ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Knowledge Acquisition Bot – KAB

Wasique Ali Ansari[1], Paritosh Diya[2], Sahishnu Patil[3], Dr Sunita R Patil[4].

U.G. Student, Department of Computer Engineering, KJ Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India[1]

U.G. Student, Department of Computer Engineering, KJ Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India[2]

U.G. Student, Department of Computer Engineering, KJ Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India[3]

Vice Principal, Professor, Department of Computer Engineering, KJ Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India[4]

ABSTRACT: The expansion of the internet since its inception is incomparable to any other data and knowledge collection methodology used to date. This expansion is noteworthy because of the quantity of data that is collected in that time period. The internet is filled with information and is getting more and more data every day. This huge collection of data requires efficient data processing techniques, which will help in refining the data and removing the unwanted and redundant information, which might be possible. This paper demonstrates a Knowledge Acquisition Bot (KAB) as a robotic process automation (RPA) product that uses two promising techniques, Natural Language Processing (NLP) and Web technologies which, when combined, enables the enterprise to obtain the unstructured and structured data in ways that were not handled efficiently by traditional tools. For a better understanding of web information, this RPA system uses a combination of NLP processing and Web portal. This paper also explores NLP APIs which can be used for the future of RPA.

KEYWORDS: RPA, process, robot, automation, machines, knowledge, NLP, tasks.

I. INTRODUCTION

According to the Google dictionary, the term “bot” is defined as “an autonomous program on a network (especially the Internet) which can interact with systems or users in order to carry out some monotonous task”. In accordance with automation, a “bot” is best described as a software or program which is instructed to carry out tasks which are repetitive, and are conducted by humans. The ideology for automation is to create bots which will handle these repetitive tasks, such as answers to a particular type of email query, and will give the output with same effectiveness and efficiency on each iteration.

The use of a bot is the base on which the technology of Robotic Process Automation (RPA) originated. The process of automating business operations with the help of robots to reduce human intervention is said to be Robotic Process Automation (RPA).If we deeply try and understand the term RPA, it can be broken into 3 main terms:

Robots: Objects or entities that perform the set of operations that are otherwise performed by humans.

Process: The set of sequence or order of instructions that are laid down which are followed by the robots to achieve some specific goal.

Automation: Any process that is done by the machines without human interference in the working.

In this paper, we have shown how the use of RPA technology and bots has helped us achieve a goal that is required: Knowledge Acquisition. The term, “knowledge acquisition” is basically gathering of information, or data, from informative resources, such as books, internet, etc. The main aim that was to be accomplished is to generate an output

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

summary on any particular topic, by scanning the data present on the internet, and filtering only the relevant and necessary data.

The problem that is tackled using the Knowledge Acquisition Bot (KAB) is related to data extraction and summarization. We have an abundance of data present online, and even if the unwanted data is removed, the vast size and amount of data present after filtering is exceptional. If this problem is faced by a human, then the time, energy and resources consumed for collecting data relevant to a specific topic seem monotonous and mundane. This routine can be seen as a repetitive pattern, which could be automated using Robotic Process Automation. This conceptualisation of automating the processes is the base on which Knowledge Acquisition Bot (KAB) product has been developed.

The purpose of the development of KAB product is to increase the efficiency of the task of data collection and summarization, which is deficient with the use of traditional tools currently. Along with data extraction, the data summarization technique that is used in KAB ensures that the output is refined and a proper summary of only the relevant data as an output. For the wide future of KAB, API's based on NLP can be used.

Along with the use of API's for improvisation of KAB, the implementation of a vernacular module, which will further improvise the scope and usability of this project. The inclusion of video and pictorial representation can also be seen as an addition in this project.

The paper covers total VII sections. Section II of the paper shows the Literature Surveyed, which would help us in the project. Section III describes the System Design that has been implemented in our robot. Section IV in the paper shows the actual implementation of our project, whereas Section V presents the Results of the implemented bot. Section VI of the paper describes the Future Scope of improvement that can be made in the project, and Section VII presents the Conclusion.

