Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 2, September 2023, pp. 521-530
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Aplikasi Web Untuk Visualisasi Web Scraping Menggunakan Metode VSM

Andi Nurkholis[1], Yusra Fernando[2], Faris Arkan Ans[3]
[1,2,3]Informatika, Universitas Teknokrat Indonesia, Indonesia
e-mail: andinh@teknokrat.ac.id[1], yusra.fernando@teknokrat.ac.id[2], faris.arkan.mhs@teknokrat.ac.id[3]

Abstract

The internet, like the workplace, is at the heart of every aspect of communal life in the modern digital era. Many platforms already provide job vacancies, especially for independent contractors. To identify relevant job openings, consumers typically need to access multiple websites to gather this information. One way to overcome this problem is to use web scraping. BeautifulSoup and Selenium libraries will be used to collect data in accordance with previous research findings. The vector space model approach is used to determine the degree of data similarity between queries and documents to perform data searches. After examining the data, a mean accuracy value of 56% was obtained and a mean perfect recall value of 100%. This is because, even if the context does not match, data searches use three parameters, increasing the likelihood of returning irrelevant material if the document contains words from the user's query. Users can manage the processes of web scraping, data processing, and data searching with the help of the Streamlit framework in Python, which displays the results of data processing. To obtain data from the Sribulancer, Project and freelancing freelancer websites, this research will use web scraping techniques. Users can search for data from multiple websites using a vector space model approach rather than accessing each loose website one by one. Web scraping results can also be processed to be displayed in a more user-friendly format and save time by using data visualization in the form of a web application built using the Streamlit framework.

Keywords: *Web Scraping, freelance job, Python, VSM method*

1. INTRODUCTION

The development of information technology is getting more advanced every time. Thanks to the internet network as a data communication medium, everything has entered the digitalization era. The internet has become a center for community activities, including in the field of work. In general, a job or a business needs an office place/location and has permanent workers/employees. However, this is no longer urgent because freelance sites can be used as platforms connecting workers and business owners. Without an office, workers can work flexibly anywhere and at any time with the provision that they are connected via the internet [1]. This profession is often called a freelancer [2], which has recently been in great demand by professional workers who can be found on the Freelance, Project, and Sribulancer websites. Indonesia is the 16th country on the list of countries with the largest economies, with 55 million professional workers. By 2030, this number is estimated to increase to 113 million people. This is what makes Indonesia predicted to be the seventh-largest economy in the world [3].

Currently, many third-party platforms specifically provide vacancies for freelancers. Problems arise when users need to open multiple websites to find information about suitable job vacancies. Few freelance vacancies provide information about their vacancies on only one website. Thus, many freelance websites will make it difficult for freelancers to summarize all vacancy

information. These problems can be solved using web scraping techniques that aim to extract the web page framework to collect data [5] that can be processed into summary information in a short time [6], [7].

Many studies have examined the implementation of web scraping. The first study implemented Web Scraping on Hospitality Sites, which succeeded in automatically providing large amounts of data samples for further analysis [8]. Other research applies crawling techniques using Selenium, which automatically performs navigation and user input commands on web pages [9]. In other studies, the results of implementing web scrapping can also be followed up by applying the vector space model (VSM) method for visualizing karaoke song information based on user searches combined with sorting based on the level of word similarity [10]. VSM is used to see the degree of closeness or similarity of terms using term weighting. Based on previous research, the web scrapping technique combined with the VSM method can extract data, which is then visualized according to user needs.

This study aims to implement web scraping techniques for data extraction on three freelance websites, namely Freelance, Project, and Sribulancer. In addition, the VSM method is also used to visualize results in web-based applications by utilizing the Streamlit framework. To complete the research conducted is divided into two main stages, namely web scraping and data visualization. As an implication, a summary of all freelance job vacancies can be displayed efficiently according to user needs to save more time.

2. RESEACH METHODOLOGY

The object of this research consists of the three largest freelance websites in Indonesia, namely Freelance, Project, and Sribulancer. The most significant amount of traffic evidence this from several other freelance websites accessed by the people of Indonesia. The main stages of this research are divided into 2, namely web scraping and data visualization. The web scraping stage consists of the process of data collection and data storage. At the same time, the data visualization process consists of Data Preprocessing and Query, TF-IDF Weighting, Document Ranking with Vector Space Model, and Data Visualization. The following are the stages of the research which can be seen in Figure 1.

