Web Scraping for Scientific Discovery: Strategies for Secure Data Retrieval, Structured Transformation, and Relevant Content Selection*

***Asfandyar Ahmed* , Muhammad Ayub Khan, Dr Atif Ishtiaq**

[1]Computer Science Department, Iqra National University, Peshawar,Pakistan; [2]Computer Science Department, Iqra National University, Peshawar, Pakistan. [3]Computer Science Department, Iqra National University,
Email: asfandyarahmed@inu.edu.pk
Email: ayub@inu.edu.pk
Email: atif.ishtiaq@inu.edu.pk

ABSTRACT

This research article presents a comprehensive exploration of web scraping techniques for the automatic collection and analysis of scientific literature from prominent web repositories like Google Scholar. The primary objectives include transforming unstructured web data into a structured format, accommodating secure HTTPS and Ajax-enabled websites, and identifying best practices to bypass security measures effectively. A key focus lies in employing a locality-sensitive hashing text similarity algorithm to discern the most pertinent research papers. By achieving these objectives collectively, this study offers valuable insights for researchers, librarians, and data enthusiasts seeking efficient methods to extract and curate relevant scholarly content from the web. These findings have significant implications in streamlining information retrieval processes and ensuring that researchers can access the most crucial literature in their respective fields while adhering to ethical and legal considerations.

Keywords : Web Scraping; Scientific Literature; Google Scholar; HTTPS Websites;Ajax-Enabled Sites; Locality Sensitive Hashing

1 INTRODUCTION

TYpical review that is literature are the ones that work on uncovering what is already understood in the body of knowledge, aiming educational debates towards advanced advancements. These procedures tend to be very important towards starting any considerable scientific tests. These processes had been around for many years to help researchers through the rules of scholastic learnings [1–3]. The execution can still be classified as daunting despite the benefits attained from effective methodological processes. Academic authors, especially when literary works this is certainly performing, suffer with a few issues. The following are between the conditions that tend to be top faced [4]:

1. Absence of consistency when one looks at the reported results
2. Potential towards uncovering defects within previous study (predicated on design, data collection instruments, sampling, interpretation, etc.)
3. Research may have already been performed on various information populations, which could trigger doubt about explanation of past scientific studies' results Etc.

This departs a few available concerns within the minds of educational article writers, which then reveal some of the main qualities of academic journals, hence hindering high quality for example. These characteristics are: level and breadth, rigor and persistence, clarity and brevity, furnishing the beds base for effective evaluation and synthesis [5].

To modify the running headings, select View | Header and Footer. Click inside the text box to type the name of the journal the article is being submitted to and the manuscript identification number. Click the forward arrow in the pop-up tool bar to modify the header or footer on subsequent pages.

1.1 Background and Motivation

In the field of scholarly examination, information assumes a significant part: a scientist, market examiner, or scholar gathers information from various sites for a better turn of events. Researcher face difficulty in accessing data for their research. Today, increasing digitalization processes of body of knowledge has led to zettabytes (billions of gigabytes) of data available on the World Wide Web (Web). This data provides real-time representations of many processes, relationships, and communication in the literature. Therefore, these large Webdata bases provide academic researchers with data collection opportunities to answer new and old research questions with robustness, accuracy, and time and improve organizational performance. In addition, experts can use Web data by developing better customer perceptions and developing better strategies based on these findings. Web scraping can fit in very nicely in this process.

Web scraping is the process of automatically extracting web data instead of copying it manually. The interaction by which coherent information is extracted from web sites and put away in a nearby data base. It is done with web scrapers with the assistance of uncommonly coded programs. It may have been traditionally compiled for an accurate website or efficiently designed to work with any site.

The process of the Web scraper focuses on the collection of random data, while 70% of the facts promoted on the Internet are found in PDF documents that are random and difficult to manage. Additionally, a web page is a structured format (containing HTML code) is a structured format. To ensure this text's possibility (HTML documents), its structured nature produces the potential expressed through scrubbing processes. Web-based removal tools and tools rely on the organization and resources of HTML language. This will empower the errand of scratching apparatuses and robots, thus empowering people without the tedious tasks of physically accessing data. In conclusion, these tools jeopardize information in available organizations for cutting edge time and combination: JSON, XML, CSV, XLS or RSS [1].

Semantic Web addresses current web problems by editing content on the Web, incorporating semantics, and extracting the maximu m benefits of web and web processing power. As explained in [2], "The semantic web is an extension of the current web In the information provided by a well-defined definition, which allows better computers and people to work together" [1]. It is an idea: the concept of having data on the Web is defined and connected in such a way that it can be used by machines not only for display purposes but by automation, integration, and reuse of data in various applications [2]. Web Mining plays a crucial role in achieving this as it can quickly and easily find the information we need. Web mining means the discovery and analysis of useful information on the Wor ld Wide Web. It is primarily a source of helpful information and information from many web pages and can be considered continuous data mining on the Web, automatically drawing, measuring, analyzing, and explaining data [3].

There are various types of web mining where web digging is a worsening process used to obtain automatic information access for user access patterns from multiple web servers. Web usage mines are defined as removing logical user patterns from logs to access a web server using data mining techniques [4]. Web scraping is the procedure of collecting helpful information from the web automatically [5]. It has been rewritten as a web data extraction and extracting useful information from HTML pages in various ways. It may be done as a content rendition for UNIX initially and can utilize a prearranging language known as Prolog Server Pages (PSP) in light of Prolog language where PSP is implanted in HTML language to erase HTML pages.

A significant element of the semantic site is incorporating proper design and semantics in the current or less organized substance of the current Web, and semantic comment is an approach to make the data work better or more effectively. The semantic modifier gives a more exact meaning of the data contained in the content and is semantics behind the scenes [6] [7].Information examination is an approach to separate answers for issues by posing inquiries and deciphering information. The investigation interaction has problems recognizing issues, tackling admittance to applicable information, figuring out which technology can help track down an intriguing issue arrangement and pass on the outcome. For scientific purposes, data should be classified into various strides by beginning like its coordination, altering, cleaning, re- examination, application models and calculations, and the outcome. The web scrubber instrument is utilized for data from the website admin, and as a feature of the utilization of web orders, electronic and information mining, online acknowledgment changes, and subsidiary checking, filtering a component (seeing test), gathering world entries, climatic information testing, site page exchanging, testing, online closeness and notoriety, web mashup and, online information joining [2].

2 LITERATURE REVIEW

Research papers and their writers, however, need certainly to proceed through an experience that is exhaustive continue to pace utilizing the ever-changing needs of this publication committees. Ergo, there was need that is considerable an program that could allow researchers to quickly download, sort and keep maintaining their papers and magazines with minimum work from the individual end. The inspiration for this development is described as an exponential development in study output from the scientific neighborhood within a globe where technical development forms the forefront of our lives [8].

Main goal of Web Scraping is always to draw out information from a single or web sites which can be numerous procedure it into quick frameworks such as for instance spreadsheets, database or CSV file. Nonetheless, along with be an extremely task this is certainly difficult Web Scraping is resource and time- consuming, mainly when it's performed manually [9]. Earlier studies have created several solutions being computerized. The goal of this article is always to revisit different Web that is existing scraping, categories, and tools, but in addition its aspects of application [10].