II. LITERATURE REVIEW

1. Authors, Rayner, Gan, Chin and Patricia present a robust framework for Web Information Extraction. They give information for retrieval for both generic and specific search. The framework will look for the data in a specific URL if provided by user else they go to search engine and get all the relevant URLs for the data. This URL will be stored in a database for referencing and future searches and update periodically. The URLs are individually crawled for the data and store in a text and content is check and filter to remove redundancy
2. In this paper, Piotr Andruszkiewicz and Henryk Rybinski categorise the data into two types, the process starts by gathering structured and reliable data from the digital libraries. To address the issue of not always up to date data in digital libraries. the process also extracts all the URL file to gather them all the first page contain the keyword and then the URLs are filtered by the SVM classifier to identify potentially useful pages. then this up to date data is also extracted from the web.to enrich the data sources knowledge they use machine learning algorithm of crowdsourcing to get correct information
3. Authors, Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni, show an Open Information extraction from the web as an unsupervised paradigm. This process uses relation specific extraction in favour of a single process extraction by discovering the relation of interest in the cost of corpus analysis with naming every relation. It also shows TEXTRUNNER as a fully implemented system to extract a massive amount of from a data repository in a corpus. The textrunner will match the recall value for the already ready information thus managing its high precision.
4. Santiago Aguirre and Alejandro Rodriguez compare the efficiency of Robotic Process Automation (RPA) on the front office and back office jobs. Front office jobs such as selling, support, requirements handling whereas back-office jobs include finance and human resource. It was concluded that Robotic Process Automation (RPA) is best suitable for high volume standardized tasks that are specifically rule-driven and there is no need for intellective judgment.
5. To gather data from structured sources such as data libraries, conferences, publications are structured but incomprehensive and many times not updated. The authors have suggested gathering data from the internet. It is unstructured but up to date.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

6. The paper by Galina V. Rybina, Ivan D. Danyakin focuses on the use of Robotic Process Automation (RPA) on creating Intelligent Systems. This paper focuses on data extracted from real-time system considering time as an object for data extraction
7. Tomek Strzalkowski and Barbara Vauthey use various natural language processing components that assist the system in database processing e.g indexing, restriction, word clustering, which is demonstrated in this paper; and translate the user query into an effective one. The database is first parsed with the syntactical parser for getting a certain type of words and then they are clusters between smaller sub-phrases and word occurring in them. Then the relation is used to expand the query with the new terms after removing the highly ambiguous word from the sub-phrases.
8. This book by Un Yong Nahm and Raymond J Mooney focuses on text mining and Information Extraction Methodologies. It uses pattern matching for the extraction of useful information from unstructured text. They have used DiscoTEX (Discovery from Text EXtraction) framework for extracting information from unstructured data.

III. SYSTEM DESIGN

As said above, acquiring information about any related topic or a domain require repetitive task. In the traditional system, if a user wants to get any information about a domain needs to search the internet from the topic and select the website which contains the information and then navigate to the individual website and grab the information, even after grabbing all the information, this information is reviewed and then formatted accordingly. This formatted information serves either as a summary or as a keynote to the user about the topic.
Using this bot, the repetitive task of searching keyword and getting all the website and extracting information from each website will be automated in an efficient way. The Knowledge Acquisition Bot (KAB) system design consists of various modules namely UiPath Bot, Nodejs Front-end, Output files and images.

