COMPILER
Vol. 10, No. 2, November 2021, pp. 71-78
ISSN: 2252-3839 (Print)-2549 2403 (On Line)
Accredited Fourth Grade by Kemenristekdikti, Decree No:200/M/KPT/2020
DOI: 10.28989/compiler.v10i2.1064

Analysis of web scraping techniques to get keywords suggestion and allintitle automatically from google search engines

Aris Wahyu Murdiyanto[1,*], Adri Priadana[2]
[1]Department of Information System, Universitas Jenderal Achmad Yani Yogyakarta, Indonesia
[2]Department of Informatics, Universitas Jenderal Achmad Yani Yogyakarta, Indonesia

Article Info

Article history:

Received August 14, 2021
Accepted September 9, 2021
Published November 30, 2021

Kata Kunci:

Automatic keyword suggestion
Automatic allintitle
Keyword research
Google search engine
Web scraping

ABSTRAK

Getting keywords suggestions and allintitle from Google search engine will not be effective, efficient, and economical If we do manually for relatively extensive keyword research. It will take a long time to decide whether a keyword is needed to be optimized or not. Based on these problems, this study was aimed to analyze the implementation of the web scraping technique to get relevant keyword suggestions from the Google search engine and the number of "allintitle" that are owned automatically. The data used as an experiment in this test consists of ten keywords and each keyword would generate a maximum of ten keywords suggestion. Therefore, from ten keywords, it will produce at most 100 keywords suggestions and the number of allintitles. Based on the evaluation result, we got an accuracy of 100%. It indicated that the technique could be applied to get keywords suggestions and allintitle from Google search engines with outstanding accuracy values.

Corresponding Author:

Aris Wahyu Murdiyanto,
Department of Information System,
Universitas Jenderal Achmad Yani Yogyakarta,
Jl. Siliwangi, Ringroad Barat, Banyuraden, Gamping, Sleman, Yogyakarta 55293, Indonesia.
Email: *ariswahyumurdiyanto@gmail.com

1. INTRODUCTION

The development of technology-based businesses has grown rapidly in the last five years. It is indicated by the growth of information seekers through search engines by 4 billion every month with a spread of 62% accessing online shops and 34% accessing online businesses. Based on a survey conducted by the Indonesian Internet Network Providers Association (APJII) in 2019, 93% of online business visitor traffic came through search engines, where 62% came from the first page of Google search. The data shows that there are millions of business opportunities that can be pursued to be potential customers. One way to optimize online business opportunities is by implementing Search Engine Optimization (SEO) techniques on the website so that it can appear in the top position of search results on a search engine based on a particular targeted keyword [1].

Skilled SEO can have a significant impact on the business, including increasing the SEO technique in search engines and potentially getting more traffic [2] from potential customers who are already interested in the business through the Google search engine [3]. Keywords play the most important role because they will determine the direction of the content theme to be developed. Therefore, it is necessary to select keywords that are relevant to the content to be marketed to potential customers through organic search engine searches [4]. Keywords are important words or phrases that are used by a user to find everything that they want to know in search engines, and then read the content on web pages based on search results.

Keyword research is one of the most important activities in SEO [5]. One of the techniques in doing keyword research is to find out how many articles titles on a website indexed by the Google search engine contain a particular keyword or so-called "allintitle" [6]. This technique can be done by typing the word "allintitle: name of article title" in the Google search field. Then you will find the number of titles published to be used as the basis for the number of competitors with the same title for content to be published based on the targeted keywords. Moreover, search engines are also able to provide keywords suggestion that

we have entered into the search field, approximately 10 relevant sub-search suggestions [7], and this can be an insight or input for keyword researchers to get the best keywords are that many are looking for relatively few competitors and trigger potential customers. Keyword search using this technique will not be effective, efficient, and economical if it is done manually for relatively large amounts of keyword search because it will take a long time to decide whether a keyword is needed to be optimized or not.

Based on these problems, this study was aimed to analyze the implementation of web scraping technique to get relevant keyword suggestions from Google search engine along with the number of "allintitle" that are owned automatically. Suggested keywords and the number of "allintitle" are data that will appear when searching through the Google search engine website pages. To get keyword suggestion data and "allintitle" automatically, this study implemented the web scraping method to extract data from the Google search engine website. Web scraping method is one of them which is implemented using the Python programming language.

