ANALYSIS OF THE INDUSTRIAL CONTRIBUTION TO THE AI STATE-OF-THE-ART

Submitted by:

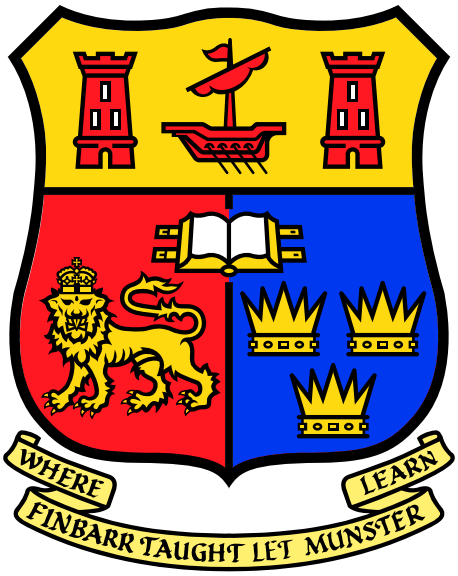
ANIMESH SHARMA

Supervisor:

DR ANDREA VISENTIN

Second Reader:

DR CATHAL HOARE



MSc Computing Science

School of Computer Science & Information Technology
University College, Cork

September 3, 2023

Abstract

Artificial intelligence is undergoing rapid evolution, leading to significant changes in various aspects of human life. This evolution has sparked increased fascination due to its enhanced efficiency and potential risks. The realm of literature has witnessed a substantial surge in contributions from researchers across diverse industries and universities. As a result, a vast repository of scholarly data has emerged, offering valuable insights for market growth, strategic planning, collaborations, and a deeper understanding of AI trends. Scholarly articles, containing crucial metadata and bibliographic information, are now widely published online. Analyzing this information can yield essential insights and patterns that contribute to a comprehensive comprehension of the AI landscape. The primary goal of this research is to comprehensively understand the contributions made in this field through systematic bibliometric analysis. This analysis involves examining the volume of contributions, international collaborations between countries, contribution by industries and universities, and emerging trending topics that captivate researchers. To achieve this, a meticulous method is employed to extract and organize the data accurately. Subsequently, the data is subjected to thorough analysis and visualization techniques, facilitating a clearer depiction of trends and patterns in the field of artificial intelligence. A selection of diverse Artificial Intelligence conferences was made, including AAAI, FOGA, ICCV, IJCAI, NeurIPS, SIGKDD, UAI, and Python was employed as the programming language for data analysis. The findings derived from this analysis indicate that universities have consistently maintained a high level of contribution with over 60%. Notably, Google and Microsoft stand out as the primary contributors in the corporate sector. The conference with the highest contribution is AAAI. The United States emerges as the most prominent international collaborator. Lastly, the analysis highlights that deep learning, natural language processing (NLP), and machine learning have emerged as trending topics.

Declaration

I confirm that, except where indicated through the proper use of citations and references, this is my original work and that I have not submitted it for any other course or degree.

Signed: _____

Animesh Sharma
September 3, 2023

Acknowledgements

I would like to express my heartfelt gratitude to Dr. Andrea Visentin for his invaluable guidance and unwavering support. From the very beginning of my journey, every week he provided innovative feedback, guiding me through each small step towards the achievement of my grand aspirations. Dr. Visentin unlocked my hidden potential and consistently encouraged me without applying undue pressure. His mentorship created the perfect environment for my growth and success, allowing me to blossom and shine. Furthermore, I extend my deepest appreciation to my parents, who have been my pillars of strength, continuously motivating me to work diligently and efficiently. Their unwavering belief in my abilities has been a driving force in helping me reach my goals. They have stood by my side every single day, offering their support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The rapid advancements in artificial intelligence technology are revolutionizing every aspect of human existence, making it more fascinating, efficient, and dangerous at the same timng, learning, adapting, understanding the senses,e. AI can be described as the ability to mimic human intelligence through computational means, which enables the highest accuracy possible to achieve a specific goal. This goal includes human cognitive abilities such as reasoni and participation. It provides the opportunity to address difficulties, but it can also become a problem. Natural language processing, speech recognition, and other similar technologies are more common and have been incorporated into our daily lives. Many companies are capitalizing on the AI trend by rebranding their solutions and utilizing the hype to enhance their brand recognition.

The achievements are undoubtedly impressive, but they can be easily overestimated and misunderstood. AI encompasses characteristics such as big data processing (structured or unstructured data), reasoning (context-driven awareness), learning (developing models based on historical data by recognizing patterns), and complex problem-solving. (Stefan van Duin 2018) describes that AI can be categorized into *narrow AI* and *general AI*. The majority of AI applications fall under the category of narrow AI, implying they are designed for specific tasks such as chess, mathematical calculations, translation, facial recognition, and speech understanding. On the other hand, general AI refers to the concept of a single system capable of learning about various problems and subsequently providing solutions, similar to how humans specialize in particular domains and become experts in them.

In recent years, artificial intelligence has seamlessly integrated into our daily lives with AI-driven solutions that simplify tasks, enhance safety measures, and reduce the time and energy expended in various domains. Notably, Generative AI has achieved remarkable progress this year, with breakthroughs like *dall-E* and *chatGPT*. In less than a year, AI has significantly enhanced the efficiency of its models, effectively doubling workers' productivity. It demonstrates its versatility by creating different types of content, including images, code, text, audio, and video. The figure 1.1 shows how the integration of generative AI has led to a noticeable improvement in the productivity of software developers. The time required to accomplish coding tasks has been reduced to 55% of

the original time.

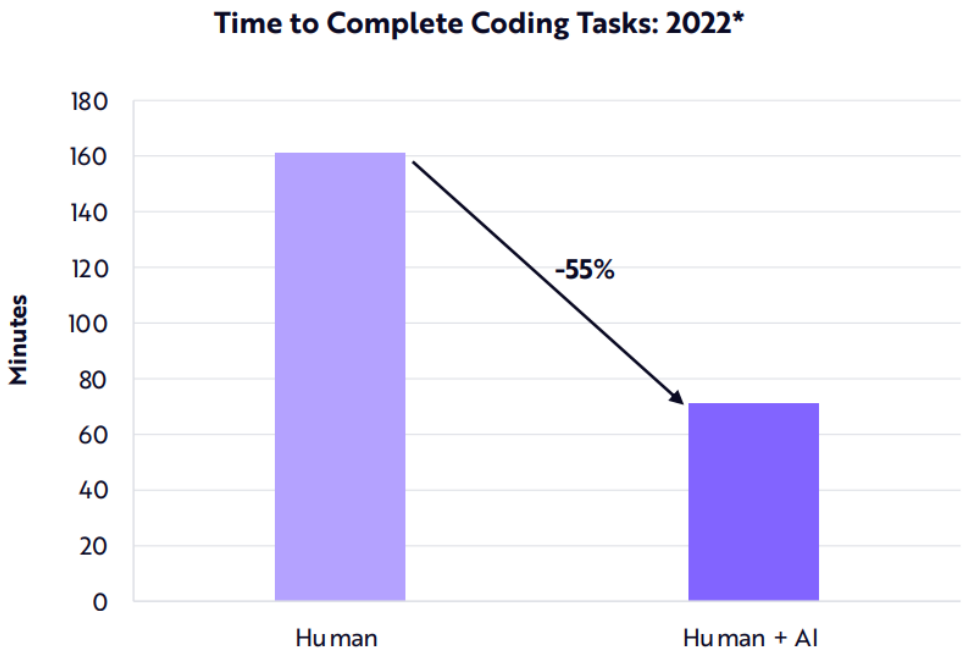


Figure 1.1: Developer Productivity Through Generative AI Technology (LLC 2023)

Understanding the achievements of diverse organizations across different sectors is crucial, as the impact of AI on improving human life cannot be underestimated. The organizations and academic universities contributing to this technology consist of highly skilled researchers and experts who extract insights from the technology and make critical decisions for the development of their organizations and society. The transformation of society brought about by AI helps us to understand the complex challenges we face and how they can be overcome for progress and innovation in this new era. However, a more in-depth review is necessary and understanding of the trends associated with artificial intelligence in every field. The amount of data collected in healthcare, e-commerce, weather, and many other sectors requires less maintenance time and cost to be more effective. According to (Jimma 2023), Healthcare in particular, will undergo a drastic change. Improving diagnosis of problems, predicting diseases, and providing better therapy solutions that benefit both healthcare providers and patients. Every year, the volume of information in academia and industry is growing rapidly. This is mainly due to the increased involvement of researchers who are using advanced technology in artificial intelligence to make significant contributions. These contributions are made through a variety of venues, including academia, journals, papers, conferences, publications, and articles.

1.1 Motivation

Research is carried out in various fields, and researchers contribute to studies on specific topics or areas of interest. These researchers, coming from various backgrounds and possessing different skills, have collectively produced millions of papers. These papers are published at various conferences and contain informative knowledge with valuable metadata. Thus, the question arises: 'What are the current and emerging sub-fields and discipline-specific trends and patterns in artificial intelligence research that may be examined globally?' It is also important to understand 'how the contributions made in the field of artificial intelligence by various organizations and universities differ among conferences, countries, authors, citations, and topics?'. With the use of data mining methods and systematic analysis, researchers can find valuable information and discover patterns that help them understand how various studies are connected.

Artificial intelligence has made its way into various domains, influencing and shaping research in numerous areas. A systematic analysis reveal the evolution of keywords, the distribution of topics, research associations, and the movement of research across different subjects. The collective efforts of researchers across disciplines have led to a vast body of knowledge and insights, with emerging trends in specific fields highlighting the dynamic nature of research and its impact on various domains.

The AI Index Report (Maslej et al. 2023) is produced by the Stanford Institute for Human-Centered Artificial Intelligence (HAI), which offers thorough and insightful information about artificial intelligence. In this report, they have presented a graph 1.2 that illustrates the number of AI-related publications worldwide between the years 2010 and 2021. These publications include various types of contributions, such as academic journals, articles, conference papers, repositories, and patents. The graph demonstrates that the number of AI publications has experienced significant growth over the years. In 2010, there were around 200,000 AI publications, but by 2021, this number had nearly doubled, reaching close to 500,000. This indicates a substantial increase in research and development related to artificial intelligence during this period.

Figure 1.3 represents the number of AI publications (in thousands) from 2010 to 2021. Each line represents a specific sub-field of AI and its corresponding publication count during this period. The increasing trend in these lines reflects the growing interest and research activity in areas like pattern recognition, machine learning, computer vision, algorithms, data mining, natural language processing, control theory, human-computer interaction, and linguistics.

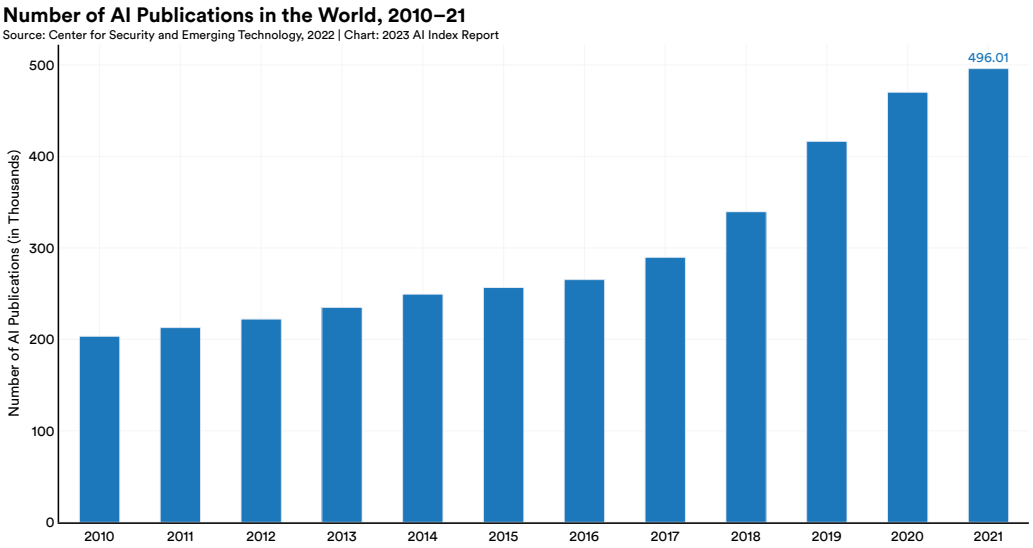


Figure 1.2: Number of AI Publications in the world [2010-2021] (Maslej et al. 2023)

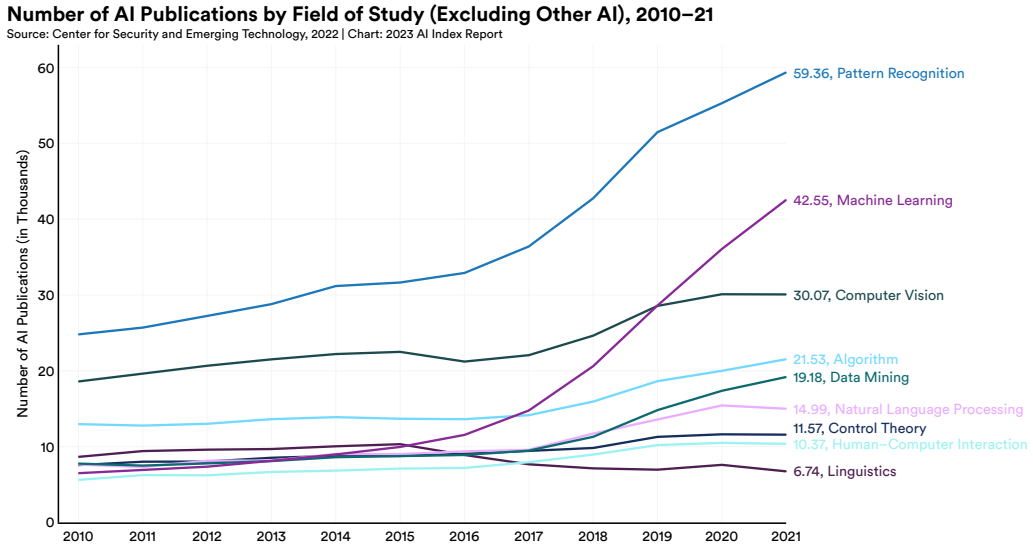


Figure 1.3: Number of AI Publications by field of study [2010-2021] (Maslej et al. 2023)

1.2 Goals of this Work

The main objective of this research is to conduct an in-depth and organized bibliometric analysis within the realm of artificial intelligence and its associated domains. By utilizing statistical analysis and systematic review techniques, the objective is to gain valuable insights into the historical contributions that have shaped the development of artificial intelligence over time. This research aims to provide a thorough understanding of the trends and topics that have emerged within the field, shedding light on its evolution and current state.

To achieve this goal, a well-defined process is used that involves data extraction, transformation, cleaning, analysis, and visualization. The openly accessible DBLP dataset and Scopus will serve as the main source of data, ensuring the availability of a vast and diverse set of scholarly information on artificial intelligence. By analyzing this open-source bibliographic data it will be easy to understand the state of the art, including the various sub-fields of artificial intelligence and their contributions to academic publications and conferences.

One of the key aspects of this research is the development of a method to extract relevant information from the web efficiently. This approach will help in gathering a substantial amount of data to form a comprehensive and representative dataset. Through careful pre-processing and data cleaning, the quality of the dataset will be enhanced, ensuring that the subsequent analyses are accurate and reliable.

A particular focus of the research involves conducting a descriptive analysis, which will enable the identification of patterns and trends within the field of artificial intelligence. This investigation will help in developing a deeper understanding of the field's benefits, disadvantages, and possible growth areas. The research will strive to identify the most efficient methods for data acquisition, pre-processing, post-processing, and the selection of appropriate models for forecasting future trends which will help in potential breakthroughs and areas for future exploration.

Chapter 2

Background

In this chapter, we will discuss the key strategies and techniques involved in developing trend analysis. Methods involved in data extraction, analysis is covered in detail in section 2.1. Section 2.4 shows the evolution of trend analysis and how it is changed through time is discussed with the exploration of historical growth, significant turning points, and innovations that have influenced the discipline of Artificial intelligence.

