilized for our project. We employed a range of powerful resources including Python libraries such as NumPy, Pandas, Matplotlib, Plotly, and Requests to perform thorough analysis on our dataset.

1. Programming languages and libraries

1.1 Python

Python is a high-level programming language known for its simplicity, readability, and versatility. It was created by Guido van Rossum and first released in 1991. Python has a clean syntax and emphasizes code readability. It supports multiple programming paradigms and has a large standard library and extensive third-party ecosystem. Python is widely used in web development, data analysis, machine learning, and scientific computing. It is cross-platform compatible and has a strong community support.

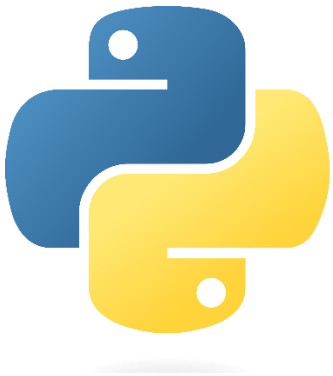

Figure 10 Python logo

1.2 NumPy

NumPy is a Python library for numerical computing. It offers a powerful multidimensional array object and a range of mathematical functions for efficient numerical operations. NumPy is widely used in scientific computing, data analysis, and machine learning. Its features include N-dimensional arrays, mathematical functions, broadcasting, integration with other libraries, efficiency, and strong community support.

Figure 11 NumPy logo

1.3 Pandas

Pandas is a Python library for data manipulation and analysis. It introduces the DataFrame, a powerful data structure for working with structured data. Pandas provides functions for data manipulation, handling missing data, time series analysis, integration with other libraries, input/output capabilities, performance optimization, and has strong community support.

Figure 12 Pandas logo

1.4 Matplotlib

Matplotlib is a Python library for creating visualizations. It offers versatile plotting functions, customization options, integration with NumPy and Pandas, exporting capabilities, and a supportive community.

Figure 13 Matplotlib logo

1.5 Requests

The Requests library is a popular Python library for making HTTP requests. It simplifies the process of interacting with web services and APIs by providing easy-to-use functions and methods. Requests supports customization of requests, handling responses, session management, error handling, SSL certificate verification, proxy support, and has a supportive community.

Figure 14 Requests logo

Conclusion

This project focused on various aspects, beginning with the data collection phase. In this chapter, we discussed the techniques employed, specifically in Python, for retrieving data from APIs, particularly Scopus and OpenAlex. We explored data extraction methods, including accessing APIs, handling data formats using Python libraries and tools, and implementing effective strategies for data storage. By comprehensively examining these aspects and utilizing Python, we aimed to gain a thorough understanding of the data retrieval and preservation processes for Scopus, OpenAlex, and similar APIs.

Next, we delved into the data analysis phase. In this chapter, we analyzed the article data obtained from the Scopus and OpenAlex APIs. Leveraging these APIs provided us access to a rich dataset that encompassed citation numbers and references. Through statistical analysis and visualizations, we uncovered valuable trends and patterns within the scholarly literature. Our analysis offered insights into citation impact, influential authors, knowledge flow, and emerging research areas. The utilization of multiple plots enhanced our understanding of the data and enabled us to draw significant conclusions. This analysis contributed to the overall understanding of the research landscape and provided valuable insights for future studies.

Lastly, we presented an extensive overview of the development tools and programming languages employed in our project. We utilized a range of powerful resources, including Python libraries such as NumPy, Pandas, Matplotlib, Plotly, and Requests, to conduct thorough analysis on our dataset.

By focusing on data collection, analysis, and employing appropriate development tools, we have successfully conducted a comprehensive study that enhances our understanding of the research domain and informs future research endeavors.