Scientific web repositories tend to be central cyber areas where documents which can be academic saved and maintained. Using the nature associated with unstructured and information/metadata this is certainly se mi-structured these repositories, literary works analysis for scholar

writing becomes a challenge. Correspondingly, applying CRISP-DM presents a stance to address this challenge through formulating a rather augmented procedure for the literature search this is certainly relevant. Nonetheless, nearly all repositories never forward have a right strategy where metadata could possibly be removed for initial information handling being applied included in the CRISP-DM process [11].

Additionally, many repositories usually do not follow accessibility that is open. This report ended up being published, the main topics the augmented, relevant literature search had seen a methodological development just, using the inability to use the root practices on a bigger scale, offered information access limitations to open access repositories until the time [12]. The World of Web Scraper, Web scraping is linked to internet indexing, whoever task will be list information on the net with the help of a web or bot crawler. Right here the aspect that is legal both positive and negative edges are taken into view.

Some instances about the issues that are legal additionally taken into consideration. The Web Scraper's creating concepts and methods are compared, it tells just how an operating Scraper was created [13]. The implementation is divided into three components: the Web Crawler to fetch the desired links, the data extractor to fetch the information through the backlinks and storing that data right into a csv file. The Python language is employed for the execution. On incorporating all these with the great knowledge of libraries and deali ng experience, one can possess a scraper that is fully-fledged.

Because of community this is certainly vast collection support for Python as well as the beauty of coding form of python language, it's the most suitable for scraping information from websites [14]. The information and knowledge that is standard are made on the root and effect commitment, shaped an example minuscule evaluation, subjective and quantitative examination, the rationality approach of making extrapolation assessment. The Web Scraper's conniving ethics and treatments are juxtaposed, it explains in regards to the doing work of how a scraper is premeditated [15]. The means of its allocated into three fragments: cyberspace scraper draws the specified links from internet, and then the information is removed to get the data through the origin links and finally stowing that data in to a csv file.

The Python language is implemented for the carrying aside. In that way, connecting all of these aided by the ethical understanding of libraries and working knowledge, one could have a satisfactory Scraper in one hand to make the effect this is certainly desired. As a result of a community that is enormous collection resources for Python as well as the exquisiteness of coding elegant of python language, it really is most suitable one for Scraping desired data through the desired site [16].A massive information in World Wide Web and social media marketing features open opportunities for business and company to get the price this is certainly considerable contributes to efficient operations.

Because of this, Web Data Extraction is a device this is certainly crucial gathering and translating semi- structured papers into important information [17]. Nonetheless, one of many difficulties which are major working with changes from web papers, specially rising of JavaScript Web development technology that has notably impacted the way to embed and making data of Web pages. In this paper [18], researchers suggested a design and utilization of a Web this is certainly brand-new Data system that intends for remove data from JavaScript web programs. The recommended system allows people to choose information which are valuable web papers by defining information removal guidelines and information transformation patterns. The extraction engine automatically scrapes and transforms data being semi-structure relational data. The evaluation that is initial showed that one suggested system has successfully extract data from modern JavaScript Web programs [19]. Huge amounts of information tend to be created by various people, entities, and applications and disseminated online. This volume that is copious of data is distributed across an incredible number of internet sites and is available for various programs. Search motors do provide a device this is certainly simple accessibility this information. Opening this information search that is using takes a user to expend time and resources to manually click and download [20].

Clearly, this type of method that is manual maybe not scalable for the great majority of actuality applications in the enterprise and organization degree. There occur a true wide range of automatic ways to information removal from the net. Many of these approaches are domain and ad-hoc certain. Therefore, the need for a sturdy, automatic, user friendly framework for extracting content from the net through a minimal work this is certainly personal domains seems enticing [21]. The entire process of examining an offered pair of data by examining, cleaning, modeling and transforming is known as Data Analys is.

To interpret and evaluate data this is certainly wealthy, information researchers make use of the methods of data analysis to draw out important information or insights from data in different forms. This is a procedure of obtaining information being raw then transforming it into understanding helpful for people to make choices [22].Different levels of data analysis are data collection, processing, cleansing, evaluation, and modeling. A data which can be raw be acquired from vario us sources like interviews, newspapers, magazines, surveys, the Internet, etc. The Web is amongst the biggest sourced elements of obtaining fixed information which are raw is easily offered. Hence, to draw out data from the oddly structured internet world, web strategies being scraping web wrappers, HTTP programming, etc. are employed [23].

It involves data which can be gathering web pages or webpages and removing information as a result. This can include internet crawler and information extractor. Web crawler crawls all the backlinks contained in an internet page and stores them within a database whereas information extractor extracts data from the stored backlinks. The target that is main of web scraping is to collect, store and analyze data [24]. Whenever amount of accessible resources of information is much as well as unlimited, the data procedure that is retrieval very difficult. Visiting information resources 1 by 1 and researching data or information from all the details resources checked out will add time that is much the entire process of rediscovering the information and knowledge [25].It will require an approach that will gather information from numerous resources as an entity this is certainly single enhance the process of information retrieval.

The study [26] utilizes 3 sites that are e- commerce a source of information. Through the use of the procedure this is certainly crawling create brand new factors that will keep information from the origin information, which in turn these data is likely to be stored in a database. Web crawling works by taking HTML tags as required, using methods being scrapping [27]. Web scraping may be the way of spontaneous assort-

ment of information through the global World Wide Web. It is an arena with vigorous advancement which shares a goal with semantic internet sight grounded for a device that traverses with abstraction of website constituents and protecting them in a data base that is local. It is seen that, commencing a point of view this is certainly appropriate internet scraping is contrary to the terms of use in few web sites: courts are organized for the preservation of authorized contents of commercial sites from objectionable usages even though level of protection of these contents is not obviously established. Subsequently two dissimilar explanations for internet scraping are designated: initial one is formerly obtainable and castoff specifically because of this test, whereas the succeeding one is still within the development phase [28].

Removing information this is certainly useful the internet is considered the most significant problem of issue when it comes to realization of semantic web. This might be accomplished by a few techniques among which Web Usage Mining, Web Scrapping and Semantic Annotation plays a role that is essential. Web mining makes it possible for to learn the appropriate results from the net and it is used to extract information this is certainly meaningful the development patterns kept back in the hosts. Web consumption mining is a kind of internet mining which mines the offered information of accessibility routes/manners of users going to the internet sites [29]. Web scraping, another strategy, is really a procedure for removing information that is of good use HTML pages that might be implemented using a scripting language known as Prolog Server Pages(PSP) based on Prolog. 3rd, Semantic annotation is really a technique rendering it feasible to incorporate semantics as well as a formal construction to unstructured textual documents, an important aspect in semantic information removal which may be performed with a tool referred to as KIM(Knowledge Information Management) [30]. Data tend to be omnipresent in the Internet. Looking the internet for of good use information and information has changed into a task this is certainly routine.

The data in the web sites are located in tables, articles, responses, nested in various HTML tags, etc. Gathering a great deal of information from the web isn't a task that is simple but it is a great way to gather information which is often utilized in further analyzes [31]. The total amount of analysis documents posted has actually dramatically increased over the past several years. Consequently, scientists spend a lot of time reviewing literature this is certainly appropriate order to better understand their particular domain of great interest and maintain new improvements. After doing literary works reviews in the area of text mining, one discovered works being numerous the ways sentence representation in machine understanding for finding sentence similarity. These include typical bag of words, fat word that is average, case of n-grams, and matrix-vector functions. Nonetheless, these techniques are limited in term ordering and evaluation that is semantic [32].