2.1 Data Acquisition

There are multiple techniques and processes used to obtain relevant information from the internet. There are numerous programming and markup languages that support free text integrated within tags. (Ferrara et al. 2012) shows that these languages provide developers and users with the flexibility to include content without strict formatting constraints. Tags, represented by angle brackets ("<>" or "</>"), play an important role in organizing and structuring the information, determining how it is presented or processed. This combination of free text and tags is widely used in web development, enabling the creation of dynamic and user-friendly websites that contain information that can be extracted. The figure 2.1 shows data flow while extraction of data.

Figure 2.1: Flow Diagram of Web Scraping
(Ruchitaa, Nandhakumar, and Vijayalakshmi 2023)

HTML is a form of semi-structured data that is used to store free text in a nested structure. It is a primary language that is widely adopted to create web pages across the internet. One common application for extraction of semi-structured data is in designing suitable wrappers. Wrappers are tools or scripts that help extract specific information from semi-structured data sources, transforming it into a more structured format that can be easily processed and analyzed. A significant portion of the data available on the internet exists in a semi-structured format, including content from sources like emails and social media platforms. But the research targeting data extraction from this type of source is beyond the scope of this work.

2.1.1 Web Crawling

The internet, or World Wide Web, contains a huge amount of data. Web crawling, or scraping, is a process that helps in data extraction in a systematic manner. A software program or script called a web crawler or spider robot retrieves the information present on the internet. A crawler simulates the behavior of humans by sending requests to the website and can grab any information that is present on that particular web page.

To extract specific information from the nested structure of the data, some of the libraries are present in the Python language, which makes it more effective to easily target the data.The below Python packages will aid in retrieving specific information from almost any web page present on the internet. The libraries help in requesting appropriate information and then parsing the results.

- **URLlib:** The URLlib package in Python is a powerful tool for working with URLs and interacting with web resources. It allows users to easily open URLs, send requests to web servers, and fetch data from them. With urllib, handling errors is smooth, ensuring program stability. It also respects website rules defined in the "robots.txt" file, promoting responsible web usage. Supporting various HTTP methods like GET, POST, PUT, and DELETE, urllib simplifies working with URLs and enables easy interaction with web resources in Python, following web standards responsibly.

- **BeautifulSoup:** BeautifulSoup is a tool for efficiently navigating through URLs and parsing HTML and XML files. This package converts web documents into a unicode representation and organizes them into navigable tree structures. By automating data extraction and providing functions, it can efficiently extract specific data from web pages. The user-friendly structure of BeautifulSoup shown in figure (2.2) simplifies the process of working with web data. The latest version is Beautifulsoup4, which offers enhanced features and improved performance. Additionally, BeautifulSoup can be utilized in conjunction with the ETree module to create XML data by parsing APIs.

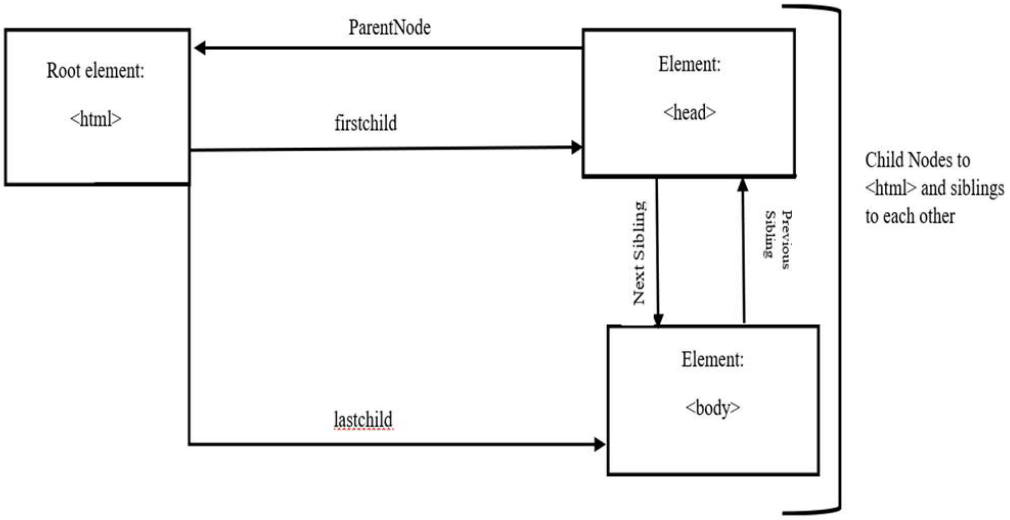


Figure 2.2: Data Retrieval using BeautifulSoup
Ruchitaa, Nandhakumar, and Vijayalakshmi 2023

- **Scrapy:** Scrapy is an efficient and powerful open-source web scraping and crawling framework. To extract particular information and navigate through web pages, this package offers a robust and fully open-source solution that handles concurrent requests effectively. It can be used in conjunction with other packages, making it even more scalable. Scrapy's architecture can be seen in figure (2.3) which allows precise control over data flow through items, spiders, and pipelines, enabling data processing and storage in various formats, including CSV and JSON. Moreover, Scrapy can be seamlessly integrated with add-ons like Splash, adding further capabilities to the scraping process. With its asynchronous networking and support for XPath and CSS selectors, Scrapy continues to be a popular choice for web scraping projects, enabling fast, versatile, and responsible data extraction from websites.

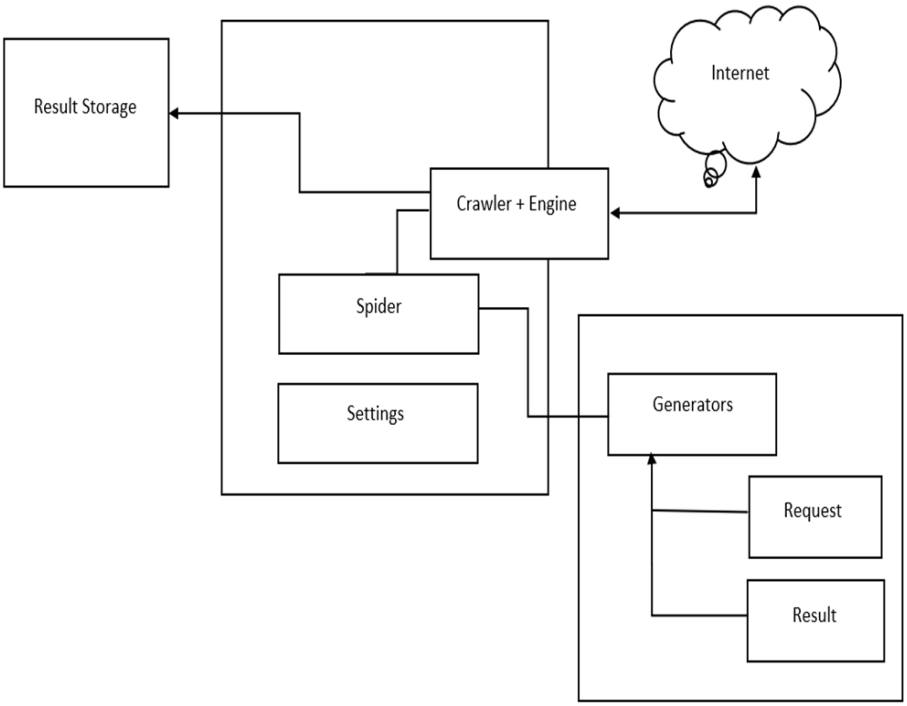


Figure 2.3: Data Retrieval using Scrapy
(Ruchitaa, Nandhakumar, and Vijayalakshmi 2023)

- **Selenium:** It is an open-source package for building browser automation tools primarily used for testing. It consists of a core component called the web driver, allowing developers to interact with web browsers. It also includes Selenium IDE, an extension for testing in multiple environments. Additionally, it can act as a web scraper or crawler. figure (2.4) shows the extraction of information from HTML elements. It is particularly useful for handling dynamic content. The process involves setting up the Selenium web driver with the associated web browser and providing a seed URL as a string. The web driver then opens the URL and interacts with HTML elements through actions like form filling, button clicking, and data extraction. Specific information can be extracted by specifying the Xpath as an argument. It can handle websites with heavy Javascript or those requiring complex interactions effectively.

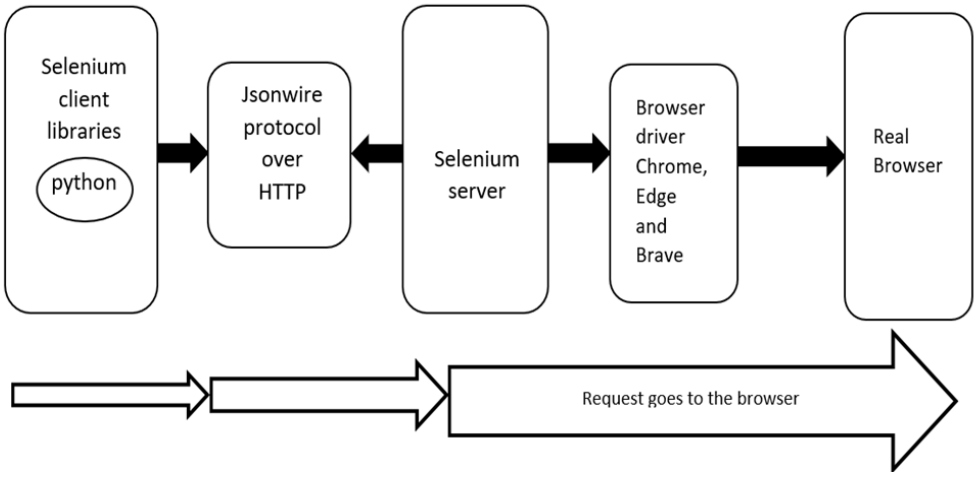


Figure 2.4: Data Retrieval using Selenium
(Ruchitaa, Nandhakumar, and Vijayalakshmi 2023)

2.2 Natural Language Processing

The internet is filled with various types of documents, including text files with extensions like .txt, .doc, .html, .xml, and more. When the information is extracted from these documents, we need to analyze them effectively.

Natural language processing (NLP) is a sub-field of AI that helps in understanding or making better sense of language from regular expression using statistics. There are numerous ways for natural language processing described below.

2.2.1 Regex

A regular expression is a string with special syntax and can be referred to as regex. In Python, there is a special library for regex which is the "`re`" library. In this library, there are different methods such as:

- **Split:** This function divides the input string into substrings based on where the pattern appears in the input string.
- **Findall:** Looks through the input string for every instance of the pattern and returns a list of matches.
- **Search:** If a match is discovered, the pattern in the input string produces a match object.
- **Match:** Returns a match object if there is a match and the pattern matches at the start of the input string. Only at the start of the string, a pattern can be recognized.

There are various ways to identify patterns in a string or text, as shown in the table below.

2.2.2 Tokenization

Tokenization is another process to transform a string or document into smaller chunks. It is a step that is performed in pre-processing of text data. There are multiple theories and rules to perform tokenization and it is possible that we can create our tokenization. Natural language toolkit library (NLTK) is the most commonly used library to tokenize the document. It helps to understand and map parts of speech and match common words.

Below are some common techniques used in Tokenization.

- **Sent_tokenize:** This function helps to split a string or sentence into individuals.
- **Regexp_tokenize:** It is used to tokenize a text using regular expressions. A custom regular expression pattern to identify the tokens in the text can be used.
- **Tweet_tokenize:** It is a special class used for tweet tokenization which allows separating hashtags, mentions, and exclamations.

2.2.3 Stemming / Lemmatization

Stemming is a process that shortens words in a text to their root stem or base form by removing suffixes and prefixes. The main purpose of stemming is to consolidate words with similar meanings, enabling efficient usage in models and text analysis tasks. Let's consider the following input words:

- "running,"
- "jumps,"
- "jumping,"
- "happiness."

After applying stemming, we get the following results:

- "running" -> "run"
- "jumps" -> "jump"
- "jumping" -> "jump"
- "happiness" -> "happi"

Stemming may not always achieve a perfect result, as seen with the word "happiness" being overgeneralized to "happi." However, despite its limitations, stemming remains a crucial preprocessing step for enhancing text classification and understanding the meaning of text. By using stemming, we can reduce various inflected forms of words to their common base form, allowing text analysis models to focus on essential patterns and insights within the text.

2.2.4 Lowercasing of Words or Tokens

Lowercasing is a process that transforms all uppercase text to lowercase. This step is essential because it ensures consistency among all the words in a document. However, it's important to note that lowercasing may not be suitable in certain situations, such as when dealing with "names of people, place names, or organizations." In such cases, preserving the original capitalization is necessary.

2.2.5 Stop Words

Stop words are common words that frequently appear in a text but don't carry much meaning on their own, like "the," "of," "a," "in," and "for." There's no universal list of stop words since it depends on the context and language. The reason we use stop words is to reduce the dimensionality and complexity of a document and make processing more efficient. By removing these commonly occurring but less meaningful words, we are left with only the words that are more informative and essential for tasks like text classification, sentiment analysis, and information retrieval. These important words can provide valuable insights into the meaning and context of the text.

2.2.6 Feature Generation and Selection

Feature generation involves transforming the text into numerical values for analysis.

- **Boolean representation:** Converting words to binary values (0 and 1) to indicate their presence or absence in a document.
- **Bag of words representation:** Counting the frequency of words in a document without considering their order.

These techniques enable the conversion of text data into a format suitable for analysis, facilitating tasks such as text classification, clustering, and information retrieval.

2.2.7 Document Similarity

Lexical similarity examines how similar documents are based on the words and phrases they share. It looks at the overlap of words and other lexical features between documents. The Cosine Similarity method measures the similarity between two documents by calculating the cosine of the angle between their word frequency vectors. A higher

cosine similarity score indicates a greater lexical similarity between the documents. Lexical similarity can be subdivided into calculations for single terms and calculations for phrases between documents. Semantic similarity, on the other hand, focuses on understanding the meaning and context of words and phrases within documents, aiming to capture the underlying relationship between words.

2.3 Clustering

Clustering refers to the grouping of entities or variables that share commonalities or similar attributes. It involves applying artificial intelligence techniques to automate the discovery of clusters. This process is categorized as unsupervised learning since it does not rely on pre-existing categorization during training. There are multiple clustering algorithms techniques available, including K-means, DBSCAN, and Spectral. These algorithms aim to partition the data in different ways which allows identifying underlying patterns and structures the data without manual labelling. Each clustering algorithm has its own strengths and limitations, making it suitable for different types of data and clustering objectives.

2.4 Literature Review

This section delves into a comprehensive exploration of trend analysis within the realm of Artificial Intelligence (AI). The primary objective here is to carefully examine the accomplished implementations of trend analysis techniques. This involves a thorough discussion and analysis of the various methods and solutions proposed in the existing literature for conducting trend analysis in AI applications. The examination of literature falls into two categories:

- Literature related to conducting trend analysis in various sectors.
- The current techniques being used for web scraping, preprocessing, and visualization.

2.4.1 Trend Analysis Across Industries in AI

We are witnessing the emergence of new technological methods known as Industry 4.0, which represents the fourth industrial revolution. This revolution is characterized by the widespread digitization and integration of artificial intelligence in various industrial processes. The industries are adopting advanced technologies such as automation and robotics, cloud platforms, Blockchain, and virtual reality, which provide modern approaches for making life less complex. Big Data enables industries to make complex decisions for market growth. This involves the collection, filtering, and modeling of data to drive product growth. The integration of human intellect and AI into a unified framework is currently in high demand within innovation systems (Vaio, Hassan, and Alavoine

2022). Analysis of data with aid of AI and ML approach have been a research focus for both practitioners and academics because they enhance usage and overcome complex issues that aid the decision-making process. A bibliometric analysis of 161 papers published in the years 2017-2021 is utilized. From Web of Science (WoS), Scopus, and Google Scholar to understand that artificial intelligence could be useful for data and business intelligence in the public sector for better decision-making and the role of the human-AI interface with respect to skills and applications in ethics design. The primary objective of this study is to examine the current state of data, business intelligence, and the human-AI interface in order to establish a more effective framework for implementing AI in the public sector. The study focuses on analyzing relevant literature in the field of AI, including books, articles, chapters, and conference papers. The research design and methodology are structured into four categories: data collection and processing, utilization of the PRISMA framework (an open-source tool for managing databases) to gather relevant details, performance analysis, and science mapping. Specific search queries were used to collect the necessary data, and the PRISMA framework facilitated the selection of 161 papers for analysis. The analysis involves examining the annual growth of published papers, the citation patterns of these papers, the interconnection between papers in terms of countries, topics, and sources, the top keywords, and the top references by citation. These analyses will provide insights into the current trends and developments in the field of AI, specifically within the public sector.