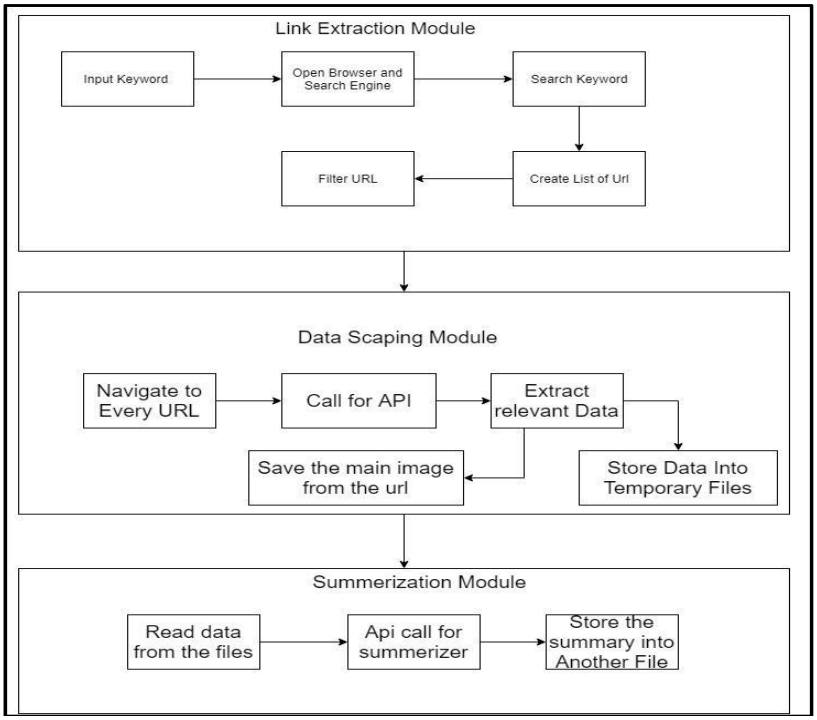

Figure 1. The KAB System Design

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

IV. IMPLEMENTATION

In order to preserve the security and optimal output of the project. The bot must be fully automated with a report log in every step and should efficiently scrape the article on the website. The KAB project is designed in UiPath Studio, which is an upcoming and leading software for creating software bot with Nodejs for a patron system. For sub-security, the UI is designed with Nodejs with the express framework, thus, providing better routing control and scraping of the data for the bot. Since the Capableness of the bot is highly dependent on accumulating data from the website, hence this task was achieved by using API Aylien, which uses NLP processing for extracting appropriate information from the website.

This bot only requirement is the keywords to be searched for and generated an output summary of the keywords. The bot consists of three main modules called the Link Extraction module, Data Extraction module and Summarization module.

Figure 2. Shows the complete flowchart sequence of the KAB project developed in UiPath Studio