The implementation of the web scraping technique has been carried out by several previous researchers. Akrianto, et al., in 2019 [8], applied the web scraping method to extract data from the Instagram website page. Priadana and Murdiyanto in 2020 [9], applied the web scraping method to extract data from the Instagram website page. from both studies, the web scraping method was proven to be used to extract data from the Instagram website page. The web scraping technique has also been applied to extract data from a web page in some previous researches. Priyanto and Ma'arif in 2018 [10], applied the web scraping technique to acquire many web pages that provide information about hydroponics. Yani, et al., in 2019 [12], applied the web scraping technique to collect information from many marketplace web pages. Tandra, et al., in 2020 [11], applied the web scraping technique to collect food promotion information from many web pages.

The implementation of data extraction techniques from search engine web pages by applying the web scraping method has also been carried out by several previous researchers. Pranav and Chauhan in 2015 [13], applied the web scraping method to extract data from search engine website pages. Upadhyay, et al., In 2017 [14], applied the web scraping method to extract data massively from search engine website pages. Slamet, et al., In 2018 [15], applied the web scraping method to extract job data from search engine website pages. Arsyad, et al., In 2019 [16], applied the web scraping method to extract data on goods published by several E-Marketplaces from search engine website pages. It proves that the web scraping method can be used to extract data from search engine web pages.

Based on the literature studies and discussion of previous studies, it could be concluded that it has not been any research that applied web scraping techniques to get keyword suggestions as well as "allintitle" automatically to do keyword research. This research provided novelty, namely in terms of techniques and object, the implementation of web scraping technique to get keyword suggestions as well as "allintitle" from Google search engine result.

2. RESEARCH METHOD

The web scraping method is a method used to extract data from a web page [17]. In this study, the extraction process was carried out to extract search result data on the Google search engine. This process aims to extract the keyword suggestion data, as well as "allintitle". The stages of the extraction process using the web scraping method in this study are as follows [18]:

a. In the analysis phase, the HTML and JSON structure of the Google search engine website were studied. This process aimed to determine the data structure and elements that would be extracted from search results on the Google search engine based on certain keywords. The "allintitle" location of the Google search engine website is shown in Figure 1 and the keyword suggestion of the the Google search engine website is shown in Figure 2.
b. Creating a crawl engine for parsing HTML and XML documents used a library called Beautiful Soup, which was a library that was available in the Python programming language.
c. The data extraction process on the Google search engine website using the web scraping method was carried out by sending a request to the Google search engine website page address then extracting data both in the HTML tag and data in the form of JSON (JavaScript Object Notation) which contained keyword suggestion as well as the "allintitle". The data extraction process from Google search engine with web scraping techniques is shown in Figure 3.