AI has made significant advancements in various sectors, including healthcare. The integration of artificial intelligence in healthcare holds great potential for improving the accuracy of diagnosis and treatment, benefiting both patients and healthcare providers. The integration of AI in the field of healthcare leverage difficulties by improving the efficiency in maintaining the electronic health records. The AI models are capable of predicting and emulating human physicians capabilities. For instance, a clinical trial study found that AI systems contributed to a 50% increase in medical adherence among stroke patients undergoing treatment.

In order to examine the usage of Artificial Intelligence in the healthcare sector from 2000 to 2021, a bibliometric analysis was conducted (Jimma 2023). The focus of the analysis was on publications related to the field of AI in healthcare. The study utilized data extracted from Scopus, a comprehensive database of scholarly literature. To perform the analysis, a structured search was conducted in Scopus, targeting journals specifically covering artificial intelligence technologies and their applications in healthcare. The retrieved documents were thoroughly examined, and various aspects were analyzed, including the growth of publications within specific subject areas and the contributions of different countries to the field. The findings were visualized using the VOS software, which generated visual representations of frequently used author keywords, offering insights into the prevailing topics and themes in the literature. It is important to note that this study solely focused on the Scopus database. To gain a comprehensive understanding of the trends surrounding AI in the healthcare sector, it would be necessary to analyze additional databases as well.

A research study (Kathiria, and Arolkar 2022) was conducted on Indian universities

and research organizations using data obtained from Scopus. The study involved analyzing a collection of published papers in the field of computer science to understand the research trends and activities carried out by Indian researchers. The data spanned the years 2010 to 2019. The total number of published papers that were extracted is 54,051. In addition to analyzing past trends, the researchers also utilized forecasting techniques to predict the research activities for the years 2020–2023. To ensure the accuracy of their forecasts, the data is divided into a training dataset consisting of the years 2010 to 2017 and a test dataset consisting of the years 2018 and 2019. The repository of data undergoes a preprocessing step to prepare it for analysis. This involves transforming the data into numerical measures using the Tf-Idf matrix, which incorporates a shared phrase graph structure known as the document graph index graph model. The Tf-Idf matrix enables the calculation of cosine and phrase similarity based on the shared phrases within the documents. These similarities are then combined to create a hybrid similarity measure. The corpus of documents is clustered to group similar documents together, aiding in the identification of unique topics. The clustering technique used for this is DBSCAN, which takes into account the hybrid similarity measure. For each cluster, the topic is identified using LDA (Latent Dirichlet Allocation), and automatic labeling is performed using CSO (Computer Science Ontology). An ARIMA model with a grid search approach is employed to estimate and forecast the trend of each topic for the next four years. This allows for insights into the future trajectory of the identified topics. The authors of the study have identified that keywords such as internet, artificial intelligence, data mining, and software engineering were popular trends in the years 2018 and 2019, and they are expected to continue being popular from 2020 to 2023. The forecasting of these trends was based on the MAPE measure, which resulted in a value of 18.34% (the average percentage difference between the predicted values and the actual values). Recognizing these trending topics will be beneficial for researchers who are starting their investigations, as it provides a starting point for further exploration and research in these areas.

A bibliometric analysis (Ampadi Ramachandran et al. 2023) was conducted using a web crawler that focused on curating published records for data retrieval from a database in the field of veterinary or human medicine. The research aimed to understand the rate of depletion of drugs and chemicals in animals used for food production. A software architecture was developed, utilizing the internationally accepted Anatomical Therapeutic Chemical (ATC) classification for drugs. Text and Data Mining (TDM) techniques were employed to retrieve data from API service providers such as Scopus, Springer, CrossRef, and ArXiv. These providers granted API keys to access their data through different search queries. Authorization was obtained through institutional login. For the crawling process, the Python library Selenium was used. The crawler extracted data from websites and APIs using Selenium and the provided APIs, generating a collection of Digital Object Identifiers (DOIs) as input. These DOIs were then used to extract full-text content from the API sources. To ensure efficient data retrieval, the Scheduler library was employed for periodic downloading. API services often impose limitations on the amount of data that can be retrieved at once, so periodic downloading

helped overcome these restrictions. Finally, the extracted data was organized into a dataframe and saved in CSV format for further analysis and interpretation.

In the rail industry, a scientometric analysis (Yong, and Lee 2022) was conducted to understand the impact of machine learning and deep learning methods in the sector. The analysis used a dataset of 12,675 published papers obtained from Scopus. To retrieve relevant papers, a search query was used that included keywords related to machine learning, deep learning, and rail transportation. The initial list of publications was refined by excluding papers from irrelevant journals, those without abstracts or index keywords, and papers that specifically focused on review or framework work. Furthermore, papers from 2021 were excluded to avoid incomplete results, as it was the ongoing year at the time of the analysis. After applying these filters, 6,717 papers remained for further analysis. To conduct the scientometric analysis, four main goals were established: research trends, research targets, influential studies, and leading journals, authors, and countries. In order to achieve the targets, multiple methods were employed. Time series analysis. This statistical technique was used to analyze and interpret the data collected over time, providing insight into the trends and patterns in the published papers. Frequency distribution. This method counted the occurrences of unique values or ranges of values in the dataset, helping to summarize and understand the distribution of specific keywords or topics within the articles. Binary classification system, This system likely assisted in categorizing the papers into relevant and irrelevant categories based on specific criteria. Major research topics: The analysis probably identified the main areas of study within machine learning applied to the rail industry, such as maintenance activities, traffic management, rolling stock, rails and passengers. Co-occurrence keywords network and Louvain method. These techniques helped identify the relationships and communities within the dataset by analyzing the cooccurrence of keywords. The Louvain method, a community detection algorithm, partitioned the network into distinct groups or communities. Degree centrality analysis, This analysis assessed the importance or centrality of nodes (e.g., authors, papers) within the network, identifying influential authors of papers in the field. Citation analysis, A tree mapping method was employed to analyze the citation patterns among the papers, revealing the impact and influence of specific studies. h-index, This metric quantified both the productivity and impact of researchers' work, indicating the number of highly cited papers they had. Collaboration network: By examining coauthorship relationships, the collaboration network likely provided insights into the structure and dynamics of research collaborations in the field. The results of the scientometric analysis revealed several findings, including the focus of machine learning studies in the rail industry on maintenance activities, traffic management, rolling stock, rails, and passengers. The analysis also identified the distribution rankings of research targets, future challenges, top countries, leading authors, highly cited papers, influential publications, and popular keywords.

2.4.2 Web Scraping and pre-processing of data

Web data extraction involves various methods for gathering information from websites. To retrieve and interpret unstructured data, a diverse range of tools and techniques are employed. These methods enhance our understanding of digital content.

(Liu, Zhu, and Guo 2021) demonstrated strategies for extracting unstructured data from internet using a specific web scraping techniques. A scraping framework was developed to connect with HTTP/HTTPS or web browsers to extract the desired information from web pages. These web pages are present in different formats such as XML, HTML, and CSS, and each website has its own unique layout, which is crucial to understand in order to extract the desired information. Scraping software includes a crawler that follows a script to locate specific HTML elements on a website. If the website's layout changes, the crawler may stop working, and the only solution is to update the code by modifying the element accordingly. To retrieve the data, The author used two techniques in order to extract the information which is beautifulsoup and SERP API and described which method is best in different scenarios.The HTML web page contains elements in recursive order and to extract the information such as request, beautifulsoup and scrapy can be used. The HTTP protocol is utilized to send and receive requests. This is done using libraries such as CURL and traverse, which facilitate sending an HTTP GET request to a specific URL of a website. Once the request is received and the webpage is accessed, various techniques can be employed to extract the desired information, such as HTML parsing, DOM parsing, XPath, and text matching. Once the data is retrieved, it often appears in an unstructured format. It is very critical to pre-process the data before storing them in a desirable format such as CSV, spreadsheet, or PDF. SERP API is an API that provides real-time data, unlike traditional APIs, where you need to make a request to establish a connection between the client and the server. This API is especially useful for dealing with captchas. The Google Scholar API allows users to scrape data from search queries by using it as the endpoint in the script "/search?Engine=google scholar". To utilize the information it is important to create an SERP API account and obtain a secret key. With the SERP API, you can extract specific fields such as author information, author citations, articles, and coauthors. The author of the statement discovered that crawlers efficiently extracted the information and stored it in a CSV file. Python proved to be an efficient programming language for this task, as it offers various packages and libraries for data scraping. Using different techniques, such as Beautiful Soup and the SERP API, can be beneficial depending on specific requirements.

The authors (Moral Munoz et al. 2020) focused on exploring the tools that can be utilized for bibliometric and scientometric analysis. These types of analysis involve evaluating and quantifying various aspects of scientific publications, such as their impact, performance, and relationships. The authors identified and examined both theoretical and practical tools that are commonly used in bibliometric and scientometric analysis. These tools serve different purposes, including measuring general bibliometric indicators, conducting performance analyzes of researchers or institutions, creating scientific maps to visualize research areas and collaborations, and studying different aspects of libraries in relation to scientific publications. Using these tools, researchers can

gain valuable insights into the impact and influence of scientific publications, identify patterns and trends in research fields, assess the performance of individuals or organizations, and examine the role of libraries in supporting scientific endeavors. In order to find an adequate state of art the source of database, bibliometric software, SMA tools, R and python libraries are used. The source of information to create a database is from scopus, pubmed, wos, google scholar, microsoft academic and dimensions. The paid software that can be used, which can serve as an alternative, is altmetrics but it requires authorization with organization. In the context of performance analysis, several tools are commonly used. These include CRexplorer, Publish or Perish, and ScientopyUI. For scientific mapping purposes, a range of tools are available, such as Bibexcel, Biblioshiny, BiblioMaps, WebCiteSpace, CitNetExplorer, SciMAT, and Sci2 tool. The analysis is often conducted using programming languages like R and Python. During their analysis, the authors discovered that the Web of Science (WoS) only allows retrieval of 500 records per query in plain text and tab-delimited format. On the other hand, Scopus allows extraction of up to 2000 records in RIS and CSV formats. Google Scholar and Microsoft Academic do not allow direct data downloads, although Microsoft Academic offers an API for accessing the data. To overcome these limitations, performance analysis tools like Publish or Perish enable direct data download by utilizing API keys, which can be obtained from the respective websites. Additionally, Dimensions allows downloading up to 50,000 records; however, its web service lacks proper conference filtering. Each tool and library offers its own approach to data pre-processing and filtration. Notably, Publish or Perish software and the R language do not incorporate data preprocessing capabilities. However, the SciMAT preprocessing technique stands out as a comprehensive software library for this purpose. In relation to specific functionalities, tools such as CRExplorer, Bibliometrix/Biblioshiny, and Meta knowledge can be employed to create spectrograms. Burst detection is supported by tools like CiteSpace, Sci2 Tool, Metaknowledge, and Bibliometrix/Biblioshiny. For geographical analysis, tools such as Bibexcel, Bibliometrix/Biblioshiny, BiblioTools/Biblio Maps, CiteSpace, and Sci2 Tool prove useful. These tools collectively offer a range of features to analyze bibliometric data, identify research activity bursts, and perform geographic analyses.

(Dastani et al. 2020) used a method called text mining, which is used to discover patterns in large datasets. The researchers extracted information from PubMed and Scopus databases, covering the years 1964 to 2019. Initially, they retrieved a total of 12,819 papers, but some of them did not contain abstracts, so those were removed, leaving them with 7,599 papers. The main objective of the study was to identify emerging keywords in the field of medical librarianship and information. To achieve this, the researchers used text mining techniques, specifically using the Porter root algorithm and TF-IDF (term frequency inverse document frequency). The process consisted of three main stages: data preprocessing, text mining operations, and post-processing. During the data preprocessing stage, the researchers performed tasks such as data selection, categorization, normalization, feature selection, and removal of stop words. In text mining operations, the Porter stemming algorithm was applied to find the root form of words. This algorithm prunes word suffixes, allowing different word forms to

be converted to the same root form. Following the stemming process, the researchers applied the TF-IDF technique to assign weights to the words. In the post-processing stage, we have a word cloud visualization technique, which visually represents textual data. The size of each word in the word cloud is proportional to its importance, as determined by its weight. Through this visualization, the researchers identified the most significant words in the field of medical librarianship and information, which turned out to be "librari," "patient," and "inform," with respective weights of 95.087, 65.796, and 63.386. Additionally, the researchers analyzed the weight of words across different decades. They found that the word "catalog" was trending in the 1960s and 1970s, while "patient" became prominent in the years 2015 to 2020. This analysis provided insights into the evolution of keywords in the field over time.

2.4.3 Findings from the Literature Review

The principal observations that emerge from the literature review are :

- The data for the above study has been primarily sourced from authoritative repositories, such as Web of Science, Scopus, and Google Scholar. Additionally, web crawlers and APIs have been utilized to ensure a comprehensive dataset.
- To process and refine the data, the PRISMA framework was employed. Applying advanced Natural Language Processing (NLP) techniques, including tokenization, stemming, lemmatization, and word vectorization (such as TF-IDF), has enabled the transformation of textual data into quantifiable metrics, resulting in highly refined text representations.
- Certain works within the literature have also focused on time series forecasting, employing methods like the ARIMA model to predict future trends. Other methodologies have employed techniques such as time series reversal and frequency distribution analysis to illuminate temporal patterns and understand the prevalence of specific keywords.
- For the purpose of identifying latent themes and the correlation between them, clustering techniques like DBSCAN have been employed. These techniques aid in categorizing and analyzing the primary topics present in the dataset, revealing interrelated patterns.
- The analysis has been executed through the Python programming language, utilizing libraries such as Scrapy, SciPy, NLTK, NumPy, and Pandas. These libraries enable effective data manipulation, statistical analysis, and Natural Language Processing.
- The visualization aspect of the study has been enhanced by leveraging both the Views software and Python modules, including Matplotlib. This combination of tools facilitates the creation of illustrative and informative visual representations to better convey the findings of the analysis.

Chapter 3

Methodology

In this chapter, I will present the step-by-step approach taken to achieve the research objectives. The section 3.1 discusses obtaining data using digital bibliography and library projects, as well as utilizing a Scrapy spider for data extraction. The section 3.2 describes the data flow of the extracted data. The section 3.3 refines the collected data, including the classification of conferences and affiliations. The section 3.4 outlines the attributes of the dataset. The section (3.5) describes Topic Modeling - Latent Dirichlet Allocation (LDA) applies LDA for extracting meaningful topics. The section (3.6) and (3.7) describe the techniques used to visualize the extracted information. The GitHub repository provides access to relevant code and resources.

3.1 Data Acquisition

The bibliographic data or metadata found on the Web is incredibly scattered, posing a significant challenge for researchers to understand the trends in the market. With the inclusion of multiple conferences, the task of extracting data from all these conferences becomes tedious and time-consuming due to the distinct layouts of each conference's website. However, there are organizations that offer a solution to this problem by providing comprehensive databases containing published and accepted papers from various conferences, complete with metadata and bibliographic information, along with external links to the full text of the papers.