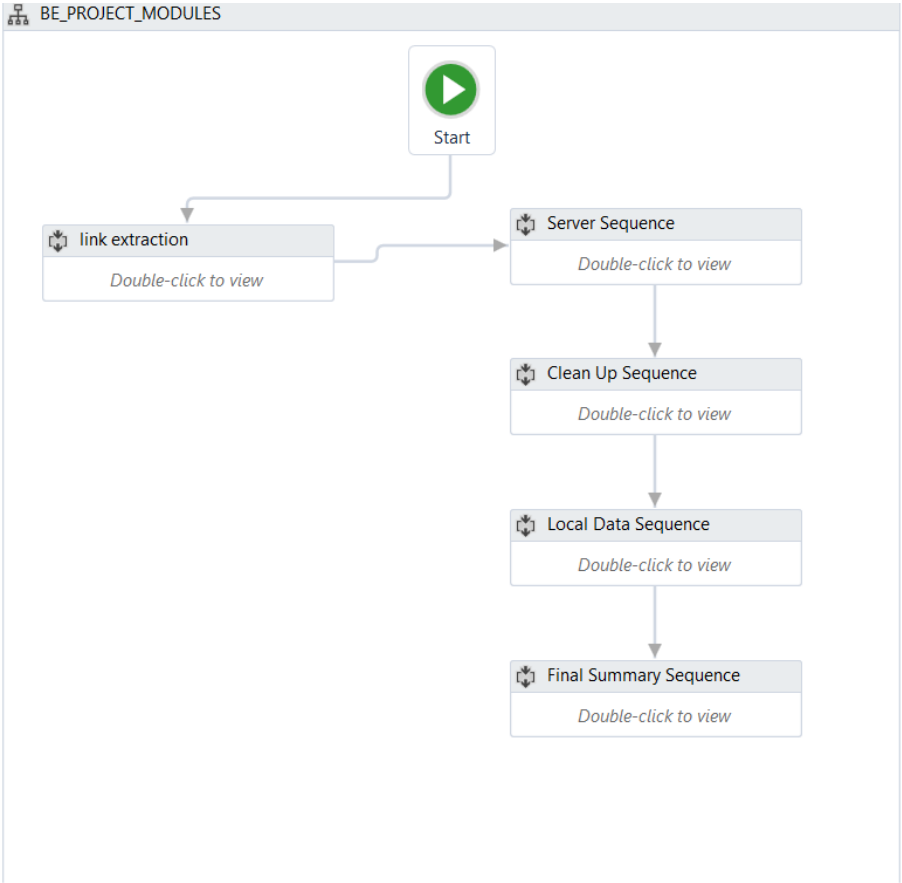

Figure 2: Modules of KAB

Figure 3 shows the link extraction module. The Link extraction module takes input from the user and further parses it to the search engine in the browser to get the list of top up to date results. After the list of all the URLs is obtained, the unwanted URLs are filtered out. Now the module consists of only genuinely required URLs to the web pages.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

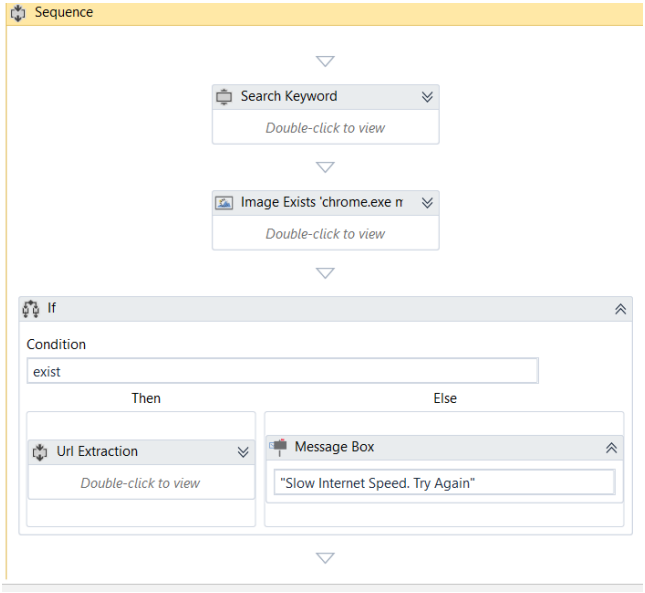

Figure 3: Link Extraction Module Sequence

The next module, Data Extraction module navigates to every final list of obtained URLs with the help of used Application Programming Interface (API) and extracts data from the web pages as shown in figure 4. It stores scraped data and images into separate files. Text data is further parsed to summarization module.

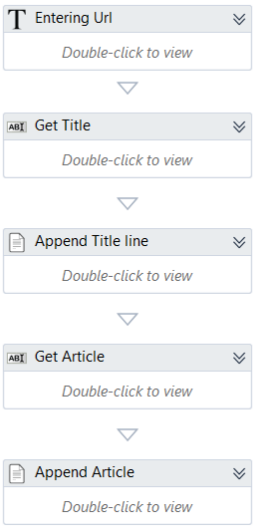

Figure 4 Data Extraction Module

The last module of the system design is the summarization module. This module processes the collected data. The textual data stored in a temporary file is read and parsed to the API for summarization purpose. After all the data is summarized the final summarized information is stored in another file.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