Data mining is the understanding breakthrough process which analyses the large volumes of data from various aspects and summarizing it into of good use information; information mining is actually an essential and component this is certainly essential various fields of everyday life. It really is utilized to recognize hidden design a data being big is an important data mining method with wide applications to classify various kinds of data used in virtually every industry of person life [33]. Searches for grey literature can need sources that are substantial undertake however their addition is important for study tasks such as for instance organized reviews. Web scraping, the extraction of patterned data from website pages on the web, is developed when one look at the industry this is certainly personal company functions, however it offers advantages that are significant those trying to find grey literature [34]. By building and sharing protocols that extract search engine results and other data from web pages, those finding. Grey literary works can increase their transparency significantly and site performance.

Numerous options exist in terms of web-scraping software and they're introduced herein [35].

The Internet may be the vastest information and repository ever before built by mankind. The World Wide Web contains all sorts of information various origins; several of those are social, financial, security and academic. Most people accessibility information over the internet for academic functions. All about the internet comes in different platforms and through various accessibility interfaces [36]. Therefore, indexing or processing that is semantic of data through web pages might be difficult. Web Scraping may be the method which is designed to address this matter. It's the means of extracting information on the internet through different internet scraping tools and technologies. Web scraping is used by many people industries to quickly collect information not available various other formats [37].

Web scraping can be used to transform unstructured data available on the web into structured information that may be stored in a central database that is local spreadsheet and that can be easily examined. The usages of web scraping will be in numerous industries, one may be a reporter, working on a fresh story, or even a data scientist extracting a dataset that is brand-new.

It is also beneficial for climate data monitoring, website change detection, research, internet data integration, contact scraping and cost comparison this is certainly online [38]. The Internet provides a quantity that is huge of data that will be typically formatted because of its people, which makes it tough to draw out relevant information from numerous sources. Therefore, the accessibility to robust, versatile Information removal (IE) methods that transform the Web pages into program-friendly frameworks like a relational database becomes a requisite this is certainly great [39].

Although many approaches for data extraction from Web pages being developed, there has been limited energy to compare tools which can be such. Unfortunately, in just several cases can the results produced by distinct tools be straight compared because the extraction this is certainly addressed will vary [40]. Web Data Extraction is a proble m that is important has been studied by way of various scientific tools plus in an extensive range of application domain names. Numerous methods to information being removing the Web being designed to resolve specific issues and operate in ad-hoc application domains.

Other techniques, alternatively, heavily reuse strategies and formulas created in the field of Information Extraction [41]. This survey in paper [42] is aimed at offering a structured and overview that is comprehensive of analysis efforts made in the world of Web Data Extraction. The fill rouge of one work is to supply a classification of present methods with regards to the programs which is why they have been used. This differentiates one work from other studies dedicated to classify techniques which can be present the cornerstone regarding the algorithms, practices and resources they normally use [43].

One categorized Web Data removal approaches into groups and, for every single group, one illustrated the basic practices with their variants which can be main. One grouped programs which can be existing two main areas: applications during the Enterprise amount and at the Social Web level. This type of category relies on a reason why is twofold on o ne side, Web Data Extraction methods emerged as a crucial device to execute data analysis running a business and Competitive Intelligence systems as well as for business procedure re-engineering [44]. On the other hand, Web Data Extraction methods enable gathering a great deal of organized information continuously created and disseminated by Web 2.0, Social Media and Online Social Network people and this provides unprecedented possibilities of examining human actions for a scale this is certainly big.

One talked about additionally about the potential of cross-fertilization, i.e., regarding the possibility of re-using Web Data Extraction practices initially designed to operate in a given domain, various other domains [45]. Access to amount this is certainly huge of sources on the web is limited by searching and searching due to the heterogeneity together with not enough framework associated with the web information sources [46]. It has resulted in the necessity for automatic Web Information removal (IE) tools that evaluate the Web pages and harvest of good use information from loud content for almost any analysis that is more.

The aim of this review [47] is provide a review that is extensive of major Web IE resources which used for Web text and based on Document Object Model for representing the internet pages [48]. Text mining is a procedure that is semi-automated of real information from a large amount of unstructured information. Considering the fact that the total amount of unstructured data being generated and saved is increasing quickly, the necessity for automated way to also process its increasing. In this scholarly study [49] one presents, negotiates and assesses the techniques used to perform text mining on collections of textual information [50].

An instance study [51] is presented text that is using to identify groups and styles of related study subjects from three major journals within the administration information methods field. Based on the results for this example, it's proposed that this type of analysis could be valuable for potentially scientists in just about any field [52]. Web mining strategies seek to draw out understanding from Web information. This article [53] provides a breakdown of previous and work that is present the 3 main areas of Web mining study content, structure, and consumption along with promising work with semantic Web mining [54].

Today web is the method this is certainly most readily useful of interaction in modern-day business. Many companies tend to be redefining their business strategies to enhance the continuing business result. Company over net offers the chance to clients and partners where their products or services and business that is specific be located. Nowadays online business breaks the buffer of time and room when compared with the office that is real. Huge businesses across the global globe tend to be recognizing that e-commerce isn't just exchanging over Internet, instead it gets better the efficiency to contend with various other leaders on the market.

For this purpose information mining often called as knowledge finding can be used. Web mining is information mining method this is certainly applied to the WWW. One can find vast degrees of information offered on the Internet.The automated retrieval of data from the internet, commonly known as web scraping, has become a prevalent practice in both industry and academic research endeavors. Numerous tools and technologies have been developed to facilitate this process. However, the ethical and legal considerations surrounding the use of these tools for data collection are often disregarded.

Neglecting these aspects of web scraping can lead to significant ethical dilemmas and legal disputes. This paper conducts a comprehensive review of legal literature as well as ethics and privacy literature to identify overarching areas of concern. It also presents a set of specific questions that researchers and practitioners involved in web scraping should address. By carefully considering these questions and concerns, researchers can proactively mitigate the risk of encountering ethical and legal challenges in their work.Scientific, political, and bureaucratic elites employ epistemic techniques such as "big data analysis" and "web scraping" to construct depictions of the populace and validate policy decisions. I introduce the concept of "demos scraping" to describe these methods of acquiring information about citizens (the "demos") through automated examination of digital traces, repurposed for political objectives.

This article critically examines the discourse advocating demos scraping and offers a conceptual evaluation of its democratic consequences. It scrutinizes the assertions of demos scraping proponents who claim it can narrow the divide between political elites and the public, asserting that it offers a more superior means of gauging the "will of the people" and enhancing democratic legitimacy. Consequently, this prompts a critical exploration of the repercussions of demos scraping on political representation and citizen participation. Currently, demos scraping exhibits technocratic and de-politicizing characteristics, and its operation within the broader political and economic framework makes it improbable to bridge the gap between elites and citizens. Adopting a post-democratic perspective, demos scraping appears as an endeavor within late modern and digitized societies to grapple with the democratic dilemma posed by rising citizen expectations amidst a profound crisis of legitimacy.