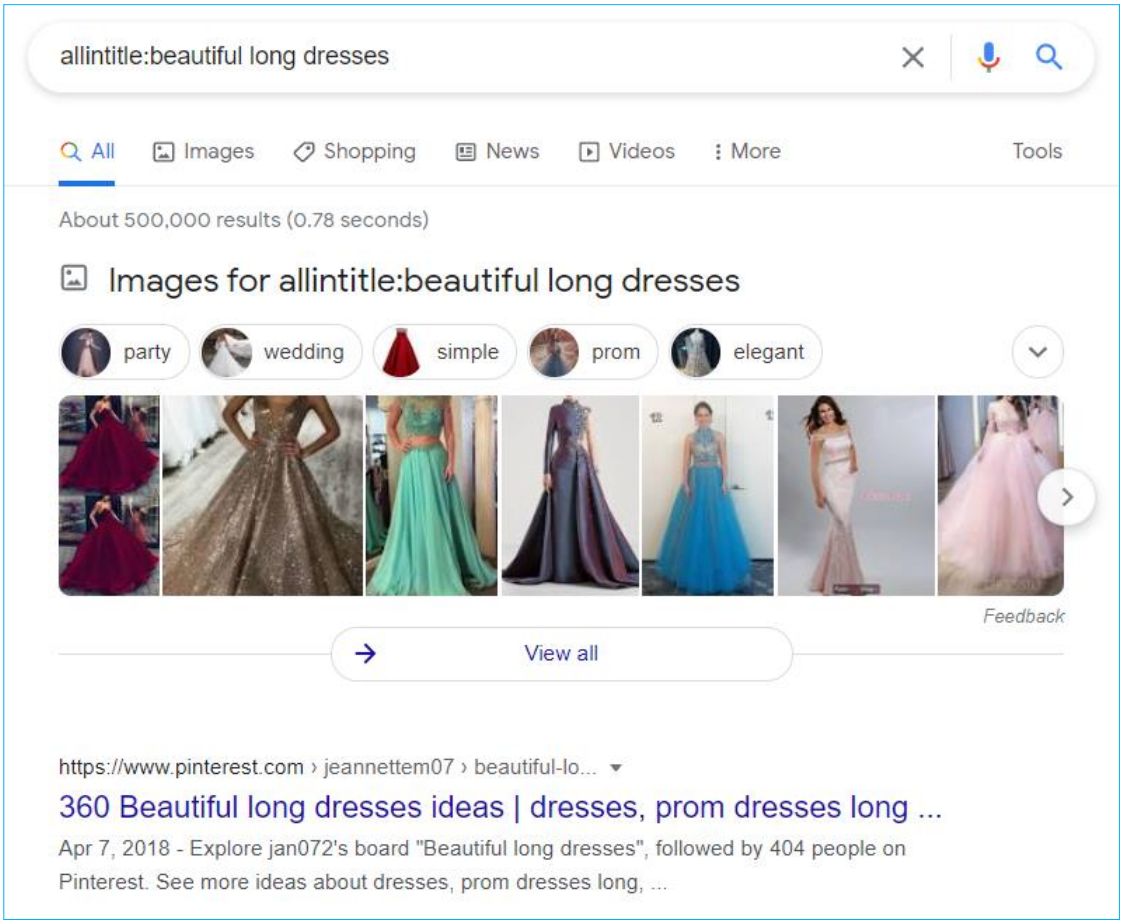

Figure 1 The allintitle's location of the Google search engine website

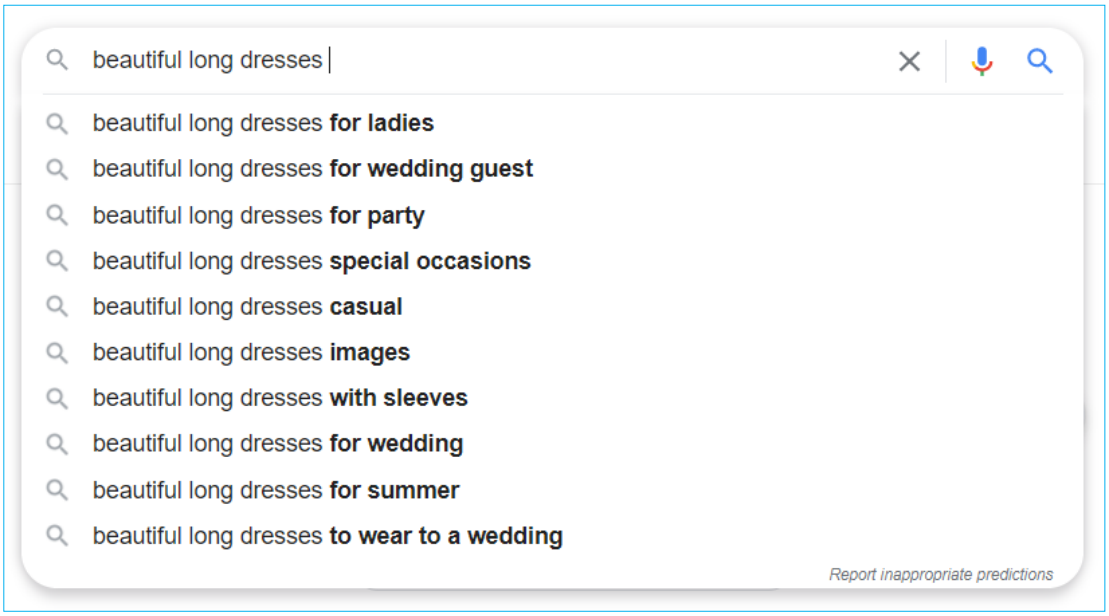

Figure 2. The keyword suggestion's location of the Google search engine website

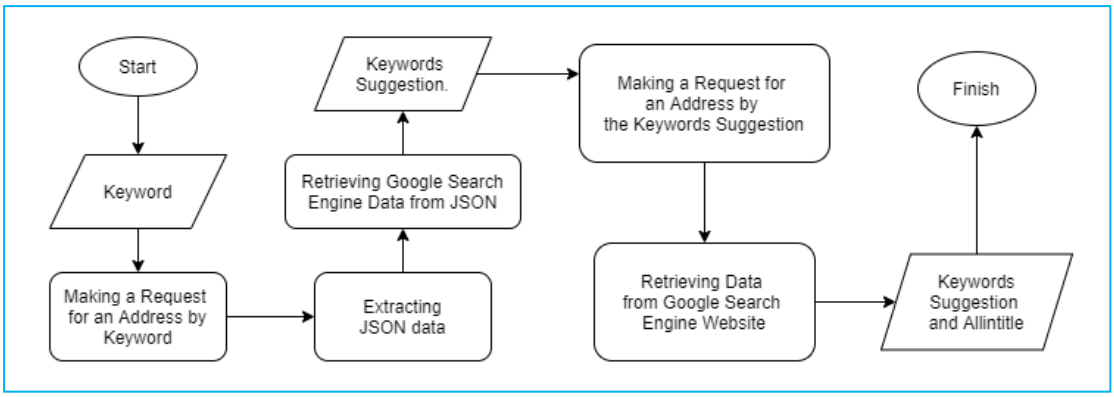

Figure 3. The data extraction process from Google search engine

2.2. Evaluation the result

In this study, the evaluation was carried out by measuring the extent to which the web scraping method was applied to extract the keyword suggestion data and the "allintitle" following the actual. Measurement of the evaluation of the results was done by calculating The Percentage Correct Classification (PCC) of the system or can be called accuracy. To measure the accuracy of the results was carried out using Equation 1.

$$accuracy = \frac{the\ appropriate\ amount\ of\ extraction.}{the\ number\ of\ the\ all\ data}\ x\ 100\% \quad (1)$$

3. RESULT AND DISCUSSION

This chapter explains the results and discussion related to the results of implementing the web scraping method and analyzing the result comparing with the actual result. The first step in implementing the web scraping technique was data structure analysis in the form of HTML and JSON which contained keyword suggestions and "allintitle" data which was done by utilizing the Inspect feature available on the Google Chrome browser. Through this feature, we could see the HTML and JSON structure of a web page. The HTML structure of a website page could be seen through the Element menu, while the data structure stored in a JSON can be seen through the Network menu.

Based on the analysis results, the keyword suggestions data from search results on the Google search engine was in the form of JSON. The JSON data displayed in an Online JSON Viewer tool is shown in Figure 4. On the other side, the "allintitle" data from search results on the Google search engine was in an HTML tag, namely in the div tag with *id*, namely "result-stats" which is shown in Figure 5.