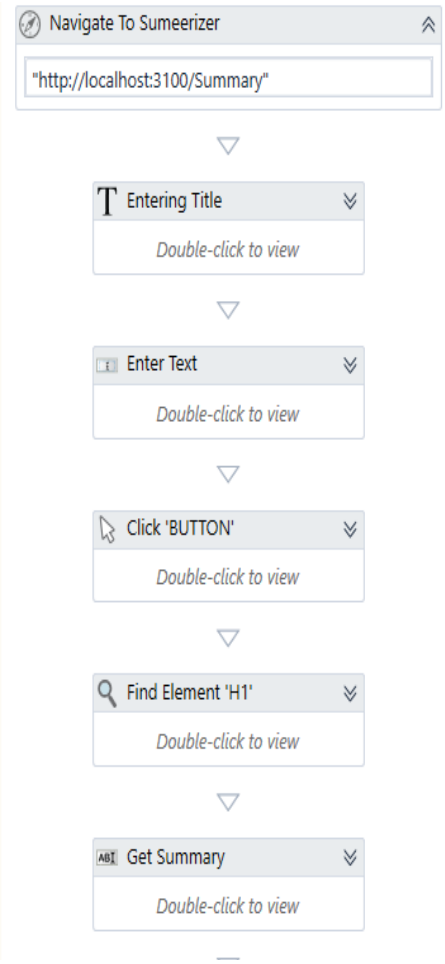

Figure 5. Summarization Module

V. RESULT

The result shows that the user expected articles (knowledge) is extracted from the given URL in terms of text and image files from all the links which are available on the internet related to the search keyword. Figure 7 shows the results acquire from KAB. The whole 'KAB' is fully automated, the system provides report log for the bot for debugging purpose. Thus, the 'KAB' will create two text file and one folder as a result.

One of the text files contains all the information including the Title of the article scrape from the web pages without any processing. This data file is to be called as RawData, another text file will contain the summary of the said "RawData" encompassing all the information of the topic using NLP processing. Along with these two files, a folder containing the main theme image from the website will be acquired after searching the keywords in the search engine. Figure 6 shows the outcome as an Article Extraction from the given URL.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

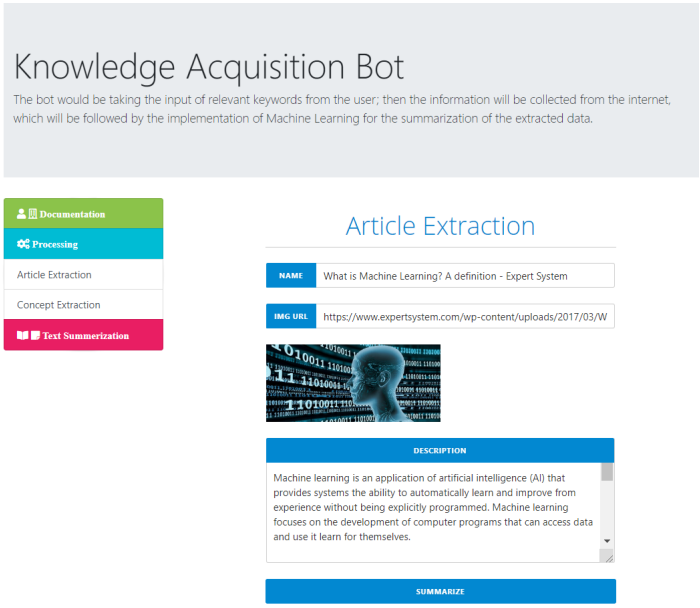

Figure 6: The result of article Extraction.

Figure 7 shows the final summary generated by the Nodejs front-end by NLP summariser using an API call which will be copy by the bot and save it to the document. This document will be the result which will be provided to the end user.