Web scraping refers to the practice of extracting valuable text information from web pages. Most of the current research in this area primarily focuses on automating the process of collecting web data. In these studies, the typical approach involves constructing a Document Object Model (DOM) tree and then accessing the required data through this tree structure. The time required for creating this tree can significantly increase depending on the complexity of the DOM structure. Unfortunately, the existing literature on web scraping often overlooks the importance of time efficiency

3 RESEARH METHODOLOGY

3.1 Scrapy Shell

The Scrapy shell is an interactive shell where the scraping code is debugged very quickly, without having to run the spider. It's meant to be used for testing scraper code, but it can be used to test any kind of code as it is also a regular Python shell.

The shell is used for testing XPath or CSS expressions to see how they work and what data is extracted from the website that is needed to be scraped. It allows the scraper to interactively test the expressions of a scraper while writing the spider, without having to run the spider to test every change.

For developing and debugging the spiders the Scrapy shell is an invaluable tool If IPython installed on the operating system, the Scrapyshell will use it (instead of the standard Python console). Smart auto completion and colorized output among other things is by the powerful IPython console.

It is highly recommended to install IPython, especially for UNIX systems (where IPython excels).

3.2 Web Crawling Framework

Scrapy, overall, is a web crawling framework written in Python. One of its main advantages is that it's built on top of Twisted, an asynchronous networking framework, which in other words means that it's: a) really efficient, and b) Scrapy is an asynchronous framework. So, to illustrate is supported under or uses Python 2.7 and Python 3.3. So a scraper can pretty much, good to go. So Python 2.6, important thing to note support was dropped starting at Scrapy 0.20. So scrapers have to bear that in mind, and Python 3 support was added in Scrapy 1.1 with each version of Python the support added in Scrapy.

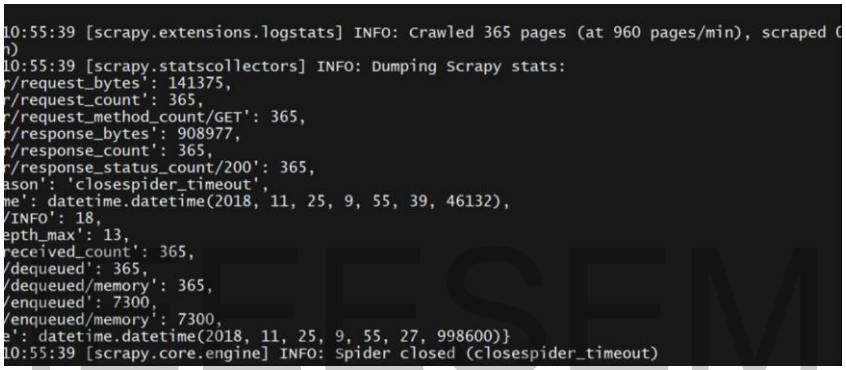


Figure 3.1 Scrapy's Benchmark

The is the benchmark of scrapy which shows that scrapy is able to scrape at 960 pages/min this can be different from a pc to another regarding to one's CPU Performance how much ram does one have and internet speed.

So Scrapy has 5 components and here how it works the engine gets the demands which are initial crawl from the Spider. The Engine schedules the demands in the Scheduler and wants the requests which are next crawl. The Scheduler comes back the requests which can be next the Engine. The demands are sent by the Engine to the Downloader, passing through the Downloader Middleware's (see process request ()). Once the page completes getting the Downloader yields a reply (with that web page) and delivers it to the Engine, moving through the Downloader Middleware's (see process response()). The Engine gets the Response from the Downloader and sends it towards the Spider for processing, passing through the Spider Middleware (see process_spider_input ()). The Spider processes the Response and returns scraped items and needs that are newt follow) to your Engine,

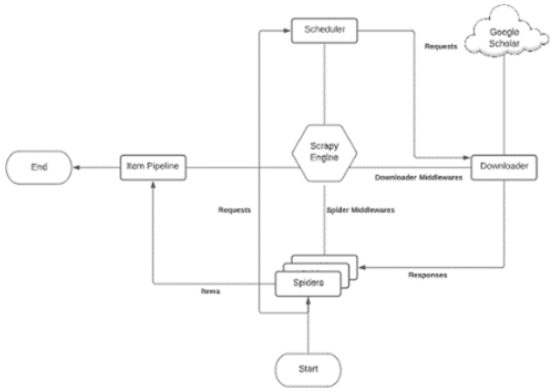


Figure 3.2 Scrapy's Architecture

3.3 JavaScript Rendering Service

Splash is like a lightweight browser. The way we interact with that browser is by writing some code that splash can understand not by using icons like Chrome for example and it's meant to be used with scrapy. JavaScript requires an engine to be executed. So each engine each browser use including splash so Chrome has what we call the V8 engine. Firefox has spider monkey Safari has Apple Web Kit. That's the same engine used by splash and Microsoft Edge has Chakra.

3.4 Extraction of Articles from Google Scholar

This section simplifies the data preparation phase for the automatic release of educational textbook content [14-16]. Earlier articles discussed how easily accessible web information might be recycled by cycling as part of several research procedures at various stages and overcoming obstacles. However, the process of automatically extracting material is broken down into more than three phases (1). Access to the Website (2). Extraction of Hyper Text Markup Language and content and (3). Exhaust structure [17].

High-tech fixes and technologies provide a variety of easy-to-use aids. These fixes are a far cry from the past, when information distribution was fraught with problems, particularly for middle-skilled users. To extract information from selected sources, many of these approaches rely on API and code- based method websites. The integrated technology underlying three internal process processes can explain the complex conventional issues in handling the Application Programming Interface and architectures that are code based.

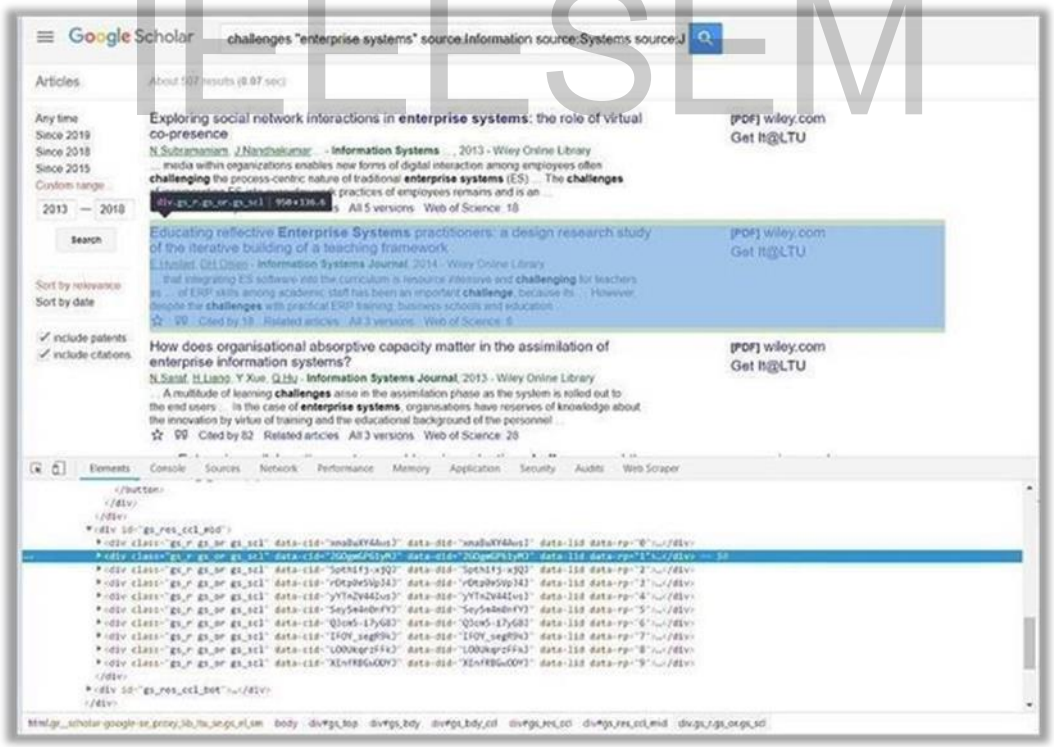
To begin scrapping or crawling the site, a user must frequently enter complicated code in a variety of parameters, beginning with the URL of the page to be crawled (Accessing the site). Given that website URLs generally contain tight coding, the issue here is the lack of robust URL parameters.