3.1.1 Digital Bibliography and Library Project (DBLP)

The Digital Bibliography and Library Project (DBLP) is an organization that addresses the issue of maintaining metadata from multiple publications and conferences. DBLP serves as an online reference for bibliographic information and is widely recognized as a popular open data service. The DBLP offers free access to high-quality metadata and links to electronic editions of publications. The scalability of the DBLP database is impressive, indexing more than 4.4 million publications authored by more than 2.2 mil-

lion authors, spanning 40,000 journal volumes and 39,000 conferences. It is important to note that DBLP primarily serves as a bibliographic service rather than a document repository. Therefore, it does not provide the full text or abstracts of the research papers. This limitation is due to copyright laws, which prohibit the unauthorized distribution of copyrighted material. DBLP has found a workaround by incorporating hyperlinks within its database. These hyperlinks lead users to the external sources where the full text or research papers can be legally accessed. To facilitate the extraction of data from DBLP, the service offers various strategies and techniques. Users can employ web crawling or scraping methods to extract the desired information from the website.

Web crawling involves systematically browsing web pages, collecting data, and traversing the links to discover new pages. This approach can be utilized to navigate through the different conferences and extract the required metadata. Additionally, users can employ extraction strategies specific to the DBLP's layout and structure to enhance the efficiency of the data extraction process.

DBLP is a user-friendly service to assist in finding authors, profiles, conferences, journals, and individual publications. It aids in the enhancement of the navigation experience through the DBLP web pages, making it more efficient.

We can extract the information in multiple ways which are discussed as follows :

- **1. Prefix Search:** The prefix search feature allows a specific prefix in the search query. Entering a prefix, the service will return results that contain words or sentences matching that prefix. For example, if you search for "sig," it will show the results containing words starting with "sig" anywhere in them.

- **2. Exact Word Search:** By appending the "$" sign to the search query, It can perform an exact word search. This means the service will only return results the desired text which has been given as input. For instance, if searched for "graph$," It will retrieve results specifically containing the word "graph."

- **3. Boolean AND Search:** The Boolean AND search helps to combine multiple keywords in the search query. It Simply separates the words with spaces, and the service will find results containing all the specified keywords. For example, if the search query contains the word "codd model," It will return the results that include both "codd" and "model" in them.

- **4. Boolean OR Search:** Using the pipe symbol (|), It can perform a Boolean OR search. This means the service will find results containing either of the specified words. For instance, if query contains "graph|network," it will retrieve results containing either "graph" or "network" in them.

  DBLP offers three major services for searching with the help of API:

- **Publication Queries:** The API service for publication queries by accessing the following URL:
  https://dblp.org/search/publ/api

- **Author Queries:**The author API service by visiting:
  https://dblp.org/search/author/api
- **Venue Queries:** The venue API service can be accessed through:
  https://dblp.org/search/venue/api

The best part about the service is that it is freely available to all users without the need for an API Key. The dblp repository allows crawlers to follow the guidelines specified in the robot.txt file.

The data on DBLP is available in various formats, such as XML, JSON, JSONP, and Bibtex. These formats are used to structure and organize the data for easy retrieval and processing. The metadata includes information about each publication, such as a Digital Object Identifier (DOI) or URL. However, the data available on the DBLP website is incomplete or may be missing information for some conferences. To address this issue and obtain more detailed information about the papers published in conferences, it is crucial to ensure accurate metadata responses to achieve precise results from multiple metadata repositories.

The primary objective of scraping data from the DBLP webpage is to obtain DOIs and URLs, which are used as query parameters when parsed to the "start URL." Digital Object Identifiers (DOIs) are unique codes assigned to research papers to make them easily identifiable and accessible on the internet. These DOIs are linked to external services that store metadata information about these papers. Some of these services include Crossref, OpenAlex, and Opencitations, which maintain metadata such as publication details, authors, abstracts, and more. Each of these sources has its own API or data access methods, which can be initialized in a crawler to gather information.

In our implementation, we have chosen to utilize two of these services, Crossref and OpenAlex, to access the metadata related to research papers. These services act as repositories where all the necessary information about a paper can be found, as long as its DOI is available. However, there might be cases where the DOI for a particular paper is not present in DBLP. In such situations, we adopt an alternative approach. We follow the URLs provided in DBLP, which lead us to the main webpage of the conference where the paper was originally published.

The crawler will extract the DOI present in the JSON response. If the JSON response does not contain the DOI, it will extract the external link that is linked to the main page of the conference where the paper is published. Once we are on the conference webpage, the crawler, which contains a different spider file created for a particular conference, will extract the DOI from that page. These spider files help extract the required information directly from the conference pages, ensuring that we still gather all the necessary data despite its absence in the usual DBLP database. This DOI is then used to fetch all the relevant metadata from Crossref or OpenAlex, whichever service contains the information.

It is important to highlight that during the data collection process, we carefully cross-verified the metadata using multiple reliable repositories. In cases where disparities existed between acceptance and publication years, we took proactive measures to rectify

any inaccuracies. As a result of our thorough cross-verification process, the dataset is accurate and consistent.

3.2 Scrappy Spider: Data Flow

The dataflow figure 3.1 illustrates essential components of a web scraping process on the DBLP website using Scrapy. The following section shows the core elements of the scrapy library.

- **Engine:** This is the core component of the spider, which is used in managing the execution process and allows coordination with different components of Scrapy. It gets requests from spiders to crawl the DBLP webpage.
- **Spider:** We have defined a Spider to facilitate web scraping on the DBLP website. It specifies a list of start URLs from which the spider will initiate its crawling process. The DBLP website offers data in JSON format as a response. We have set the JSON URL within the start URLs list. Additionally, the spider can navigate through links, including those leading to the URLs of published papers. URLs are extracted similarly to DOIs, and another spider created for a specific conference can access these extracted URLs.
- **Scheduler:** It retrieves URLs from the queue and generates requests. A request comprises essential information about the data to be fetched from the URL, including the DOI, URL, title, authors, and abstract. It ensures the systematic visitation of each URL while preventing the generation of duplicate requests. Additionally, it facilitates rate limiting or delays during crawling, preventing any potential overload on the target server or violations of usage policies. Specifically, when extracting metadata from metadata repositories, a customized delay of 1 second is implemented for every request.
- **Downloader:** The metadata is downloaded with the help of the downloader component. The downloader receives requests from the Scheduler to retrieve metadata from HTML or other formats, such as XML or JSONP. The downloader employs various techniques, like traversing using XPath or CSS selectors. Once the data is scraped, the downloader stores it in the database.
- **Item Pipeline:** The Item Pipeline is a component that receives an item and performs various actions on it. These actions include data cleaning, validation, duplicate checks, dropping data, and storing data in the database. During the extraction of DOIs or URLs, if any duplications are detected, they will be automatically removed.

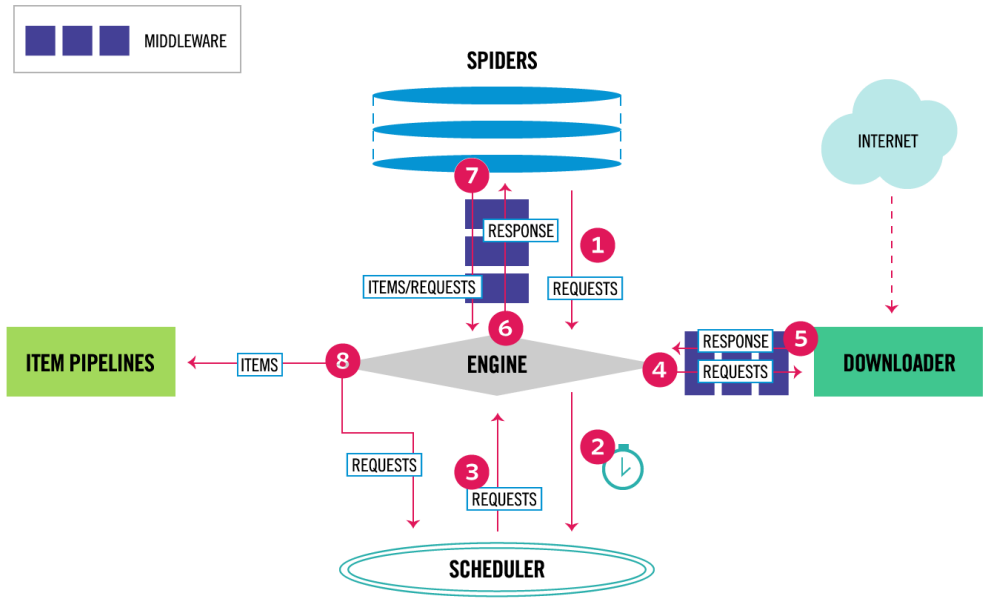


Figure 3.1: Dataflow of Scrapy
(Scrapy 2023)

3.3 Data Preparation

Once all the required information has been successfully collected, the next step is to eradicate all the unnecessary information. This involves excluding data such as author identification, volume and issue specifics, article numbers, pagination details (such as starting, ending, and total page count), funding information, sponsorships, conference dates, and the publication stage.

We have collected data from various sources and organized it into multiple CSV files. These files have been standardized to follow a consistent global structure. However, each CSV file uses a different delimiter, making it essential to align the rows and columns properly to ensure consistency when merging the data. For example, the table (3.1) shows author names, affiliations, and countries and table (3.2) how it is separated into distinct columns for a published paper with authors details. To handle cases where certain papers may not have authors listed, those columns with "NA" data to maintain uniformity and avoid inconsistencies.

Author Name, Affiliation, and Country
Verdiesen I., Delft University of Technology, Delft, Netherlands Liu Y., Department of Psychology, Centre for Technomoral Futures, University of Edinburgh, Edinburgh, United Kingdom Helm P., Technical University Munich, Munich, Germany Sabbaghi S.O., George Washington University, Washington, DC, United States

Table 3.1: Author Names, Affiliations, and Countries

Author 0 Name	Author 0 Affiliation	Author 0 Country
Verdiesen I.	Delft University of Technology, Delft	Netherlands
Liu Y.	Department of Psychology, University of Edinburgh, Edinburgh	United Kingdom
Helm P.	Technical University Munich, Munich	Germany
Sabbaghi S.O.	George Washington University, Washington, DC	United States
Hullman J.	Northwestern University, Evanston, IL	United States

Table 3.2: Author Information

The column names are maintained and kept consistent while splitting because it enables easy aggregation and further preprocessing methods. For instance, if there are four authors, the "Author" column will be labeled "Author 0 Name" until "Author 4 Name" to accommodate each author's information. This consistency simplifies tasks like applying for loop and combining data or performing necessary adjustments in the future.

3.3.1 Classification of Conferences

When gathering data from the same conference, a situation aroused where the data includes various source titles for the same conference. This particular situation can pose challenges when tried to create graphical displays of the data. To effectively communicate the information and provide clarity to the audience, it becomes crucial to appropriately label each source title which have same conferences but variations in their titles. Ensuring accurate and consistent labeling is essential for avoiding confusion and misinterpretation. It helps the audience immediately recognize the source of the data and comprehend any differences.

Conference names are associated with the information in the 'Source title' column to facilitate the tagging of publications in the DataFrame. The primary goal is to categorize the publications based on their respective conference affiliations. To achieve this, a new column named 'conference' is added to the DataFrame and initialized with blank values.

The .loc indexer is utilized to filter rows where the 'Source title' column contains specific strings corresponding to different names of the same conferences. When a row's source title' matches a particular string, the corresponding 'conference' column is updated with the relevant conference name. This process is repeated for multiple conferences, each having its own unique set of identifying strings within the 'Source title' column. Notable conferences include 'AAAI', 'FOGA', 'ICCV', 'ICLR', 'ICML', 'SIGKDD', 'NIPS', and 'UAI'. Consequently, the data frame is modified to incorporate the new column arrangement.

Source Title	Conference
AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society	AAAI
Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022	AAAI
33rd AAAI Conference on Artificial Intelligence, AAAI 2019, IAAI 2019, EAAI 2019	AAAI
Proceedings of IEEE International Conference on Computer Vision	ICCV
ICLR 2022 - 10th International Conference on Learning Representations	ICLR
6th International Conference on Learning Representations, ICLR 2018	ICLR
Proceedings of Machine Learning Research 37th International Conference on Machine Learning, ICML 2020	ICML
Advances in Neural Information Processing Systems	NIPS
Advances in Neural Information Processing Systems 24, NIPS 2011	NIPS
Proceeding of Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	SIGKDD
Proceedings of 11th International Workshop on Multimedia Data Mining, MDMKDD'11	SIGKDD

Table 3.3: Source Title and Conference

3.3.2 Classification of Affiliations

The dataset includes information about the organizations to which authors belong when they contribute to the literature in various areas of artificial intelligence. These

sources of literature come from a wide range of sectors. By figuring out whether these sources are from academic institutions or industries, we can better understand which sectors are producing a larger volume of information in the field of artificial intelligence. This helps us identify the trends in which sectors are contributing more to the literature.

It's crucial to categorize affiliations and assign tags based on their sources. The method employed for this classification is a rule-based keyword approach. This method involves sorting the text into specific categories using carefully designed language rules. These rules are organized within a class that guides the system in examining terms for their label, allowing it to recognize patterns as outlined by the predefined rules. This approach is easy for humans to understand and can be enhanced by making adjustments to its rules, which in turn enhances its performance. In the figure 3.2 below, you can observe that the input (a collection of text) is fed into the text classification model. This model then produces a label based on the predefined rules.

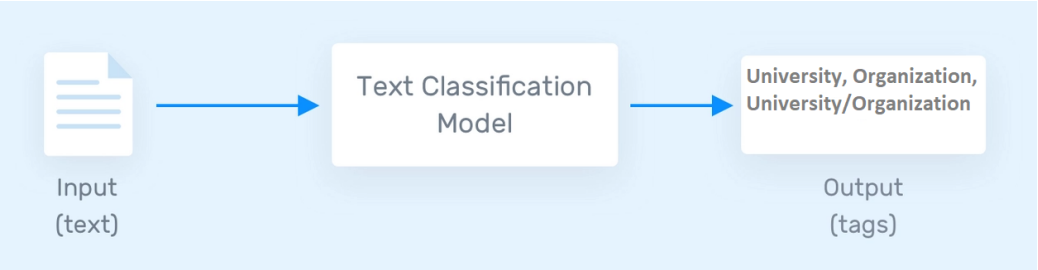


Figure 3.2: classification of Affiliation

In the table 3.4 below, if the affiliation string contains the keywords "University," "univ," or "universete," it will be labeled as "University" if the source is from an academic area. If the affiliation string indicates both an academic area and research community, it will be labeled as "University/Organization." If the affiliation string doesn't match any of these rules, it will be classified as "Organization

Affiliation	**Classification**
Telecom Paris, Institut Polytechnique de Paris, Palaiseau	University
Carnegie Mellon University, Pittsburgh, PA	University
University of Texas at Dallas, Richardson, TX	University
IBM Research AI	Organization
Beijing Lab of Intelligent Info Tech, Beijing Inst. of Tech, Beijing	Org./Univ.
Karlsruhe Institute of Technology	Univ./Org.
Advanced Analytics, Swiss Re	Organization
Amazon Alexa AI	Organization

Table 3.4: Affiliation Classification

3.4 Data Set Description

The raw data is sourced from various bibliographic repositories accessible on the internet. This data is then carefully cleansed and pre-processed to enhance its structure. The resulting dataset consists of numerous columns, each described as follows.

- **Title:** The title of the published work.
- **Year:** The year of publication when it is accepted.
- **Source Title:** The title of the source from which the work originates.
- **Conference:** Abbreviated source title.
- **Cited by:** The number of citations received by published paper.
- **Abstract:** A concise summary of the research paper and their objective.
- **Author Keywords:** Keywords provided by the authors to describe the main themes of published paper.
- **Index Keywords:** Keywords assigned to the published paper for categorization and retrieval by the source.
- **Conference Location:** The location where a conference took place.
- **Language of Original Document:** The language in which the published paper is written.
- **Author Name:** The name(s) of the author(s) of the publication.
- **Author Affiliation:** The institutional or organizational affiliation(s) of the author(s).
- **Author Affiliation Classification:** Classification of the author's affiliation(s).
- **Author Country:** The country of origin of the author(s).