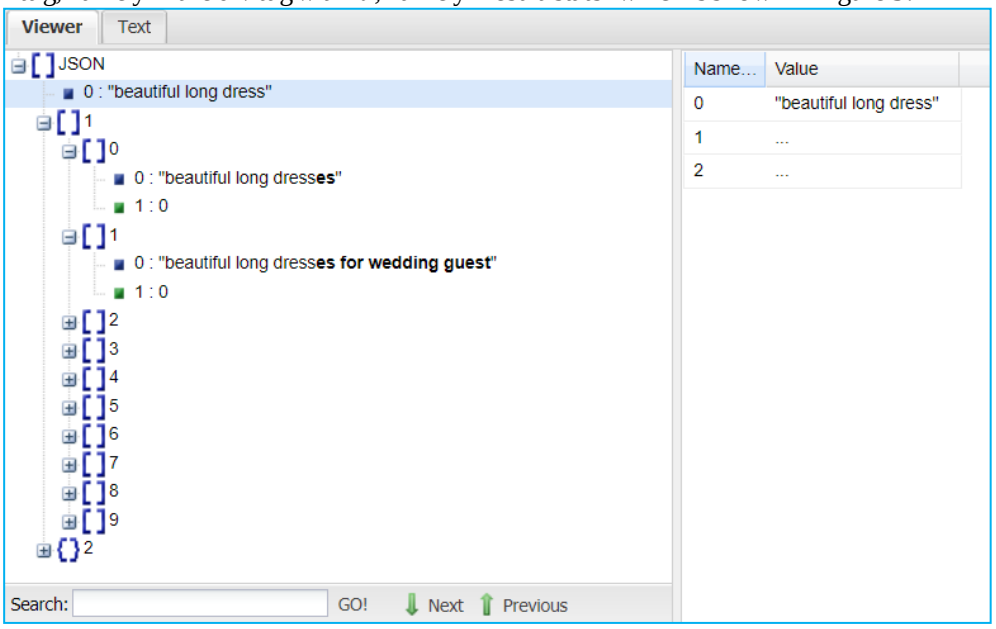

Figure 4. JSON data displayed of keyword suggestions

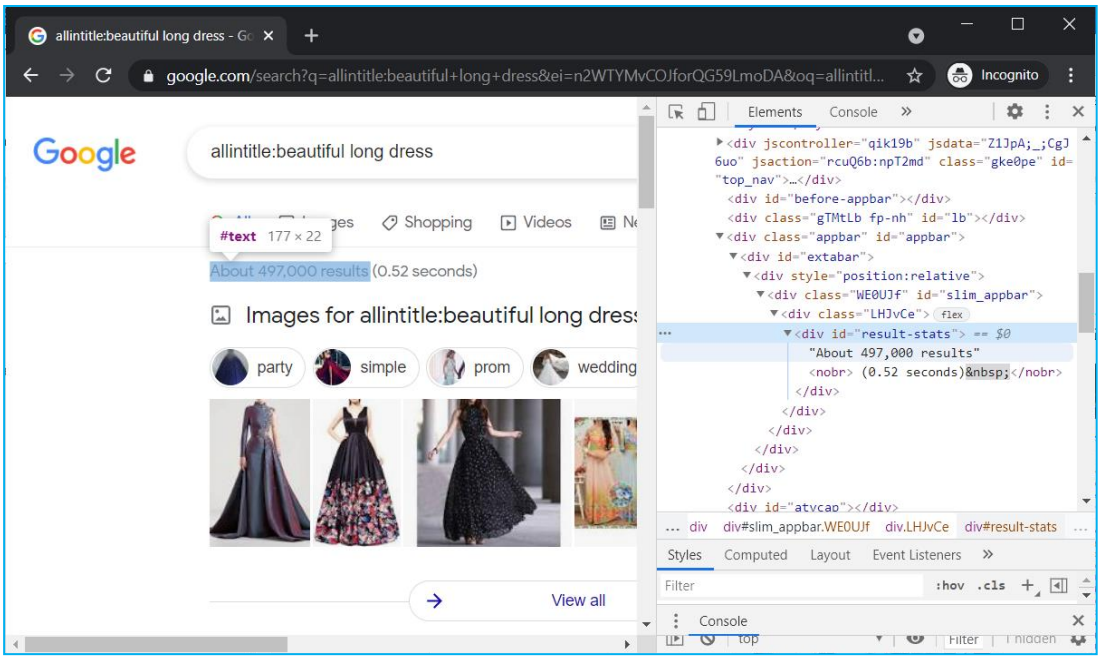

Figure 5. HTML data displayed of "allintitle"

The next step was creating a crawl engine using a library called Beautiful Soup, which is a library that available in a Python programming language. The crawl engine was created based on the result of data structure analysis in the previous step. The algoritma of the crawl engine is as follows:

Algoritma of the crawl engine

```
keyword input
make a request based on the keyword input
get the json data of keywords suggestion
repeat
    keyword suggestion input
    make a request based on the keyword suggestion input
    get the html data of keywords suggestion
    find the tag div with id "result-stats" and get the value (allintitle value)
    save the keyword suggestion and the allintitle to the database
until all the keyword suggestion data
```