Given that website URLs generally contain tight coding, the issue here is the lack of robust URL parameters.

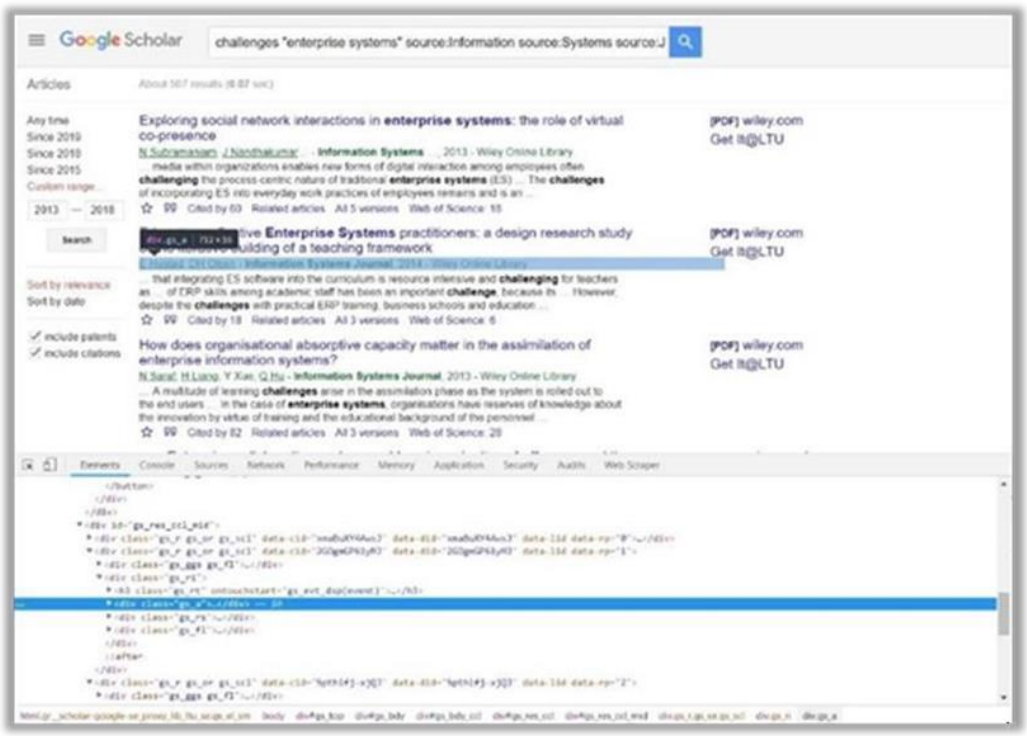
This issue surfaces again and again, particularly in the following stage of parsing of HTML and data extraction. Segmentation in HTML and content extraction often occur by analyzing the structure of websites and content domains. A classic example where a client will acknowledge the structure of an Xpath of a website and its name to add it / extract it to the content [18].

Many browsers now have built-in developer tools that allow web developers to examine the design of a web page.

As a result, web developers have access to a large variety of development tools. The method of obtaining web material from archives, on the other hand, remains difficult expected to path structures. Shown in Figure 3.3a, b.



(a)



(b)

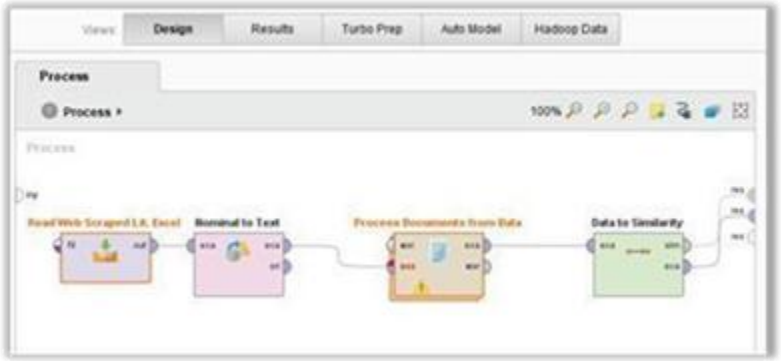
Figure 3.3. The Google Chrome test panel that analyzes the anatomy of a web page: (a)The parent line that highlights specific book searches; (b) highlighting the lower part of the parent's hit

3.5 Mining& Analyzing the Relevance of Literature (Modeling)

Creating a modeling class necessitates a powerful yet simple machine learning platform that allows consumers to successfully manage their goods. Most importantly, it empowers researchers to supplement the book search process, intending to focus on the essential objectives based on the publication criteria and the study mentioned above. The artificial intelligence forum is brimming with tools that cater to various degrees of user developmentand topic knowledge. [19].

To help the capacity to employ various documentation operations throughout excavation procedures and the analysis of appropriate literature searches, a somewhat sophisticated tool will be necessary. Rapid miner has been outgo-to tool because of its efficiency and ease of use.The following is a higher-level view of text mining as a follow-up to the web-based activity.

To help the capacity to employ various documentation operations throughout excavation procedures and the analysis of appropriate literature searches, a somewhat sophisticated tool will be necessary. Rapid miner has been outgo-to tool because of its efficiency and ease of use.The following is a higher-level view of text mining as a follow-up to the web-based activity. The aforementioned can be balanced in order to readily reproduce the accurate search strategy and summary suggested in the literature while also promoting widespread acquiring because of its simplicity. A similar analogy, Figure 3.5a, b shows a high level of textual extraction process designed to identify termsimilarities betweenthe texts used.



(a)

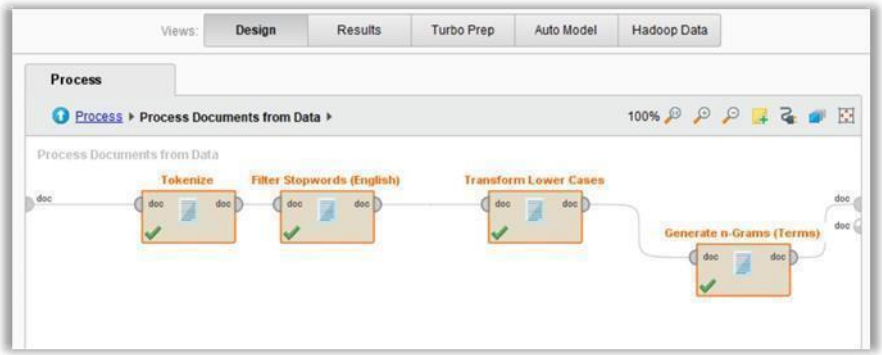


Figure 3.5. (a) Rapid miner (b) Text-mining

It can be seen that all of the items below are GUI-based items that work to increase ease ofuse and time value from the tool.