3.5 Topic Modeling - Latent Dirichlet Allocation (LDA)

Extracting "information" from raw data is crucial. Text mining involves many words, but analyzing and deriving meaning from the data is essential. In a large database of text documents, automatic topic identification is vital for detecting patterns and understanding the content. Topic modeling methods, such as LDA, enhance text mining results, leading to better decision-making. While regular expression, rule-based, or keyword-based techniques are options, LDA offers a distinct approach specifically designed for topic modeling.

Latent Dirichlet Allocation (LDA), The term "Latent" means some information is present in the document which is not yet discovered. LDA is an unsupervised learning approach that is used for extracting and observing a bunch of keywords represented as topics in a large cluster of text. A repetitive pattern of co-occurring words in a document is considered a topic. LDA can create a model that represents a particular field of interest. For instance, a model can be said to be good if the output of the model is "patient, doctor, medicine, swelling" for the topics "Healthcare" and "Wheat, flour, rice" for the topic "Farming". A model is useful when it creates a large cluster with a huge number of highly frequent words which are correlated to each other that are present in unstructured data. For text mining, multiple approaches can be incorporated with LDA, such as Term frequency, inverse document frequency, and the nonnegative matrix factorization technique, where LDA can generate more optimized results.

3.5.1 Parameters of LDA

From Figure 3.3 The collection of documents, considered as a dataset, is input into the LDA (Latent Dirichlet Allocation) algorithm. As shown in Figure 3.4 'm' represents the total number of documents within the corpus, 'n' is the total number of words in a document, 'alpha' and 'beta' are Dirichlet distributions, and 'theta' and 'phi' are multinomial distributions. By employing these distributions, we can derive the number of topics present in the corpus and determine the frequency of each topic per document.

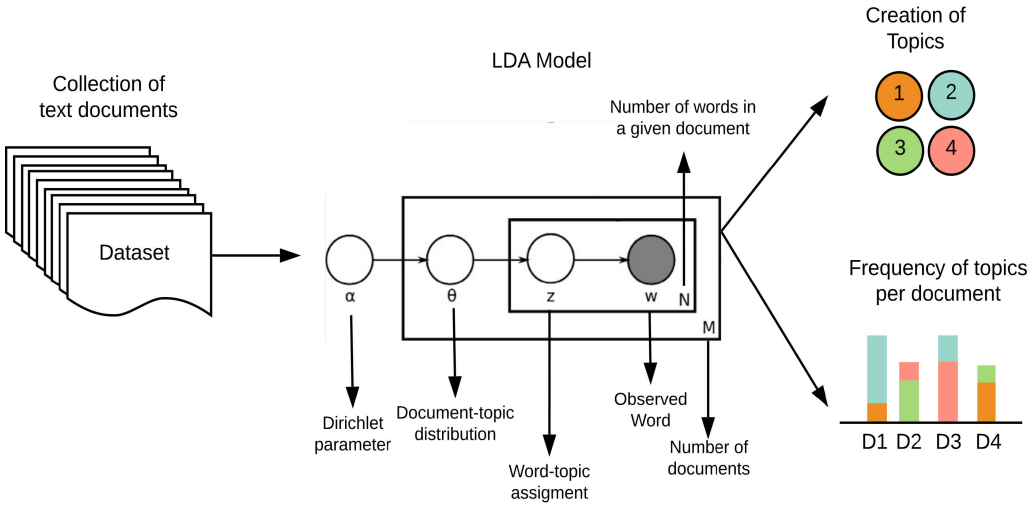


Figure 3.3: Representation of LDA Algorithm
(Buenano Fernandez et al. 2020)

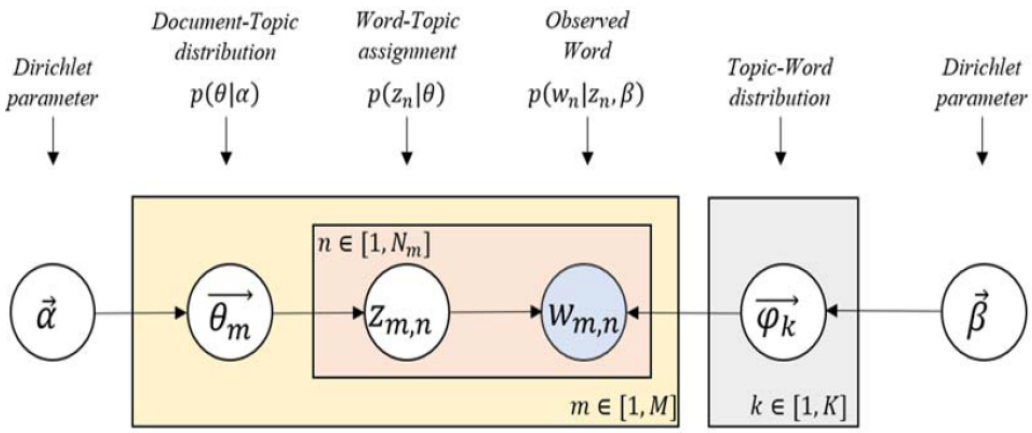


Figure 3.4: Parameters of LDA Algorithm
(Buenano Fernandez et al. 2020)

From 3.5 We can see that on the left-hand side (LHS) is the probability that a document will appear, and on the right-hand side, we have four factors: 'distribution of topics over terms,' 'distribution of documents over topics,' 'probability of a topic appearing in a certain document,' and 'probability of a topic appearing in a certain topic.' The first two are Dirichlet distributions, which are settings of the machine, and the last two are multinomial distributions, which are gears of the machine. Each of them consists of probabilities, and when these probabilities are multiplied, we will get the total probability of a document.

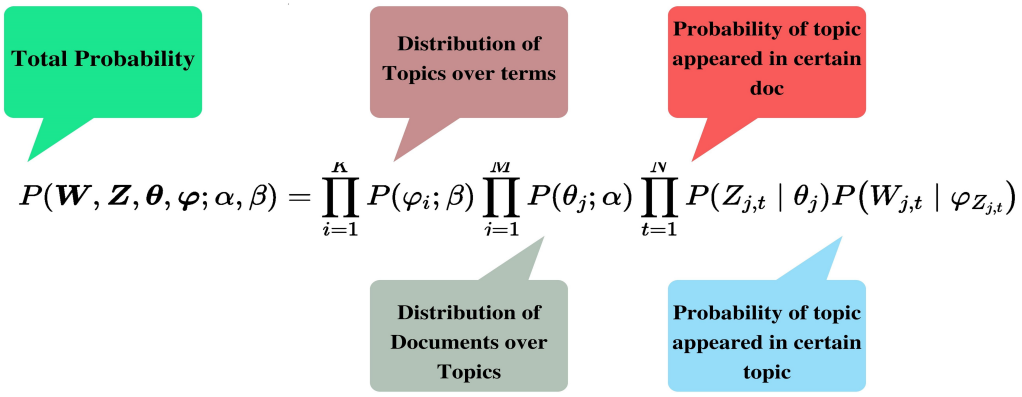


Figure 3.5: Calculation of Probability in LDA
(Buenano Fernandez et al. 2020)

A group of words or documents can be considered a document term matrix, or

Table 3.5: Document Frequency Matrix

Documents/Word	Word1	Word2	Word3	Word4	Word5	Wordm
Document 1	3	0	1	2	0	1
Document 2	0	2	1	0	2	0
Document 3	1	1	0	3	1	2
Document n	2	0	1	1	0	0

DTM. At the very initial stage, LDA is presented with unstructured documents that contain a mixture of words that contain topics. The topics will be generated based on their probability distribution. LDA is considered a matrix factorization technique that includes the matrices of words with their corresponding frequencies in every document. These matrices are very large if the dataset contains a huge amount of textual data. Therefore, It is vital to first clean the documents by excluding all the unnecessary words that do not provide any important context or meaning to the decision-making and tokenizing the words to be represented as individual keywords. With this process, it is possible to reduce the size of the matrix without losing significant knowledge of the data.

To create a matrix, term frequency is calculated for each document. The below matrix 3.5 shows the corpus of N documents D_n and the vocabulary size of m words W_m. W_j * D_i represents the frequency count of words in a document."

LDA converts the term frequency matrix into a lower-dimensional representation, below are example of the two matrices 3.6 and 3.7 : $M1$ (document-topic distribution, $N \times K$) and $M2$ (topic-term distribution, $K \times M$), where N is the number of documents, K is the number of topics, and M is the vocabulary size. In $M1$, each row represents a document, and each column represents a topic, showing the probability of each document belonging to a specific topic. In $M2$, each row corresponds to a topic, and each column corresponds to a term in the vocabulary, showing the probability of each term being associated with a specific topic. It's worth noting that the initial topics are generated, but LDA performs an iterative method to refine and improve the clustering of relevant topics.

The probability is calculated as follows :

Table 3.6: Weighted Document-Topic Matrix

$M1 =$

Documents/Topics	Topic 1	Topic 2	Topic 3
Document 1	0.4	0.2	0.4
Document 2	0.1	0.8	0.1
Document 3	0.3	0.6	0.1
Document 4	0.6	0.1	0.3

Table 3.7: Weighted Topic-Words Matrix

$M2 =$

Topics/Words	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 1	0.2	0.1	0.3	0.2	0.2
Topic 2	0.3	0.2	0.2	0.1	0.2
Topic 3	0.2	0.2	0.1	0.3	0.2

P1 – p(topic t|document d) = The proportion of words in document d that are currently assigned to the topic.

P2 – p(word w|topic t) = The proportion of assignments to topic t over all documents that come from this word w.

3.5.2 TF-IDF (Term Frequency-Inverse Document Frequency)

The TF-IDF technique is a common approach used to highlight important terms and prioritize search results. It measures the significance of a term within a particular document in relation to the larger document collection. It is a natural language processing technique used for information retrieval. This method provides a numerical representation for each term based on its frequency and efficiently assesses the commonness and rarity of the term within a document.

- TF (Term Frequency) for a term in a document:

$$TF(term, document) = \frac{\text{(Number of times the term appears in the document)}}{\text{(Total number of terms in the document)}}$$

- IDF (Inverse Document Frequency) for a term across the document collection:

$$IDF(term) = \log\left(\frac{\text{(Total number of documents in the collection)}}{\text{(Number of documents containing the term)}}\right)$$

- TF-IDF (Term Frequency-Inverse Document Frequency) score for a term in a document:

$$TF-IDF(term, document) = TF(term, document) \times IDF(term)$$

3.6 Visualization with Python

To visualize the trends in conferences and understand the contributions of various proceedings over the past years, we have utilized the Matplotlib library. Python offers the Matplotlib library as a powerful tool for visualization. This library is known for

its flexibility and effectiveness in exploratory data analysis, enabling the creation of interactive 2D plots, including bar plots, scatter plots, line plots, pie charts, and more. Matplotlib operates seamlessly with essential data structures like NumPy and Pandas, allowing us to craft visually appealing representations. Its customization options help users define colors, markers, labels, and titles, enhancing interactivity and engagement. It is easy to save plots in diverse formats, such as PNG and JPEG.

Matplotlib offers a number of techniques for generating data visualizations. Table 3.8 lists some of the commonly used methods along with their descriptions.

Method	**Description**
`plt.plot()`	Creates line plots to visualize data trends.
`plt.scatter()`	Creates scatter plots to explore relationships between two variables.
`plt.bar()`	Creates vertical bar charts to compare categorical data.
`plt.barh()`	Creates horizontal bar charts.
`plt.hist()`	Creates histograms to display data distribution.
`plt.boxplot()`	Generates box plots to visualize data spread and outliers.
`plt.pie()`	Generates pie charts to represent proportions within a whole.
`plt.legend()`	Adds a legend to identify different elements in the plot.
`plt.title()`	Sets the title of the plot.
`plt.xlabel()`	Adds a label to the x-axis.
`plt.ylabel()`	Adds a label to the y-axis.
`plt.xlim()`	Sets the limits for the x-axis.
`plt.ylim()`	Sets the limits for the y-axis.
`plt.grid()`	Adds grid lines to the plot for better readability.
`plt.savefig()`	Saves the current figure to a file in various formats.
`plt.show()`	Displays the plot on the screen.

Table 3.8: Commonly Used Methods in Matplotlib

3.7 LDA Visualization

LDA Visualization is a web-based, user-friendly, interactive tool designed for visualizing topic models in large document collections. It provides a dynamic interface for exploring topics within the corpus. By utilizing computational techniques, it uncovers hidden patterns in the documents, exploiting multi dimensions of statistically related topics (Sievert, and Shirley 2014). This allows for the interpretation of prevalent themes and topics across the corpus.

The tool is implemented using the R programming language and D3.js (Data-Driven Documents), a JavaScript library for binding data to the Document Object Model (DOM). LDA Visualization offers a comprehensive view of multiple topics, providing insights into the corpus as a whole. Beyond topic interpretation, it also displays the relevance of

terms associated with each topic, enabling users to compare and contrast topics based on word distribution and occurrence in various documents.

LDA Visualization seamlessly integrates with Latent Dirichlet Allocation (LDA) algorithm to effectively generate visualizations. This is a significant achievement though there were challenges posed by the high dimensionality of fitted models, resulting in insightful visual representations of the topics.

3.7.1 Components of LDAvis

- **Left Panel: Global View of Topics -** In the left panel, topics are visualized as circles in a 2D plane. Each circle corresponds to a topic identified by the LDA model. The arrangement of these circles is based on the computed distances between topics, where closer circles represent more related topics and distant circles represent less related topics.
- **Center of Circles and Distance -** The center of each circle represents the central theme or "centroid" of the topic, calculated from the average word distribution. The distance between the centers of two circles indicates the dissimilarity or similarity of those topics in terms of their word distributions.
- **Multi-dimensional Inter-Topic Relationships -** Although the visualization is in 2D, it captures the multi-dimensional relationships between topics which were converted into two matrices using term frequency and presented on 2D X-Y plane.
- **Topic Prevalence and Circle Areas -** The size or area of each circle reflects the prevalence or importance of the corresponding topic in the document collection. Larger circles denote more prevalent topics, while smaller circles represent less prevalent topics.
- **Sorting by Prevalence -** The topics are sorted in descending order of prevalence. This arrangement places the most prevalent topics, represented by larger circles, at the top of the visualization, with less prevalent topics arranged towards the bottom.
- **Right panel - Bar Representation -** The visualization utilizes a horizontal bar chart to represent individual terms relevant to the selected topic. Each term is depicted by a bar, with the length or height of the bar indicating the term's importance or relevance within the context.
- **Overlaid Bars -** For each individual term, the visualization employs a pair of bars that are overlaid on one another. One bar signifies the term's frequency across the entire corpus, while the other bar represents the term's frequency within the specific topic. This visual arrangement facilitates a comparative analysis between the term's overall prevalence and its importance within the topic.

Chapter 4

Experimentation

The purpose of this chapter is to present the experimental study. The research was conducted using Python version 3.11.4, and all computations were performed on Google Colab.

The research objectives [RO] of the experimental study are:

- *RO 1:* to understand the trend in the annual count of published papers in the field of artificial intelligence across different conferences?
- *RO 2:* to understand the patterns in the annual percentage distribution of contributions from industry and academia across multiple conferences, depicting their rise and fall over time?
- *RO 3:* to understand which countries are the leading contributors in terms of providing contributions, either in academia or industry?"
- *RO 4:* to understand the specific published papers that stand out with the highest counts of citations?
- *RO 5:* to understand the top trending topics that help us understand latent themes and patterns within the published papers?

4.1 Trends in Academic-Industry Contributions

The process involved in analyzing and understanding the evolution of publication trends across conferences and the dynamic interplay between academia and industry. In this context, below two experiments shed light on these facets.

4.1.1 Experiment 1: The Evolution and Impact of Prominent AI Conferences

The objective of the first experiment is to analyze the publication trends across various conferences. The figure 4.1 showcases the participation of conferences.