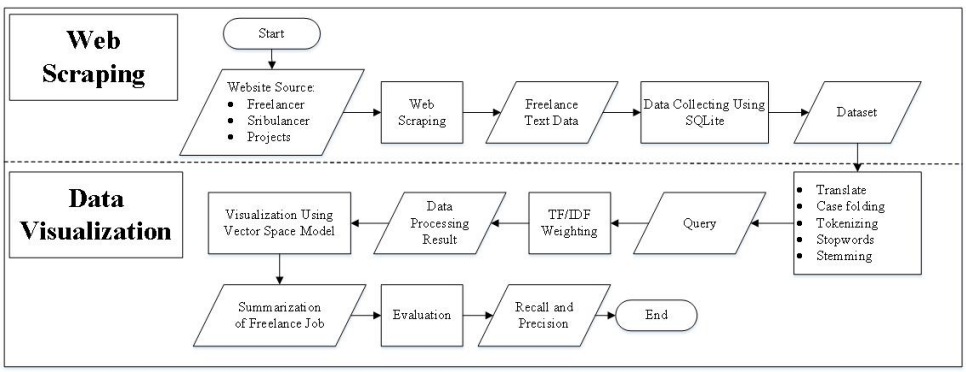

Figure 1. Research stages

Based on Figure 1, the following is an explanation of each stage:

2.1. Web Scrapping

The following are the stages carried out in the web scraping stage, shown in Figure 2.

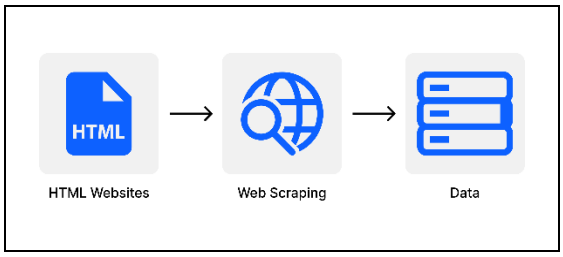

Figure 2. Stages of web scraping

a. Web Scrapping
   Web scraping is an activity carried out to retrieve specific data in a semi-structured manner from a website page [11]. In this study, the web scraping stage of freelance job vacancies was carried out on three websites using the BeautifulSoup and Selenium libraries in the Python programming language version 3.8.1.
b. Data Collecting
   Next is storing data from the web scraping results in the database. In this research, data is stored in SQLite database version 3.39.4.

2.2. Data Visualization

This stage aims to process data from web scraping and visualize the summarization results using the vector space model method. The stages of data visualization are divided into four stages, which are explained as follows:

a. Data Preprocessing
   The data preprocessing stage aims to prepare data to change raw data [12], commonly known as raw data collected from various sources, into cleaner information that can be used for further processing. This process can also be called the first step to retrieve all available information by cleaning, selecting, and combining the data. Data preprocessing is very important because errors, redundancy, missing values, and inconsistent data reduce the accuracy of sentiment analysis results. The data collection results cannot be directly modeled because there are still many unnecessary symbols and words, so data preprocessing is needed so that the data is more structured and clean to be classified [13]. Meanwhile, the data preprocessing steps include translation, case folding, tokenizing, stopwords, and stemming.
   1) The translation is a transfer process step that aims to change the source language's written text into the target language's text. In this research, the English text is translated into Indonesian, which is the ultimate goal of information visualization. This was done because a lot of freelance

work data has been written in English, so it needs to be translated into Indonesian to be uniform. The data translation process is carried out by utilizing the deep_translator library in the Python programming language.

2) Case folding is a step to convert all responses into lowercase letters. In this process, the characters 'A'-'Z' in the data are changed to characters 'a'-'z'. Meanwhile, other characters that do not include letters and numbers, such as punctuation marks and spaces, are considered delimiters.
3) Tokenizing is a step to delete data containing punctuation marks to produce sentences/words that stand alone. Entities that can be referred to as tokens, for example, words, numbers, symbols, punctuation marks, etc. That is, this stage aims to break down responses into units of words.
4) Stopword is a step to remove common words that often appear in large quantities and have no meaning using a stoplist algorithm (removing less important words) or wordlist (saving essential words). For example, conjunctions such as 'and', 'which', 'and', 'after', and others. Eliminating these stopwords can reduce index size and processing time. In addition, it can also reduce the noise level.
5) Stemming is the step of changing each affix word into a base word. This stage is needed to minimize the number of different indexes from one data so that a word with a suffix or prefix will return to its basic form. In addition, it is also used to group other words with the same base word and meaning but has a different form because they get different affixes.

b. Term Frequency – Inverse Document Frequency Weighting
Term Frequency-Inverse Document Frequency (TF-IDF) is a technique for assigning a term's relationship weight to a document. The TF-IDF method works by combining two concepts for calculating weights: the frequency of occurrence of a term in a document (TF) and the frequency inversion of documents (IDF) containing that word. The frequency with which a word appears in a given document indicates how important the word is in that document. The number of times the document contains the word means how common the word is. The fewer the number of documents containing the term in question, the greater the value in the IDF.