The data used as an experiment in this test consists of ten keywords with various fashion themes. Each of this keyword will generate a maximum of ten keywords suggestion. Therefore, from ten keywords, it will produce at most 100 keywords suggestions and the number of allintitles. The ten of keywords is shown in Table 1. In this study, the evaluation was carried out by measuring the extent to which the web scraping method was applied to extract the keyword suggestion data and the "allintitle" following the actual. The result of the comparison between the two is shown in Table 2. According to Table 2, we could calculate the accuracy using Equation 1. The accuracy of the web scraping technique to get keywords suggestions and allintitle from Google search engines was 100%.

Table 1. The Keywords Data for The Experiment

No.	Keywords
1	dresses for juniors
2	short pants outfit
3	short sleeve tops
4	casual shoes women
5	beautiful long skirts
6	long sleeves sweater
7	bags for brides
8	popular girl shoes
9	baby dress girl
10	teenage outfits girl

Table 2. The Result of The Comparison

No.	Implementation Result of The Web Scaping Techniques		Ground Truth		Evaluation
	Keyword Suggestion	Allintitle	Keyword Suggestion	Allintitle	
1	dresses for juniors	78,400	dresses for juniors	78,400	appropriate
2	dresses for juniors formal	36,000	dresses for juniors formal	36,000	appropriate
3	dresses for juniors graduation	771	dresses for juniors graduation	771	appropriate
4	dresses for juniors near me	387	dresses for juniors near me	387	appropriate
5	dresses for juniors party	12,400	dresses for juniors party	12,400	appropriate
6	dresses for juniors casual	3,480	dresses for juniors casual	3,480	appropriate
7	dresses for juniors plus size	1,330	dresses for juniors plus size	1,330	appropriate
8	dresses for juniors jcpenney	4,050	dresses for juniors jcpenney	4,050	appropriate
9	dresses for juniors semi formal	1,100	dresses for juniors semi formal	1,100	appropriate
10	dresses for juniors kohls	223	dresses for juniors kohls	223	appropriate
11	short pants outfit	12,200	short pants outfit	12,200	appropriate
12	short pants outfit ideas	183	short pants outfit ideas	183	appropriate
.....					
99	party outfits teenage girl	0	party outfits teenage girl	0	appropriate
100	fashionable teenage girl outfits	8	fashionable teenage girl outfits	8	appropriate
Total of the appropriate result					**100**

4. CONCLUSION

Implementing the web scraping technique to get keywords suggestions and allintitle from Google search engines give an accuracy of 100%. It indicates that the technique can be applied to get keywords suggestions and allintitle from Google search engines with outstanding accuracy values. In future work, we will develop it into an application that can be used freely by anyone who needs it, especially business people, to do keyword research.

ACKNOWLEDGEMENTS

The author would like to thank the Directorate of Research and Community Service Directorate General of Research and Development Strengthening (DRPM) of the Ministry of Research and Higher Education (Kemristekdikti) of the Republic of Indonesia for the support provided to the author in the form of research funding assistance in 2021 of Penelitian Dosen Pemula (PDP) scheme.

REFERENCES

[1] T. H. Sinaga, T. H. Sinaga, and E. Hadinata, "Implementasi Teknik Search Engine Optimization Dalam Meningkatkan Trafik Website Bima Utomo Waterpark," *Query J. Inf. Syst.*, vol. 3, no. 2, Oct. 2019.

[2] H. Himawan, A. Arisantoso, and A. Saefullah, *SEARCH ENGINE OPTIMIZATION (SEO) MENGGUNAKAN METODE WHITE HAT SEO UNTUK MENINGKATKAN PERINGKAT DAN TRAFIK KUNJUNGAN WEBSITE*, vol. 0, no. 0. 2017.

[3] M. Maskur, "Pengukur Parameter Search Engine Optimization (SEO) Secara On Page Pada Toko Online Untuk Meningkatkan Penjualan," *J. Adm. Dan Bisnis*, vol. 13, no. 1, pp. 83–91, 2019.