Conference Names	Abbreviations
Association for the Advancement of Artificial Intelligence (AAAI)	AAAI
Foundations of Genetic Algorithms (FOGA)	FOGA
International Conference on Computer Vision (ICCV)	ICCV
International Conference on Learning Representations (ICLR)	ICLR
International Conference on Machine Learning (ICML)	ICML
International Joint Conference on Artificial Intelligence (IJCAI)	IJCAI
Neural Information Processing Systems (NeurIPS)	NeurIPS
ACM SIGKDD Conference on Knowledge Discovery and Data Mining	SIGKDD
Conference on Uncertainty in Artificial Intelligence (UAI)	UAI

Table 4.1: Conference Names and Abbreviations

Notably, AAAI, NeurIPS, and IJCAI emerge as the predominant conferences with a consistent history of paper publications. Among these, AAAI and ICML were among the earliest conferences to publish papers. The paper count presented at the ICML conference exhibits a noticeable spike between 1991 and 1998, followed by a significant decline from 1998 to 2002. However, this decline is subsequently overcome, leading to a continuous increase in contributions. In contrast, AAAI demonstrates remarkable consistency in its publication count, with a substantial spike observed between 2015 and 2022. While ICML maintains a steady presence in terms of publication, its count is relatively lower when compared to AAAI and ICLR.

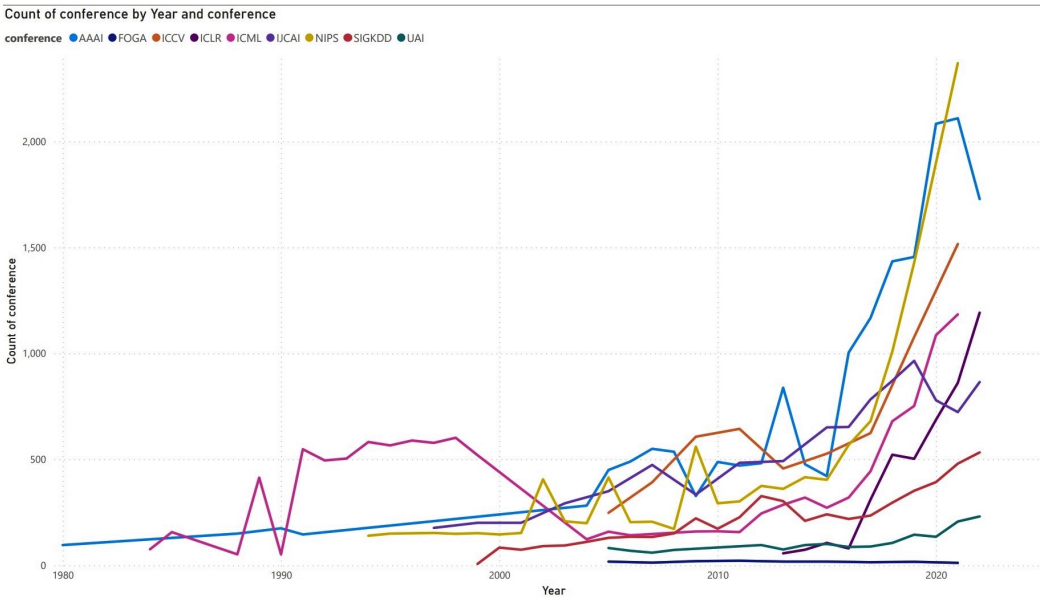


Figure 4.1: Count of conferences across multiple years

The publication count of NeurIPS is experiencing a remarkable surge. From the years 2015 to 2020, NeurIPS and AAAI had a relatively similar publication count. However, in the year 2020, NeurIPS surpassed AAAI's publication count, demonstrating a notable increase in its scholarly output.

The table below presents the total counts for various conferences, highlighting their popularity within the field of artificial intelligence. Among these, AAAI, NIPS, and IJCAI emerge as the top three highest contributors, with AAAI taking the lead with a remarkable count of 17,362 papers. There is a significant gap of 4,204 papers between AAAI and NIPS, and a difference of 8,065 papers between AAAI and IJCAI. Following these, ICML, ICCV, and ICLR contribute to the field, with an average difference of approximately 1,000 papers. This data underscores the prominence of AAAI, NIPS, and IJCAI as major contributors to artificial intelligence research. On the other hand, conferences like UAI and FOGA represent smaller communities focusing on specialized research topics.

Conference	Total count
AAAI	17,362
NIPS	13,518
IJCAI	9,297
ICML	7,088
ICCV	6,092
SIGKDD	5,223
ICLR	4,393
UAI	1,653
FOGA	154

Table 4.2: Contributions by Multiple Conferences

4.1.2 Experiment 2: Academic and Industry Contributions: An In-Depth Analysis

The data in the figure (4.2) covers a span of four decades, from 1980 to 2022. Starting in the early 1980s, we can observe that universities were the dominant participants, accounting for over 60% of collaborations on their own. Around the year 2000, the contribution of organizations began to increase, but it displayed fluctuations afterward.

Collaborations involving both universities and organizations remained relatively stable but with a lower overall contribution. In the 1980s, universities had the highest dominance at 65.6%, organizations accounted for 35%, and mixed contributions made up around 10.4%. In the 1990s, the dominance of universities increased by 7%, while organizations' contributions decreased by 3% and mixed collaborations grew by 5%.

Moving into the 2000s, organizations played a significant role, contributing 31.4%. However, universities had a remarkable contribution of 72.4% in 2003. Mixed collaborations stayed nearly the same as in the 1990s, at 14.5%. In the 2010s, organizational

contributions slightly decreased to 27.4% in 2018, while universities maintained a strong presence with a peak contribution of 70.8% in 2013. Collaborative efforts declined to 8.6% in 2014.

In the 2020s, organizational contributions resurged, reaching their peak at 33.2% in 2021, while university contributions declined to 64.0% in 2022. Mixed collaborations increased to 6.3% in 2019. These shifts illustrate a complex relationship between organizations, universities, and collaborations, showcasing how contributions evolved over these four decades.

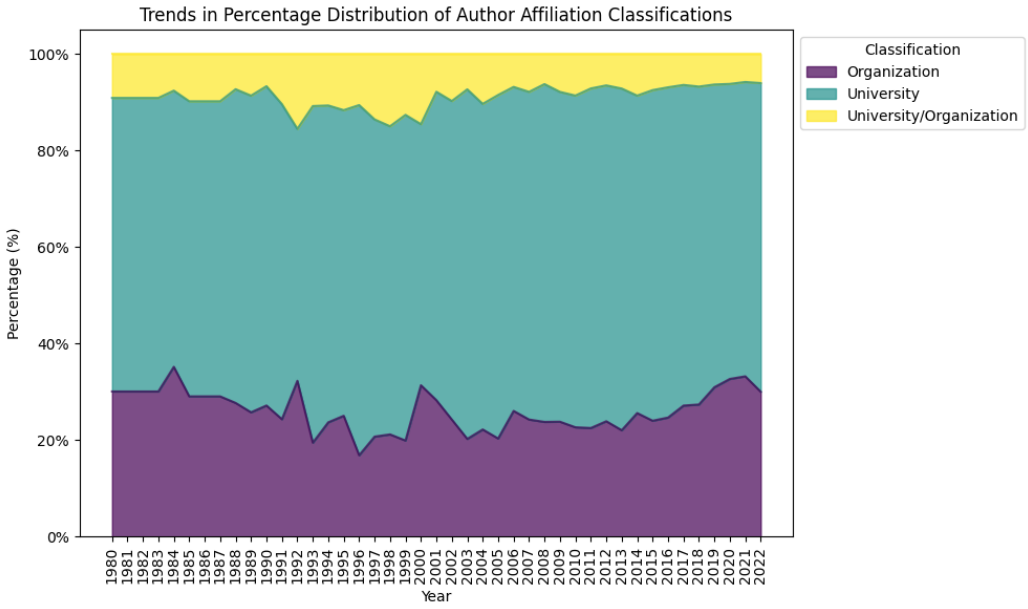


Figure 4.2: Percentage Distribution of Contributions Among Different Conferences for Different Years.

4.2 Exploration of Leading Contributions: Universities and Organizations

The top institutions and organizations in the world are directing innovation and establishing the foundations for future advancements in AI research and development. These fields of industry and academic institutions constantly expand our knowledge. They not only engage in global competition, but they also draw resources and researchers that increase their effect and reputation. We can discover new trends and learn more about the key factors advancing development by tracking the patterns of these contributions over time. In this section, we will explore the fundamental contributions made by prominent institutions and organizations, emphasizing their importance and impact on the larger AI domain. The interaction between academics and industry is highlighted

in this analysis, which also identifies the key players at present.

4.2.1 Experiment 3: Pioneering Contributions from Academic Institutions

The universities that are making the most significant contributions in their field have valuable insights to offer. The universities that rank in the top 10 globally not only compete on an international scale but also attract researchers to put more effort into their work, which in turn enhances their reputation. By keeping track of the changes in these top contributions over time, we can identify emerging trends in the areas these universities are focusing on.

The following diagram depicts the leading 10 contributors on a global scale in the realm of artificial intelligence literature. Among them, three institutions emerge as the most prominent: Stanford University, Carnegie Mellon University, and the University of California, Berkeley, accounting for 19.9%, 18.7%, and 11.8% of contributions, respectively.

While not part of the Ivy League, these institutions have made substantial contributions to advancing the field. Remarkably, the United States stands out as the preeminent contributor, showcasing three of the previously mentioned institutions within its borders. Notably, Cornell University, which does belong to the Ivy League, lags behind its peers in the top 10 with a contribution of 5.8%.

Breaking down the contributions by country, the United States leads the way with 55.25%, followed by China at 28.75%, the United Kingdom (University of Oxford) at 8.3%, and Singapore (National University of Singapore) at 6.2%. These figures underscore the profound impact and widespread engagement in artificial intelligence research within universities.

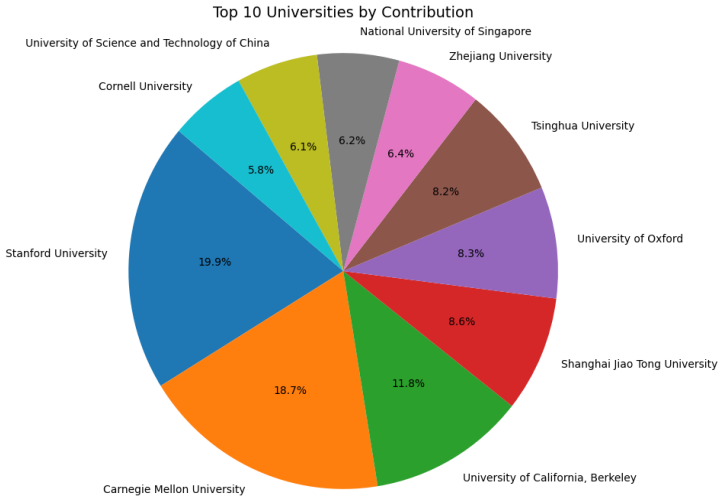


Figure 4.3: Top 10 University Contribution

4.2.2 Experiment 4: Pioneering Contributions from Industries

The ten most significant contributions from the business world provide us with a comprehensive view of how the market operates and how major tech companies invest in it. These insights help us understand the key players in the industry and how much they contribute. This knowledge also gives us an idea of how industries are involved in the AI market. By looking at the competition between these companies, we can identify potential partners for collaborations that can lead to progress and growth in the market.

The figure shows us different aspects of research in artificial intelligence. It highlights the impact of each of the top 10 organizations in this field. When we delve into the data, we see that Google and Microsoft stand out. They have made significant advancements in AI. Google has different parts, like DeepMind and Google Brain, each focusing on different areas. Overall, Google's contributions make up a large portion, 47.6%, of the top 10's total. Microsoft follows with a respectable 14.7%.

Other important organizations include Facebook AI Research with 8.7%, Alibaba Group and IBM Research with 8.4% and 6.6%, respectively, and Adobe Research and Huawei Noah's Ark Lab, both contributing 4.9%. NVIDIA, known for its innovations, has a 4.2% share of contributions.

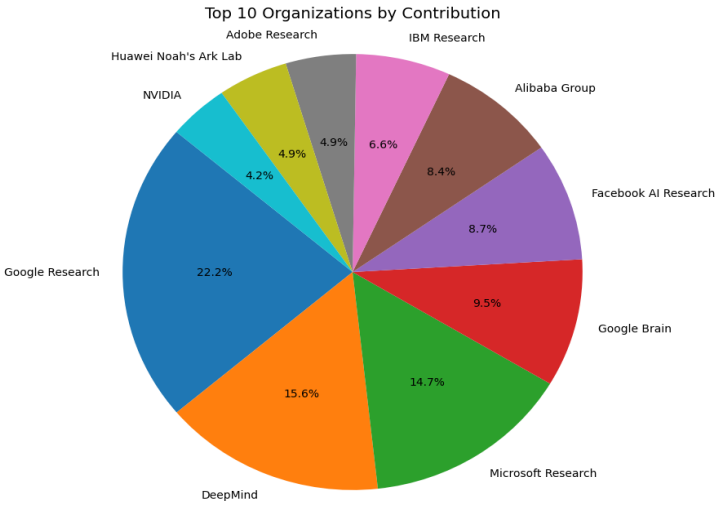


Figure 4.4: Top 10 Industry Contribution

4.3 Understanding Global Contributions and Influential Papers in Artificial Intelligence

4.3.1 Experiment 5: Global Contributions of Countries

From the figure (4.3), it is evident that the United States, China, and the United Kingdom are the most significant contributors globally. The United States stands out with

the highest count at 94,536 overall contributions. Importantly, the academic sector contributes more than the industrial sector in the United States. Similarly, China and the UK also display a dominance of the academic sector over the industrial sector. It's worth noting that there is a substantial difference of 56,989 between the contributions of the United States and China, indicating a significant gap. This substantial lead held by the United States is likely to be challenging for other countries to overcome. This observation suggests that American authors will likely continue their dominance in the coming years.

From the perspective of continents, Europe's contribution, driven by contributions from 12 countries, surpasses that of North America and Asia.

Table 4.3: Statistics by Country/Organization

Countries	Cumulative	Organization	University	Uni./Org.
United States	94536.0	28246.0	56428.0	9862.0
China	37547.0	6642.0	24179.0	6726.0
United Kingdom	13008.0	3210.0	9003.0	795.0
Canada	8999.0	1576.0	6376.0	1047.0
Germany	8447.0	3250.0	3040.0	2157.0
Japan	7282.0	3512.0	2438.0	1332.0
France	6742.0	4203.0	2462.0	77.0
Australia	5461.0	738.0	4341.0	382.0
Singapore	4068.0	405.0	3187.0	476.0
Switzerland	3899.0	2965.0	396.0	538.0
Hong Kong	3650.0	420.0	3127.0	103.0
Israel	3578.0	733.0	1932.0	913.0
Italy	3561.0	1883.0	1489.0	189.0
India	2863.0	1470.0	95.0	1298.0
Netherlands	1723.0	363.0	1128.0	232.0
Spain	1512.0	1139.0	278.0	95.0
Austria	1301.0	446.0	287.0	568.0
Belgium	1083.0	710.0	340.0	33.0
Sweden	870.0	202.0	490.0	178.0
Finland	782.0	60.0	555.0	167.0

4.3.2 Experiment 6: Influential Papers through Citation Analysis

Citation analysis identifies the most frequently cited papers in the field of artificial intelligence. This analysis provides valuable insights and fills knowledge gaps related to

crucial topics in the field of artificial intelligence. The list of these cited papers stands to benefit forthcoming researchers and the academic community.

The citation analysis conducted on the dataset involves determining the highest citation count for each respective paper. This analysis has produced the following figure (4.4), which showcases the top citation counts across four primary categories within artificial intelligence: natural language processing, object detection and recognition, generative models, optimization and machine learning techniques, and computer vision and image recognition.