4 RESULTS AND DISCUSSIONS

4.1 Results and Discussions

So after launching the spider from 4.4 Crawler Avoiding Ban 'scrapy crawl GoogleSchoalar –o data.json' after running the command the results are shown below

```
 'scrapy.extensions.logstats.LogStats',
 'scrapy.extensions.throttle.AutoThrottle']
2018-09-28 01:24:08 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'demo_crawl.middlewares.UserAgentRotatorMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats',
 'scrapy.downloadermiddlewares.httpcache.HttpCacheMiddleware']
2018-09-28 01:24:08 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
```

Fig 4.1 Execution of the spider for 4.4

And after opening data.json in web scraper we will see some of the scraped articles have the user agents '4' and others have user agent '7'

"Keyword": "Article", "user-agent": "agent_4"}, !W-
on-GoogleScholar--conquerentrequests-12", "":
"Article", "user-agent": "agent_4"}, ", "user-
agent": "agent_4"}, agent": "agent_4"},
"GoogleScholarPagination", "user-agent":
"agent_7"}, ', "Keyword": "Article", "user-agent":
"agent_7"}, concurrentrequests--ua-429", "Keyword":
"Article", "user-agent": "agent_7"}, 'Article",
"user-agent": "agent_7"},

Fig 4.2 Data received from the execution of spider

The execution of the command of fig 4.29 in 4.5 ‘Bypass 504 HTTP Error’

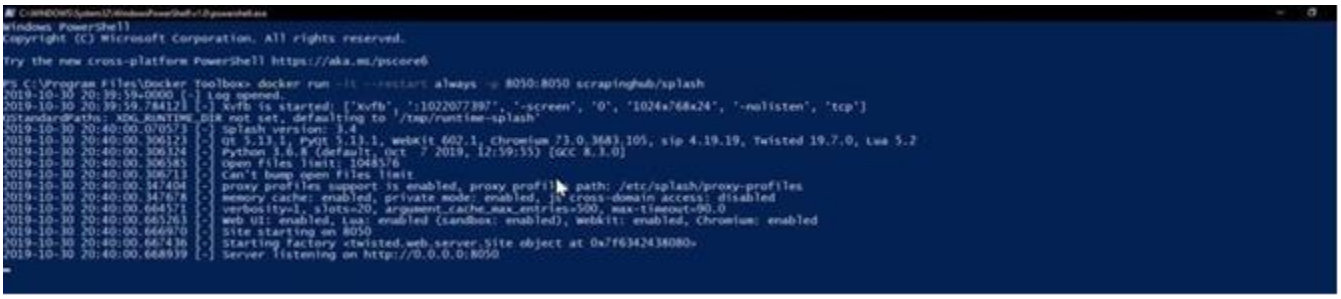


Fig 4.3 Command execution

Now executing the spider scrapy crawl Articles

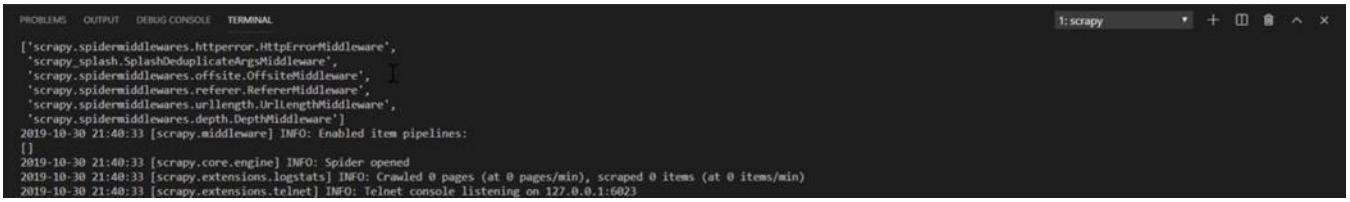


Fig 4.4 Executing the spider to reduce 504 error

Now let’s take a look at scrapy stats if you can see here we have 504 Gateway is equal to 160 this is too much and this also means that amount of memory that I dedicated to the virtual machine wasn't sufficient at all

Fig 4.5 504 Gateway Error

So overall solution to this is that this technique of bypassing 504 error will work again if a real time scraper is made by practicing web scraping techniques by studying the past research articles along with finding out new techniques to bypass 504 error and making an application in which all the operations will be carried out which will run on system that must have higher CPU and higher memory because low CPU and lower memory systems can’t handle the web scraping techniques now so for that a scraper must have a High-end CPU. Now executing the spider scrapy crawl ‘GoogleScholar’ for fig 4.41 in 4.6 to get scraped research articles.

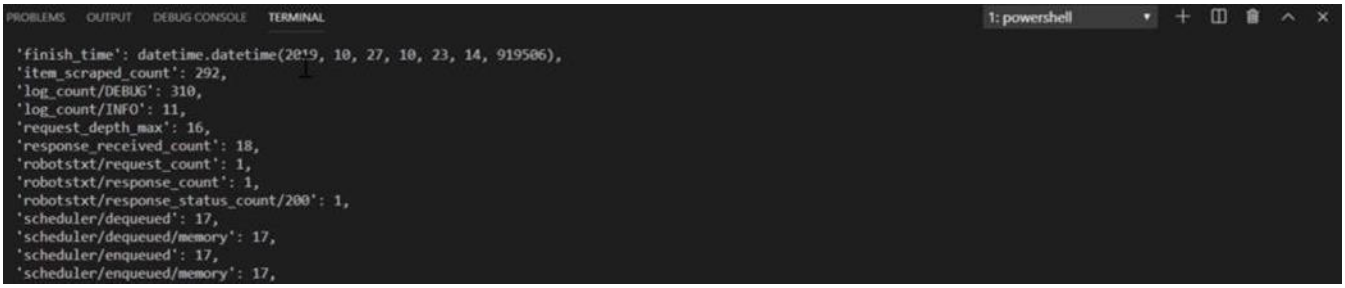


Fig 4.6 Execution of spider for Articles

Following are the research articles scraped through crawlers along with the pdf links