[4] A. Sofyan, E. Ferdianto, R. Rahmawati, and R. K. Aldi, "Pengaruh Search Engine Optimization (SEO) Dan Riset Kata Kunci Terhadap Pendapatan Toko Online," in *INCONTECSS*, 2019, pp. 351–356.

[5] A. I. Hadiana and E. K. Putra, "Model Search Engine Optimization bagi Usaha Mikro Kecil dan Menengah (UMKM) di Bandung Barat," *J. Masy. Inform. Unjani*, vol. 2, no. 1, pp. 31–38, 2018.

[6] S. N. Wahyuni and D. A. Wijaya, "Penerapan dan Optimasi Riset Keyword Dengan Teknik Allintitlle Pada Mesin Pencari Google," *J. Mantik Penusa*, vol. 2, no. 2, pp. 40–44, 2018.

[7] R. Fattahi, M. Parirokh, M. H. Dayyani, A. Khosravi, and M. Zareivenovel, "Effectiveness of Google keyword suggestion on users' relevance judgment: A mixed method approach to query expansion," *Electron. Libr.*, vol. 34, no. 2, pp. 302–314, Apr. 2016, doi: 10.1108/EL-03-2015-0035.

[8] M. I. Akrianto, A. D. Hartanto, and A. Priadana, "The Best Parameters to Select Instagram Account for Endorsement using Web Scraping," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2019, pp. 40–45, doi: 10.1109/ICITISEE48480.2019.9004038.

[9] A. Priadana and A. W. Murdiyanto, "Instagram Hashtag Trend Monitoring Using Web Scraping," *J. Pekommas*, vol. 5, no. 1, p. 23, Apr. 2020, doi: 10.30818/jpkm.2020.2050103.

[10] A. Priyanto and M. R. Ma'arif, "Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik)," *Indones. J. Inf. Syst.*, vol. 1, no. 1, pp. 25–33, Aug. 2018, doi: 10.24002/ijis.v1i1.1664.

[11] D. J. Tandra, A. Noertjahyana, and A. N. Purbowo, "Implementasi Web Scraping untuk Pengumpulan Informasi Promo Makanan Menggunakan Klasifikasi Naïve Bayes," *J. Infra*, vol. 8, no. 1, pp. 289–294, Apr. 2020.

[12] D. D. A. Yani, H. S. Pratiwi, and H. Muhardi, "Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace," *J. Sist. dan Teknol. Inf.*, vol. 7, no. 4, p. 257, Oct. 2019, doi: 10.26418/justin.v7i4.30930.

[13] A. Pranav and S. Chauhan, "Efficient Focused Web Crawling Approach for Search Engine." 2015.

[14] S. Upadhyay, V. Pant, S. Bhasin, and M. K. Pattanshetti, "Articulating the construction of a web scraper for massive data extraction," in *Proceedings of the 2017 2nd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2017*, 2017, doi: 10.1109/ICECCT.2017.8117827.

[15] C. Slamet, R. Andrian, D. S. Maylawati, S. Suhendar, W. Darmalaksana, and M. A. Ramdhani, "Web Scraping and Naïve Bayes Classification for Job Search Engine," in *IOP Conference Series: Materials Science and Engineering*, 2018, pp. 1–7, doi: http://doi.org/10.1088/1757-899X/288/1/012038.

[16] A. K. Arsyad, B. Pramono, M. . Isnawaty S.Si., M. Yamin, and I. Sarita, "Implementasi Levenshtein Distance Pada Aplikasi Pencarian Barang Di Berbagai E-Marketplace Menggunakan Teknik Web Scraping," *Semin. Nas. APTIKOM 2019*, vol. 0, no. 0, pp. 512–519, Nov. 2019.

[17] R. C. Pereira and T. Vanitha, "Web Scraping of Social Networks," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 3, no. 7, pp. 237–240, 2015.

[18] Fatmasari, Y. N. Kunang, and S. D. Purnamasari, "Web Scraping Techniques to Collect Weather Data in South Sumatera," in *Proceedings of 2018 International Conference on Electrical Engineering and Computer Science, ICECOS 2018*, 2019, doi: 10.1109/ICECOS.2018.8605202.