Table 4.4: Statistics by Country/Organization

Cited Count	Title
71113	ImageNet classification with deep convolutional neural networks
37274	Adam: A method for stochastic optimization
34954	Generative adversarial nets
32955	Attention is all you need
22609	Faster R-CNN: Towards real-time object detection with region proposal networks
20889	Distributed representations of words and phrases and their compositionality
19992	Very deep convolutional networks for large-scale image recognition
17637	Batch normalization: Accelerating deep network training by reducing internal covariate shift
17394	Efficient estimation of word representations in vector space
16781	XGBoost: A scalable tree boosting system
15051	Fast R-CNN
13031	Mask R-CNN
13011	PyTorch: An imperative style, high-performance deep learning library
12318	Rectified linear units improve Restricted Boltzmann machines
11944	Sequence to sequence learning with neural networks
11537	Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification
9967	Focal Loss for Dense object Detection
9516	Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks
7617	Semi-supervised classification with graph convolution networks
7395	Neural machine translation by jointly learning to align and translate

4.4 Exploration of Latent Themes

In this section, our primary aim is to uncover latent themes present within the documents contained in our dataset. This method is especially useful for understanding the

usage of highly trending topics and important areas in the field of artificial intelligence. Our bibliometric dataset encompasses a variety of natural language texts, presented in diverse formats such as abstracts, titles, and index keywords. The latent themes are captured in a broader context within the abstract and titles columns, while the index keywords provide more specific insights, as publishers often encapsulate the precise keywords of concentration within individual papers. Various methodologies can be employed for topic modeling, and among these, Latent Dirichlet Allocation (LDA) stands out due to its adaptability with different sets of hyper-parameters. For the purpose of this study, we have employed the LDA technique specifically.

4.4.1 Experiment 7: Implementing LDA

The resulting table 4.5 showcases ten topics generated through the application of the LDA technique. The LDA algorithm yields a collection of topics, each accompanied by the words relevant to that topic along with their corresponding weights (probabilities).

Table 4.5: LDA Topic Discovery for Titles

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
learning	linear	model	online	causal
data	discovery	time	decision	optimization
networks	classification	markov	systems	sampling
mining	algorithms	processes	gaussian	algorithm
using	active	series	policy	optimal
analysis	dimensional	graphical	regression	state
graph	local	rules	process	approximate
network	real	recommendation	constraints	multiple
neural	self	mdps	methods	patterns
information	matrix	complexity	towards	approximation

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
based	social	multi	models	efficient
bayesian	latent	clustering	via	search
inference	system	approach	detection	sparse
large	text	prediction	knowledge	web
probabilistic	user	kernel	graphs	proceedings
dynamic	relational	event	modeling	applications
non	structural	adaptive	uncertainty	conference
scale	vector	topic	variational	sigkdd
selection	exploration	hierarchical	robust	machine
structure	embeddings	discovering	conditional	privacy

LDA is a statistical method that assigns weights to topics without providing explicit interpretations. The interpretation of these topics is an important phase in the process. The topics presented below in the table 4.6 have been manually interpreted.

Table 4.6: Interpreted Topics for Titles

Topic Number	Interpreted Topic Name
1	Machine Learning and Data Analysis
2	Algorithms and Dimensionality
3	Time Series and Model Complexity
4	Online Systems and Decision Making
5	Causal Inference and Optimization
6	Bayesian Inference and Probabilistic Models
7	Social Networks and Text Analysis
8	Clustering and Prediction Approaches
9	Graph Modeling and Uncertainty
10	Efficient Search and Applications

The figure 4.5 illustrates the connections between different topics that are present in a collection of text documents.

The term "multi-dimensional" indicates a way to represent these connections in a spatial manner, condensing the many dimensions of topics into a two-dimensional or three-dimensional map. In the diagram, we can identify 10 distinct groups of topics that form clusters some of which are co-related to each other.

The horizontal bar chart on the right represents the most relevant terms for the topic, showcasing their respective frequencies. The chart effectively visualizes the top 30 most relevant terms, providing a clear and concise overview of the key concepts and their prevalence within the given context.

The libraries incorporated is the *NLTK* library which serves as a valuable tool for handling natural language text within our dataset. It provide us with an array of text preprocessing capabilities, including tokenization, stemming, and lowercase conversion. Through the *nltk.corpus* module, we can readily incorporate a stop word list, which filters out common words devoid of significant analytical value.

One of the sub-modules within *NLTK*, known as the *RegexpTokenizer*, proves essential for sentence tokenization, effectively dividing sentences into distinct terms. we have also included the gensim library which is a specialized resource for the tasks such as topic modeling and document similarity. Within gensim, two pivotal components, namely *corpora.dictionary* and *model.LDAmodel*, plays an important role.

The Corpora module within gensim facilitates the construction of a word dictionary, compiling words and assigning them unique integer IDs derived from document similarity techniques and the *models.LdaModel* method allows us to create an LDA (Latent Dirichlet Allocation) models from our preprocessed data.

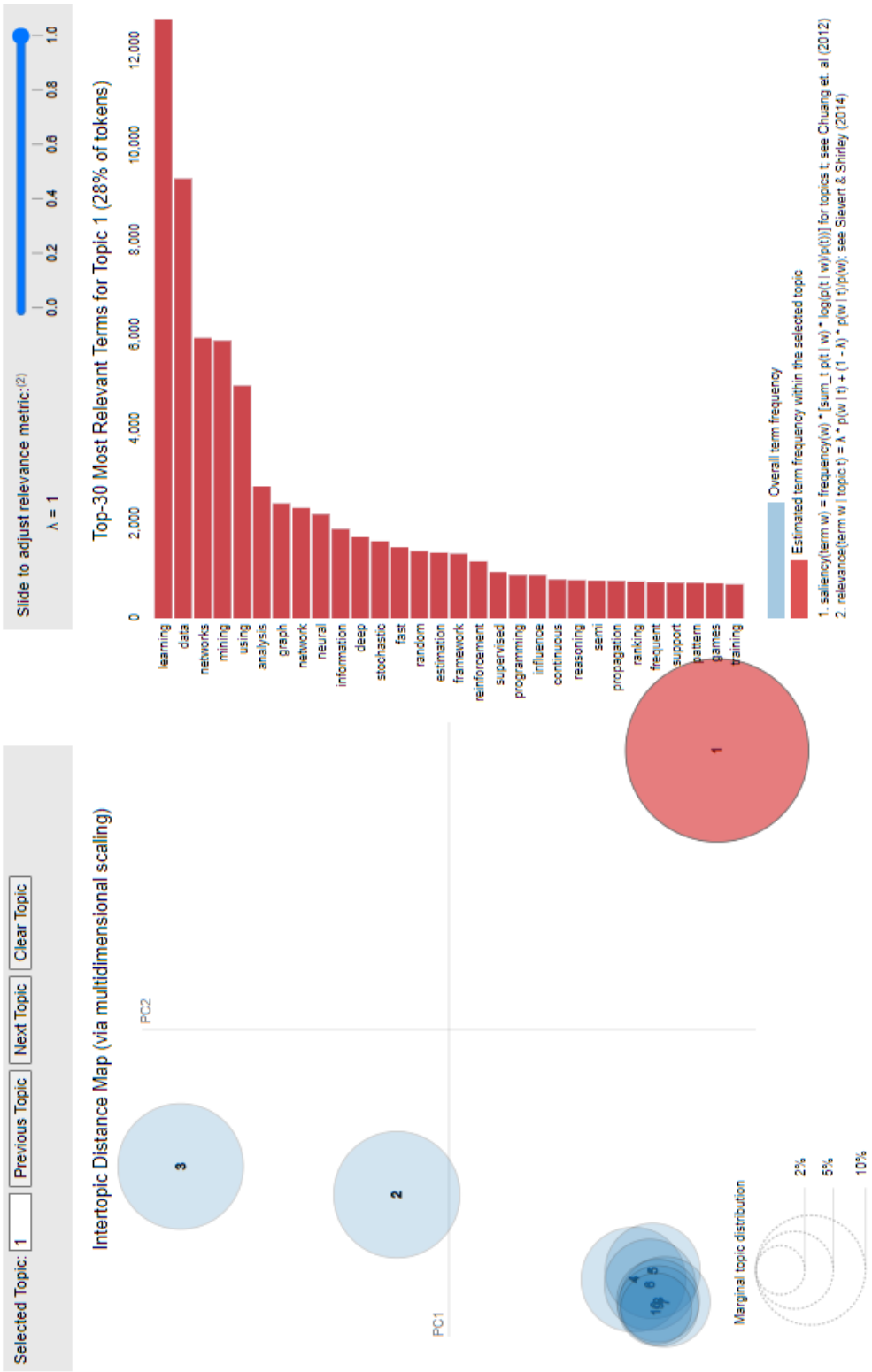


Figure 4.5: PyLDA Visualization for Titles

Similarly, we have applied this technique to index keywords. It's important to highlight that the context of these keywords is much clearer than the context of keywords derived from titles. The table 4.7 illustrates the topics generated by the LDA technique for index keywords and figure 4.6 shows the LDA visualization.

Table 4.7: LDA Topic Discovery for index keywords

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Learning	high	data	clustering	intelligence
Information	regression	networks	linear	artificial
Model	sampling	models	problems	systems
Analysis	dimensional	classification	programming	time
Classification	parameter	neural	matrix	search
Prediction	reduction	machine	database	engines
Neural	scale	analysis	representation	process
Machine	class	analysis	retrieval	series
Deep	commerce	network	solving	research
Algorithm	discovery	algorithms	semantics	algorithm

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
mining	approximation	state	algorithms	inference
graph	online	detection	bayesian	real
knowledge	approximate	feature	optimization	world
based	value	art	methods	multi
graphs	networking	image	markov	structures
control	polynomial	computer	theory	datasets
nan	machines	pattern	processes	structure
empirical	features	recognition	functions	graphical
graphic	belief	extraction	stochastic	causal
likelihood	heuristic	rules	problem	monte

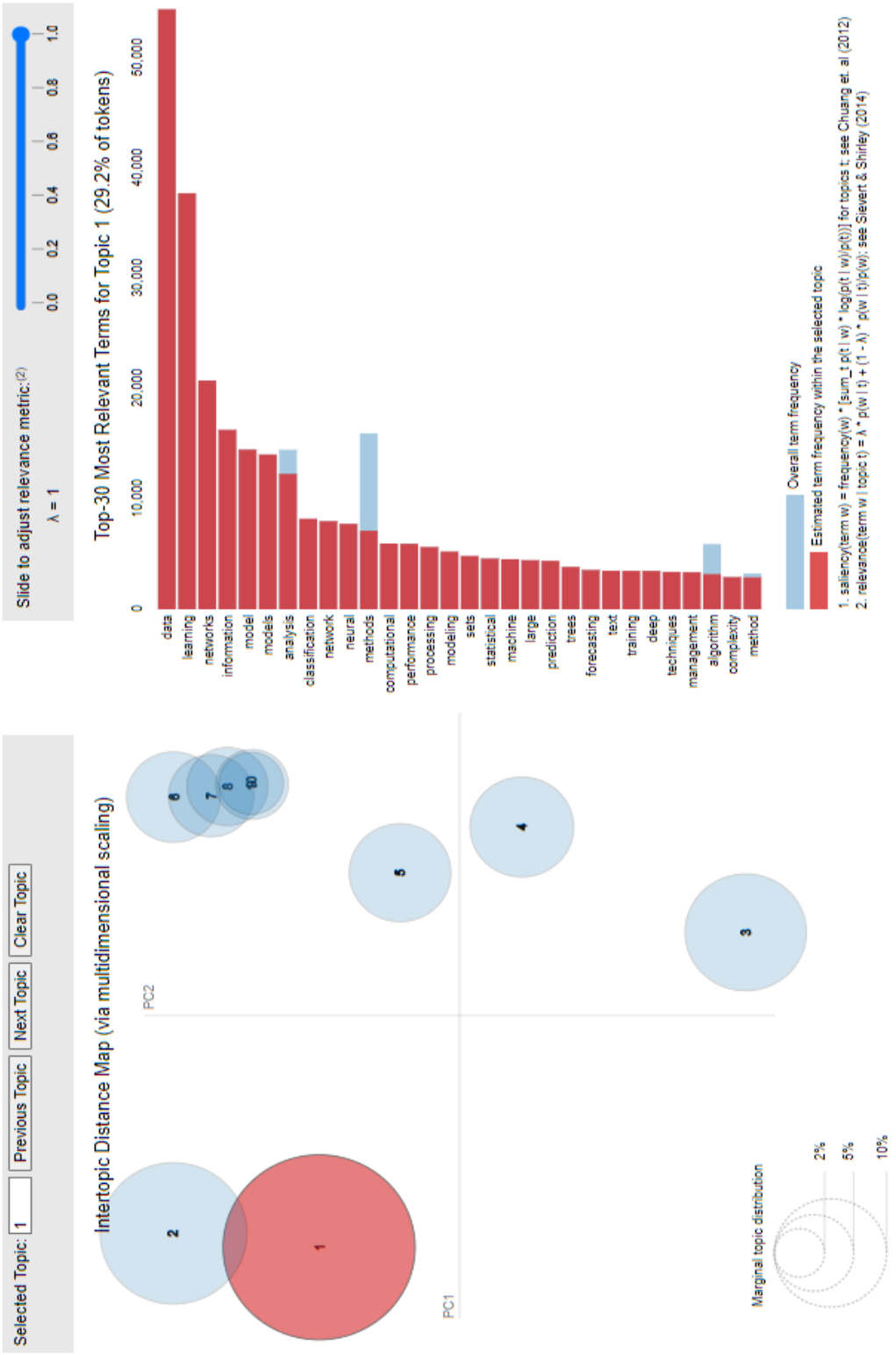


Figure 4.6: PyLDA Visualization for Index Keywords

The below table represents the interpretation of topics of indexed keywords.

Table 4.8: Interpreted Topics for Indexed Keywords

Topic Number	Interpreted Topic Name
1	Learning and Intelligence
2	High-Dimensional Data Analysis
3	Data Networks and Models
4	Clustering and Programming
5	Artificial Intelligence and Systems
6	Data Mining and Graphs
7	Approximation and Networking
8	Feature Extraction and Image Analysis
9	Optimization and Theory
10	Inference and Monte Carlo Methods

In a similar fashion, we have processed abstract text composed in natural language rather than keywords which can be seen in table 4.9. This text has been appropriately tokenized, and stop words have been removed. It's important to highlight that this process was time-consuming due to the substantial amount of text documents that needed analysis. However, the outcomes are valuable, as they offer more insightful information by delving into the greater relevance of highly frequent topics.