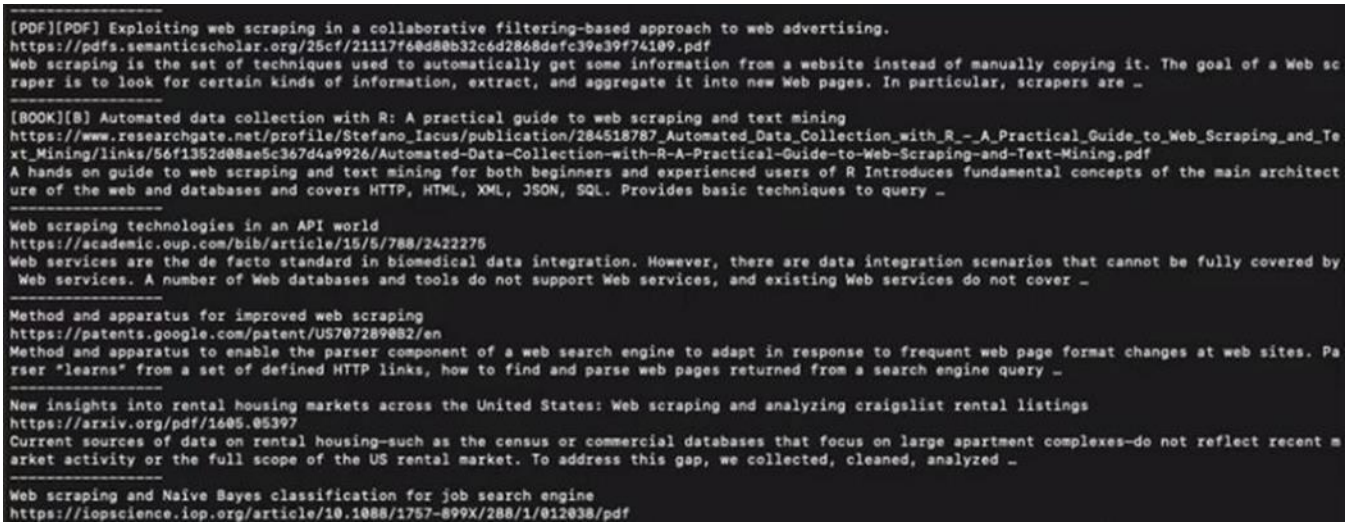


Fig 4.7 Scraped Research Articles

Now here is the data set for all the research articles scraped after the execution of the spider a scraper have to check the data.json file to make sure the crawler bypassed the 504 error and had been successfully executed. In our case the scraper or spider was successfully executed

Fig 4.8 data.json file containing scraped Research Articles

After applying modeling on the articles received from .json It can be seen right away that the top examined phrases are centered around terms like text mining , web crawling, research methodology, python, excel, data analysis, web scraping, and so on. In the case of post-implementation, however, this is not the case. It was noticed from the web scraping, although making a minor proportion of appearances across the literature extracted from journals.

Row No.	word	in documents	total	in class (MIS Quarterly)	in class (Journal of IEEE)
199	text mining	24	24	1	9
200	web crawling	3	3	0	1
201	research methodology	6	6	1	5
202	python	4	4	0	4
203	excel	4	4	0	1
204	data analysis	7	7	2	2
205	web scraping	5	5	2	2
206	Information retrieval	5	5	1	1
207	web data extraction	3	3	1	0
208	practice	7	7	0	3
209	json	7	7	0	2

Table 4.1 with just five instances in total, the phrase “webscraping" is ranked as the 205th most common term in the produced N-Grams.

Original Title 1	Original Title 2	Similarity
Web Scraping And Data Acquisition Using Google Scholar	Web Scraping Scientific Repositories for Augmented Relevant Literature Search Using CRISP-DM	0.501
Web Scraping State-of-the-Art and Areas of Application	Data Analysis by Web Scraping using Python	0.600
Towards data extraction of dynamic content from JavaScript Web applications	Articulating the construction of a web scraper for massive data extraction	0.578
Analysis Of Different Web Data Extraction Techniques	A Review on Web Scrapping and its Applications	0.597
Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation	Tools to Support Systematic Literature Reviews in Software	0.592

Consequently, the findings demonstrated how the suggested technique may be used to supplement the activities of a relevant literature search. As a result, researchers will be able to attain the objectives of breadth and depth of knowledge, rigor and consistency of comprehension and application, as well as clarity and brevity of analysis, synthesis, and assessment more simply and effectively.

5 CONCLUSION AND FUTURE WORK

So when to scrape websites and when to not and Now the very first thing a scraper have to check when scraping a website is terms of services and the robot.txt file now the robot.txt file is used by websites owners they can specify what website pages that can be scraped or not so if a scraper finds the web page that tries to scrape is forbidden in the robot.txt file In this case scraper is not allowed to scrape that website.
The second main thing a scraper have to check is does the website have a public API. If yes a scraper have to check if there are some limitations because some websites limit the amount of request. Next a scraper have to check if the API provides all the data it wants now in case of the API is paid it doesn't provide all the data and it has some limitations In this case the only solution you have is to use web scraping on it.
So a scraper have to keep itself up to date with web scraping technology to avoid any kind of wastage and an if a scraper works further in this a scraper must develop an application in which all these techniques and operations of a crawler, scrapy and splash should be carried in order to scrape websites along with reducing 504 error and additionally if a scraper wants any other technique no related to crawler to work with scrapy and splash like similarity detection technique then a separate portion must be added to that crawler application which will focus on extracting websites using all the techniques and applying similarity detection technique within that application on real time.

REFERENCES

[1] Von Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., Cleven, A. (2009). Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. In Proceedings of the 17th European Conference on Information Systems (ECIS 2009), Verona, Italy, 8–10 June 2009; Volume 9, pp. 2206– 2217.

[2] Webster, J., & Watson, R.T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly, 26, 13–23.

[3] Cooper, H.M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. Knowledge in Society, 1, 104–126.

[4] Nakano, D., & Muniz, J., Jr. (2018). Writing the literature review for empirical papers. Production, 28.

[5] Levy, Y., & Ellis, T.J. (2006). Towards a Framework of Literature Review Process in Support of Information Systems Research. InSITE, 6.

[6] Synnestvedt, M.B., Chen, C., & Holmes, J.H. (2005). CiteSpace II: Visualization and knowledge discovery in bibliographic databases. AMIA Annual Symposium Proceedings, 2005, 724–728.

[7] Hutton, B., Catalá-López, F., & Moher, D. (2016). The PRISMA statement extension for systematic reviews incorporating network meta-analysis: PRISMA-NMA. Medical Clinics, 147, 262–266.

[8] Phongwattana, T., & Chan, J.H. (2018). A Combination of Text Mining Techniques for Relevant Literature Search and Extractive Summarization. In Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval, Bangkok, Thailand, 7–9 September 2018; pp. 7–11.

[9] Chen, C. (2017). CiteSpace: A Practical Guide for Mapping Scientific Literature; Nova Science Publishers, Inc.

[10] Wirth, R., & Hipp, J. (1998). CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the Fourth International Conference on the Practical, New York, NY, USA, 27– 31 August 1998.

[11] Goyal, V.K. (2014). A Comparative Study of Classification Methods in Data Mining using RapidMiner Studio. International Journal of Innovative Research in Science and Engineering, 4, 28–30.

[12] Liao, S., Chu, P., & Hsiao, P. (2012). Data mining techniques and applications—A decade review from 2000 to 2011. Expert Systems with Applications, 39, 11303– 11311.

[13] Eaton, E. (2017). Teaching Integrated AI through Interdisciplinary Project-Driven Courses. AI Magazine, 38, 13.

[14] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’08), Las Vegas, ND, USA, 24– 27 August 2008; ACM, 990–998.

[15] Haddaway, R.N. (2015). The Use of Web-scraping Software in Searching for Grey Literature. Grey Journal, 11, 186–190.

[16] Meschenmoser, P., Meuschke, N., Hotz, M., & Gipp, B. (2016). Scientific Web Repositories: Challenges and Solutions for Automated Content Extraction. D-Lib Magazine, 22, 9–10.

[17] Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. Brief Bioinformatics, 15, 788–797.