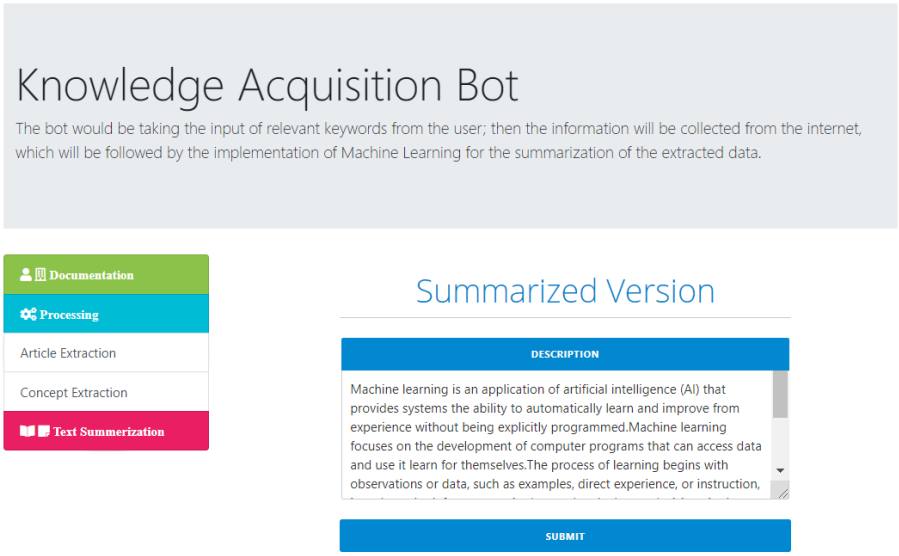

Figure 7: The front-end for the Summeriser.

VII. CONCLUSION

The presence of such an abundance of data places the need for an efficient way to sort out this data and present only the required information. Using the Automated Knowledge Acquisition System as a bot ('KAB'), the data is refined and

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

only the expected data is given as an output. The efficiency of this product is much better than the present tools used, and so it will prove to be an effective and efficient tool in this domain.
The use of API's also provides further scope for improvements, which would furthermore improve the working and functioning of the product. Hence due to the use of robotic process automation, and good API's, this system as a bot can be further expanded according to other use cases and applications.

REFERENCES

[1] Alfred, Rayner, et al. "A Robust Framework for Web Information Extraction and Retrieval." International Journal of Machine Learning and Computing, vol. 4, no. 2, Apr. 2014, pp. 146–50. Crossref, doi:10.7763/IJMLC. 2014.V4.403.
[2] Andruszkiewicz, Piotr, and Henryk Rybinski. "Data Acquisition and Information Extraction for Scientific Knowledge Base Building." 2018 IEEE 12th International Conference on Semantic Computing (ICSC), IEEE, 2018, pp. 256–59. Crossref, doi:10.1109/ICSC.2018.00045.
[3] Banko, Michele & Cafarella, Michael & Soderland, Stephen & Broadhead, Matthew & Etzioni, Oren. (2007). Open Information Extraction from the Web. IJCAI International Joint Conference on Artificial Intelligence. pp 2670-2676.
[4] Aguirre, Santiago, and Alejandro Rodriguez. "Automation of a Business Process Using Robotic Process Automation (RPA): A Case Study." Applied Computer Sciences in Engineering, edited by Juan Carlos Figueroa-García et al., vol. 742, Springer International Publishing, 2017, pp. 65–71. Crossref, doi:10.1007/978-3-319-66963-2_7.
[5] Andruszkiewicz, Piotr, and Henryk Rybinski. "Data Acquisition and Information Extraction for Scientific Knowledge Base Building." 2018 IEEE 12th International Conference on Semantic Computing (ICSC), IEEE, 2018, pp. 256–59. Crossref, doi:10.1109/ICSC.2018.00045.
[6] Rybina, Galina V., and Ivan D. Danyakin. "Combined Method of Automated Temporal Information Acquisition for Development of Knowledge Bases of Intelligent Systems." 2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA), IEEE, 2017, pp. 123–27. Crossref, doi:10.1109/ICKEA.2017.8169914.
[7] Strzalkowski, Tomek, and Barbara Vauthey. "Information Retrieval Using Robust Natural Language Processing." Proceedings of the 30th Annual Meeting on Association for Computational Linguistics -, Association for Computational Linguistics, 1992, pp. 104–11. Crossref, doi:10.3115/981967.981981.
[8] Yong Nahm, Un & J. Mooney, Raymond. (2002). Text Mining with Information Extraction.
[9] https://irpaai.com/what-is-robotic-process-automation/
[10] https://blog.aimultiple.com/rpa/