Table 4.9: LDA Topic Discovery for Abstract

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Data	Clustering	Market	Detection	Time
Models	Features	Traffic	Image	Approach
Learning	Information	Reward	Objects	Based
Performance	Patterns	Services	Spatial	Using
Networks	Representation	Program	Visual	Paper
Methods	Feature	Machines	Regions	Inference
Advancements	Multi	Forecasting	Object	Set
Dominance	Latent	Services	Mapping	Used
Geographical	Domain	Technology	Recognition	Novel
Research	Multiple	Industry		

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Algorithm	System	Data	Graph	Policy
Algorithms	User	Model	Networks	Agent
Problem	Interaction	Learning	Nodes	Agents
Uncertainty	Social	Methods	Network	Dynamics
Problems	AI	Propose	Structure	Value
Function	Technology	Performance	Embedding	Policies
Show	Platform	Innovation	Relations	Planning
Variables	User behavior	Framework	Attributes	Reinforcement
Linear	Impact	Applications	Topology	Actions
Optimal	Networking	Industry	Relations	Strategy

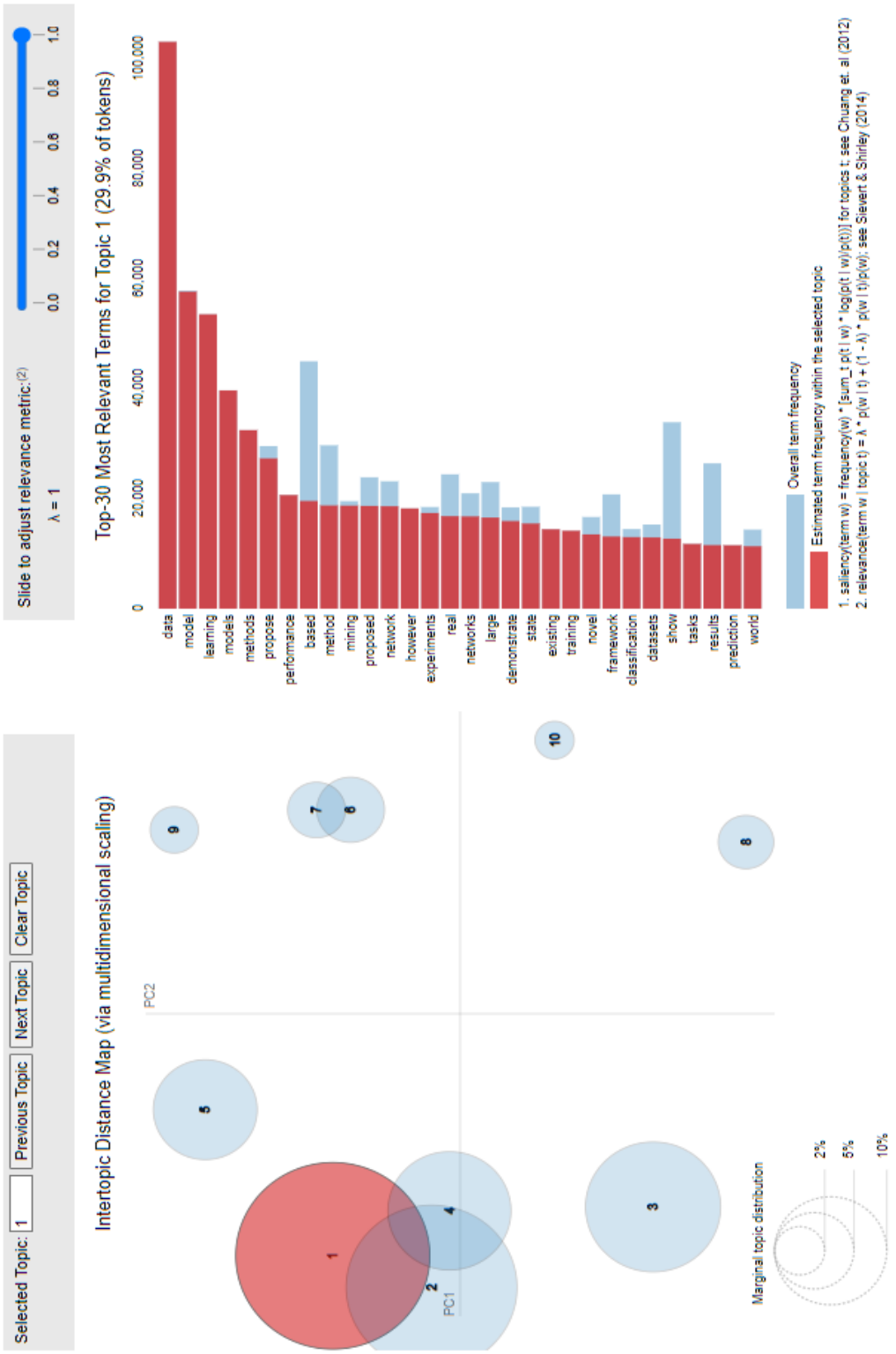


Figure 4.7: PyLDA Visualization for Abstract

Table 4.10: Interpreted Topics for Abstract

Topic Number	Interpreted Topic Name
1	Foundational Concepts in AI
2	Clustering and Pattern Analysis
3	AI Market Dynamics
4	Object Detection and Spatial Analysis
5	AI Research Approaches
6	Algorithmic Problem-Solving
7	Social Aspects of AI
8	Data Modeling and Machine Learning
9	Graph Theory and Network Analysis
10	Reinforcement Learning and Decision-Making

If we compare all three interpreted topics, the most commonly represented topics are machine learning, deep learning, data analysis, clustering, and feature extraction.

The topics inferred from these studies carry substantial influence and hold significant relevance within the specific contexts they are being studied, discussed, or researched. Notably, there is a strong focus on advanced AI techniques and algorithms, particularly within the realms of machine learning, deep learning and neural networks, underscoring the rapid technological advancements taking place.

A noteworthy trend is the shift towards solutions that adopt a data-centric approach, involving thorough data analysis through methods like clustering and feature extraction, among others. This shift is particularly pronounced in the fields of machine learning and deep learning.

This emerging pattern suggests that these key topics are poised to play a pivotal role for aspiring researchers, students, and tech developers in shaping the future of AI. By delving into these subjects, they can actively contribute to the ongoing evolution of AI and drive innovative applications across various sectors.

4.5 Experimental Findings

The essential conclusions from the experimental study are:

1. **AAAI's Leading Role and Growth:** The AAAI (Association for the Advancement of Artificial Intelligence) has demonstrated remarkable growth and dominance in terms of publication count, establishing itself as a primary driving force in AI research.

2. **Emergence of Neuro IPS:** The Neuro IPS (Neural Information Processing Systems) conference has consistently posed a strong challenge to AAAI, cementing its significance within the AI research landscape.

3. **University Contributions:** Universities have played a crucial role in contributing, consistently accounting for over 60% of collaborative efforts in AI research.

Stanford University notably stands out with the highest percentage contribution among institutions.

4. **Industry's Influence:** Both Google and Microsoft have left a significant impact on AI literature, shaping the direction of research and contributing to the dynamic landscape of artificial intelligence.

5. **Global Participation:** The United States takes the lead as the foremost contributor to AI research on a global scale, while Europe also emerges as a competitive player in advancing AI knowledge.

6. **Primary Research Focus:** Machine Learning, Deep learning and neural networks have emerged as the central areas of interest for researchers, accounting all the attention and efforts of a majority within the AI community.

Chapter 5

Conclusion

This study delves into the significant contributions made by big tech companies and universities in the realm of AI research. The data was meticulously collected from various conferences and then analyzed; creating a well-organized dataset was one of the most challenging tasks. The landscape of conferences dedicated to artificial intelligence research has undergone a transformative evolution over the years. Among these, AAAI, NeurIPS, and IJCAI have emerged as the primary contributors, maintaining consistent histories of paper publications. While other conferences have also left their mark on the AI field, their contributions have been more specialized and aligned with specific research interests. As a result, their overall impact has been comparatively considerable.

The dataset also reveals the complex connections between universities and organizations in the field of AI research. Universities' contributions have been showcased, demonstrating their influence over the years. Meanwhile, the influence of organizations has varied across decades, indicating changing research priorities. Collaborative universities have made a noteworthy impact by forming dedicated research committees to advance the field of artificial intelligence. This highlights the strong role of universities in academic literature, whereas technological progress is largely driven by industries.

When we focus on specific institutions, Stanford University, Carnegie Mellon University, and the University of California, Berkeley, emerge as highly influential players. On the other side, organizations like Google and Microsoft take the lead, investing in and shaping the AI market. Their advancements are propelled by various sub-branches such as Google Brain, Google Research, and DeepMind. Other notable competitors in the market include Facebook, Alibaba Group, IBM, Adobe, Huawei, and NVIDIA.

Considering the geographical aspect, both universities and industries are asserting their dominance in the United States. The United States maintains a significant lead over other countries in this field. Although China and the United Kingdom compete, they need to consistently make substantial efforts to establish a strong presence in the realm of artificial intelligence. In terms of sub-fields within AI, the authors of the study are particularly focusing on deep learning and neural networks, as these concepts are highly cited by researchers in the field. This pattern suggests a clear trend: the trajectory of contributions is poised to experience a significant upsurge in the forthcoming years.

5.1 Recommendations for Future Work

The data for this study was sourced from several metadata repositories. However, it's important to note that the internet contains an extensive amount of untapped data, presenting the opportunity for a more precise analysis of our work. Additionally, it's worth mentioning that the selection of conferences included in our study was within the scope of our research. Delving deeper into a wider range of conferences within the field of artificial intelligence could yield more comprehensive insights. Furthermore, potential avenues for further research involve conducting more in-depth examinations of specific sub-fields within artificial intelligence through real-time analysis.

5.1.1 Exploring Artificial Intelligence Sub-fields

Exploring sub-fields within artificial intelligence helps us gain new knowledge and make important discoveries. Analyzing topics like deep learning, natural language processing, computer vision, and machine learning closely can reveal valuable insights. The current era is highly focused on artificial intelligence and models, which are experiencing rapid growth. As a result, sub-fields may incorporate new methods that can be beneficial for new researchers and learners to understand and keep up with trends. The present time is largely dedicated to developing artificial models, leading to frequent discoveries. Therefore, a thorough analysis within sub-fields is necessary.

5.1.2 Real-time Bibliometric Analysis

Incorporating real-time analysis into bibliometric research could provide valuable insights, which involve accessing sources like academic databases, metadata repositories, and authorized databases. Automated pipelines using Python or R scripts can collect and update data continuously. Tools like Tableau and D3.js facilitate real-time visualizations. Stream processing frameworks like Apache Kafka can handle incoming data. Predictive models can forecast publication trends and identify emerging topics. This approach enables researchers to stay current with the rapidly evolving AI field.

Appendix A

Summary of the Techniques used for Bibliometric Analysis

Table A.1: Research Studies and Their Methodologies in Different Sectors

Study	Focus	Data Source	Methodology
Kathiria, and Arolkar 2022	Research trends in Indian computer science	Scopus	Data analysis, forecasting, Tf-Idf matrix, document graph index, clustering (DBSCAN), topic identification (LDA), automatic labeling (CSO), ARIMA model
Ampadi Ramachandran et al. 2023	Depletion of drugs in animals used for food production	Web-crawled databases	Software architecture, ATC classification, TDM techniques, Selenium, API data retrieval, dataframe organization, CSV format
Yong, and Lee 2022	Impact of machine learning in rail industry	Scopus	Data filtering, time series analysis, binary classification, co-occurrence keywords network, Louvain method, citation analysis, collaboration network
Vaio, Hassan, and Alavoine 2022	Emergence of Industry 4.0 and AI integration	Web of Science (wos), Scopus	Literature analysis, bibliometric analysis, PRISMA framework utilization, science mapping, specific search queries
Jimma 2023	Usage of AI in healthcare sector (2000-2021)	Scopus	Structured search, analysis of AI-related healthcare journals, growth analysis, country contributions, VOS software visualization

Table A.2: Research Studies on Web Data Extraction and Text Mining

Study	Focus	Data Source	Methodology
Liu, Zhu, and Guo 2021	Web data extraction using web scraping techniques	Internet source	Web scraping framework, HTTP/HTTPS connections, XML/HTML/CSS parsing, Beautiful Soup, SERP API
Moral Munoz et al. 2020	Bibliometric and scientometric analysis tools	Databases (Scopus, PubMed, etc.)	Bibliometric and scientometric analysis tools, R and Python libraries, Web of Science, Scopus, Google Scholar, Microsoft Academic, Dimensions
Dastani et al. 2020	Text mining for identifying emerging keywords	PubMed and Scopus databases	Text mining, Porter root algorithm, TF-IDF, data preprocessing, stemming, word cloud visualization

Appendix B

Hyperparameter Examined for the Topic Modeling

The below table B.1 shows the hyperparameters used in the LDA model for topic modelling.

Hyperparameter	**Value**
Tokenizer	`RegexpTokenizer(r'\w+')`
Stop Words	English Stopwords from `nltk`
Min Word Length	3
Number of Topics	10
Random State	100
Update Every	1
Chunksize	100
Passes	10
Alpha	"auto"

Table B.1: LDA Hyperparameters

Appendix C

General Flow Of The Analysis

The below figure C.1 represents the general flow of the idea of how the data has been extracted, analyzed, and visualized.

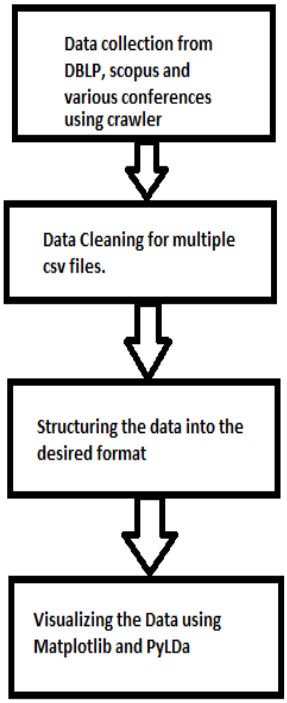


Figure C.1: General Flow of the process

Bibliography

Ampadi Ramachandran, Remya et al. [2023]. "An Automated Customizable Live Web Crawler for Curation of Comparative Pharmacokinetic Data: An Intelligent Compilation of Research-Based Comprehensive Article Repository". In: *Pharmaceutics* 15.5. ISSN: 1999-4923. DOI: 10.3390/pharmaceutics15051384. URL: https://www.mdpi.com/1999-4923/15/5/1384.

Buenano Fernandez, Diego et al. [Feb. 2020]. "Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach". In: *IEEE Access* PP, p. 1. DOI: 10.1109/ACCESS.2020.2974983.

Dastani, Meisam et al. [Nov. 2020]. "Identifying Emerging Trends in Scientific Texts Using TF-IDF Algorithm: A Case Study of Medical Librarianship and Information Articles". In: *Health Technology Assessment in Action* 4. DOI: 10.18502/htaa.v4i2.6231.

Ferrara, Emilio et al. [2012]. "Web Data Extraction, Applications and Techniques: A Survey". In: *CoRR* abs/1207.0246. arXiv: 1207.0246. URL: http://arxiv.org/abs/1207.0246.

Jimma, Bahiru Legesse [2023]. "Artificial intelligence in healthcare: A bibliometric analysis". In: *Telematics and Informatics Reports* 9, p. 100041. ISSN: 2772-5030. DOI: https://doi.org/10.1016/j.teler.2023.100041. URL: https://www.sciencedirect.com/science/article/pii/S2772503023000014.

Kathiria, Preeti, and Harshal Arolkar [2022]. "Trend analysis and forecasting of publication activities by Indian computer science researchers during the period of 2010-23". In: *Expert Systems* 39.10, e13070. DOI: 10.1111/exsy.13070. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13070. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13070.

Liu, Xiaoyu, Donghua Zhu, and Ying Guo [2021]. "Exploring the role of companies in scientific research: a case study of genetically modified maize". In: *Technology Analysis & Strategic Management* 33.4, pp. 349–364. DOI: 10.1080/09537325.2020.1814237. URL: https://doi.org/10.1080/09537325.2020.1814237.

LLC, ARK Investment Management [2023]. *Big Ideas,2023*. January 31, 2023.

Maslej, Nestor et al. [2023]. "The AI Index 2023 Annual Report". In: p. 386.

Moral Munoz, Jose et al. [Jan. 2020]. "Software tools for conducting bibliometric analysis in science: An up to date review". In: 29. DOI: 10.3145/epi.2020.ene.03.

Ruchitaa, Raj N, Raj S Nandhakumar, and Murugesan Vijayalakshmi [2023]. "Web Scrapping Tools and Techniques: A Brief Survey". In: *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pp. 1–4. DOI: 10.1109/ICITIIT57246.2023.10068666.

Scrapy, Team [2023]. *Scrapy data flow architecture*. Dataflow. URL: https://docs.scrapy.org/en/latest/topics/architecture.html [visited on 08/04/2023].

Sievert, Carson, and Kenneth Shirley [June 2014]. "LDAvis: A method for visualizing and interpreting topics". In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 63–70. DOI: 10.3115/v1/W14-3110. URL: https://aclanthology.org/W14-3110.

Stefan van Duin, Naser Bakhshi [2018]. *Artificial Intelligence*. Deloitte. URL: https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/deloitte-analytics/deloitte-nl-data-analytics-artificial-intelligence-whitepaper-eng.pdf.

Vaio, Assunta Di, Rohail Hassan, and Claude Alavoine [2022]. "Data intelligence and analytics A bibliometric analysis of human Artificial intelligence in public sector decision making effectiveness". In: *Technological Forecasting and Social Change* 174, p. 121201. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2021.121201. URL: https://www.sciencedirect.com/science/article/pii/S004016252100634X.

Yong, Gunwoo, and Ghang Lee [2022]. "Trends, Topics, Leaders, Influential Studies, and Future Challenges of Machine Learning Studies in the Rail Industry". In: *Journal of Infrastructure Systems* 28.2. Cited by: 1. DOI: 10.1061/(ASCE)IS.1943-555X.0000691.