[18] Berglund, A., Wg, X.S.L., Boag, S., Wg, X.S.L., Chamberlin, D., Wg, X.M.L.Q., Almaden, I.B.M., Fern, M.F., Wg, X.M.L.Q., Kay, M., et al. (2010). XML Path Language (XPath) 2.0; W3C.

[19] Dwivedi, S. (2016). Comprehensive Study of Data Analytics Tools. In Proceedings of the 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, India, 18–19 March 2016.

[20] Van Der Blonk, H. (2003). Writing case studies in information systems research. Formulation Research Methods in Information Systems, 2, 255–270.

[21] Sykes, T.A., Venkatesh, V., & Johnson, J.L. (2014). Enterprise system implementation and employee job performance: Understanding the role of advice networks. MIS Quarterly, 38, 51–72.

[22] Donovan, R.M. (2001). Successful ERP Implementation the First Time. Midrange ERP. Available online: http://www.midrangeerp.co m. Umble, E.J., & Umble, M.M. (2000). Avoiding ERP implementation failure. Industrial Management, 44, 25–33.

[23] Rasmy, M., Tharwat, A., & Ashraf, S. (2005). Enterprise resource planning (ERP) implementation in the Egyptian organizational context. European Mediterranean Conference on Information Systems (EMCIS 2005), 1–13.

[24] Muscatello, J.R., & Parente, D.H. (2006). Enterprise Resource Planning (ERP): A Postimplementation Cross-Case Analysis. Information Resources Management Journal, 19, 20.

[25] Wang, E.T.G., Chia-Lin Lin, C., Jiang, J.J., & Klein, G. (2007). Improving enterprise resource planning (ERP) fit to organizational process through knowledge transfer. International Journal of Information Management, 27, 200–212.

[26] Osnes, K.B., Olsen, J.R., Vassilakopoulou, P., & Hustad, E. (2018). ERP Systems in Multinational Enterprises: A literature Review of Post-implementation Challenges. Procedia Computer Science, 138, 541–548.

[27] Chatterjee, S. (Year not provided). ERP Failure in Developing Countries: A Case Study in India.

[28] Vijay Gaikwad, S., Chaugule, A., Patil, P., Y Patil, P.D., & Patil, P. (2014). Text Mining Methods and Techniques. International Journal of Computer Applications, 85, 975–8887.

[29] Lin, Y.S., Jiang, J.Y., & Lee, S.J. (2014). A similarity measure for text classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 26, 1575– 1590.

[30] Ishioka, T. (2005). An Expansion of X-Means for Automatically Determining the Optimal Number of Clusters—Progressive Iterations of K-Means and Merging of the Clusters. Proceedings of International Conference on Computational Intelligence, 2, 91–96.

[31] Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102, 16569–16572.

[32] Eggue, L. (2006). Theory and practise of the g-index. Scientometrics, 69, 131–152.

[33] Gu Chengjian, Huang Lucheng. (2008). Web Mining in Technology Management, 2008 International Seminar on Business and Information Management, DOI: 10.1109/ISBIM.2008.127.

[34] Landers, Richard N., Robert C. Brusso, Katelyn J. Cavanaugh, and Andew B. Collmus. (2016). A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research.

[35] Damasevicius, R. (2009). Automatic Generation of Concept Taxonomies from Web Search Data Using Support Vector Machine. In Proceedings of the Fifth International Conference on Web Information Systems and Technologies, Lisbon, Portugal, 23–26 March 2009.

[36] Barcaroli, Giulio, Alessandra Nurra, Marco Scarnò, and Donato Summa. (2014). Use of web scraping and text mining techniques in the Istat survey on "Information and Communication Technology in enterprises". In Proceedings of Quality Conference, pp. 33-38.

[37] Santur, Y., Mustafa, U. L. A. Ş., & Karabatak, M. (2022). A web scraping-based approach for fundamental analysis platform in financial assets. Journal of New Results in Science, 11(3), 222-232.

[38] Bache, S. M., and Wickham, H. (2014), "magrittr: A Forward-Pipe Operator for R," R Package Version 1.5, available at https://CRAN.R-project.org/ package=magrittr.

[39] Brantley, J. (2016), "ZillowR: R Interface to Zillow Real Estate and Mortgage Data API," R Package Version 0.1.0, available at https://CRAN.R-project. org/package=ZillowR.

[40] Dumbacher, B., and Capps, C. (2016), "Big Data Methods for Scraping Government Tax Revenue From the Web," in Proceedings of the Joint Statistical Meetings, Section on Statistical Learning and Data Science, pp. 2940–2954.

[41] Meissner, P., and Run, K. (2018), "robotstxt: A robots.txt' Parser and 'Webbot'/'Spider'/'Crawler' Permissions Checker," R Package Version 0.6.2, available at https://CRAN.R-project.org/package=robotstxt.

[42] Polidoro, F., Giannini, R., Conte, R. L., Mosca, S., and Rossetti, F. (2015), "Web Scraping Techniques to Collect Data on Consumer Electronics and Airfares for Italian HICP Compilation," Statistical Journal of the IAOS, 31, 165–176.

[43] Robertson, A. (2019), "Scraping Public Data From a Website Probably Isn't Hacking, Says Court," available at https://www.theverge.com/2019/ 9/10/20859399/linkedin-hiq-data-scraping-cfaa-lawsuit-ninth-circuitruling.

[44] Statistics Canada (2019), "Web Scraping," available at https://www.statcan. gc.ca/eng/our-data/where/web-scraping.

[45] Ten Bosch, O., Windmeijer, D., van Delden, A., and van den Heuvel, G. (2018), "Web Scraping Meets Survey Design: Combining Forces," in Big Data Meets Survey Science Conference, Barcelona, Spain.

[46] Wickham, H. (2014), "Tidy Data," Journal of Statistical Software, 59, 1–23.

[47] Wickham, H. (2019a), "rvest: Easily Harvest (Scrape) Web Pages," R Package Version 0.3.5, available at https://CRAN.R-project.org/ package=rvest.

[48] Zamora, A. (2019), "Making Room for Big Data: Web Scraping and an Affirmative Right to Access Publicly Available Information Online," Journal of Business, Entrepreneurship and the Law, 12, 203–228.

[49] Destatis (2018), "Methods—Approaches—Developments."

[50] Google (2019), "Googlebot," available at https://support.google.com/ webmasters/answer/182072?hl=en.

[51] IMDB (2019), "Feature Film, Released Between 2018-01-01 and 2018-12- 31 (Sorted by Number of Votes Descending)," available at https://www. imdb.com/search/title/?title_type=feature&year=2018-01-01,2018-12- 31&sort=num_votes,desc.

[52] Introduction to robots.txt (2019), https://support.google.com/webmasters/ answer/6062608?hl=en.

[53] Kearney, M. W. (2019), "rtweet: Collecting and Analyzing Twitter Data," Journal of Open Source Software, 4, 1829. R Package Version 0.7.0, available at https://joss.theoj.org/papers/10.21105/joss.01829.

[54] Open Secrets—Foreign Connected PACs (2019), https://www.opensecrets. org/political-action-committees-pacs/foreign-connected-pacs. OpenSecrets.org (2019), https://www.opensecrets.org. Woollacott, E. (2016), "70,000 OkCupid Profiles Leaked, Intimate Details and All," available at https://www.forbes.com/sites/emmawoollacott/