c. Data Visualization Using Vector Space Model
At this stage, the data will be displayed using a web application. In addition to showing the results of ranking documents and queries, this stage will also display all web scraping results from all websites integrated with the database. The Vector Space Model (VSM) is a method used to measure the degree of closeness or similarity of terms using weighting terms [14]. Documents are assumed to be vectors that have magnitude and direction. In this method, a term is represented by a vector space dimension [15]. The terms used are generally based on the terms in the query or keywords. The relevance of a document to a query is based on the similarity between the document vector and the query vector. VSM and TF-IDF weighting

represent the numerical values of documents so that the closeness between documents can then be calculated [16]. The closer the two vectors are in a VSM, the more similar the two documents represented by the vector are. The similarity between documents is calculated using a similarity measure function [17]. This size allows the ranking of documents according to their similarity in relevance to the query.

d. Recall and Precision Evaluation
Based on previous research, testing the vector space model will use recall and precision evaluations. Precision can be considered a measure of accuracy or thoroughness, while recall is perfection. In this study, the recall value was obtained based on the calculation of the summoned document according to the user's request. At the same time, the precision value is obtained based on the count of the number of records retrieved from the relevant database after being assessed by the user with the information needed.

3. RESULT AND DISCUSSION

3.1. Data Preprocessing and TF-IDF Weighting

The dataset contains text from 3 freelance websites using the Python programming language. The web scraping process for Projects and Freelancers uses Beautifoulsoup, while for Sribulancer, it uses Beautifoulsoup and Selenium because the Sribulancer website requires a login process, and almost all of the data is displayed using javascript. Furthermore, the preprocessing data includes translating, case folding, stop wording, stemming, and tokenizing. The following is an example of the final preprocessed data shown in Table 1.

Table 1. Final dataset

Before	After
perlu dokumen terjemah bahasa spanyol bahasa inggris butuh sertifikasi sedang cari terjemah dapat terjemah bahasa spanyol bahasa inggris kirim format pdf kurang halaman butuh tawar freelancer baru untuk proyek harga negosiasi	[perlu], [dokumen], [terjemah], [bahasa], [spanyol], [bahasa], [inggris], [butuh], [sertifikasi], [sedang], [cari], [terjemah], [dapat], [terjemah], [bahasa], [spanyol], [bahasa], [inggris], [kirim], [format], [pdf], [kurang], [halaman], [butuh], [tawar], [freelancer], [baru], [untuk], [proyek], [harga], [negosiasi],

After the pre-processing of the data is complete, it is followed by TF-IDF weighting to convert text data into numeric data so that the number of occurrences of words can be calculated and also to calculate the weight of each word. The following is an example of the TF-IDF weighting results, as seen in Figure 3.

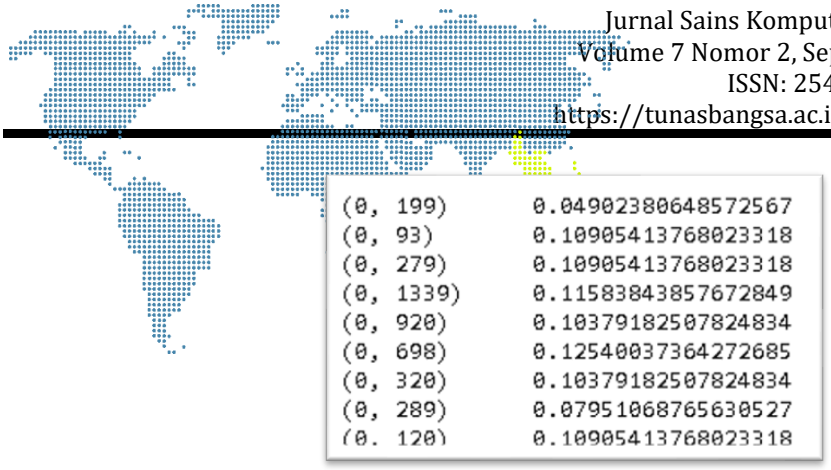

Figure 3. TF-IDF weighting

The first value in Figure 3 represents the index of the document. The second value represents the index of the term (word), while the third value represents the IDF (inverse document frequency) calculation results.

3.2. Data Visualization Using VSM Method

At this stage, a web-based user interface was developed. This stage aims to display scraped data and make it easier for users to navigate the process of scraping data and processing data. The user interface development uses the Python Streamlit library consisting of three main menus: data search, data preprocessing, and scraping results. Here is an example of a page that has been developed.

a) The Scraping Data menu functions to carry out the web scraping process and download the results of imported scraping data. Users can select the data source to be scraped, the amount of data, the option to export data from the database to excel, and the option to empty the database when scraping. The data scraping menu can be seen in Figure 4.

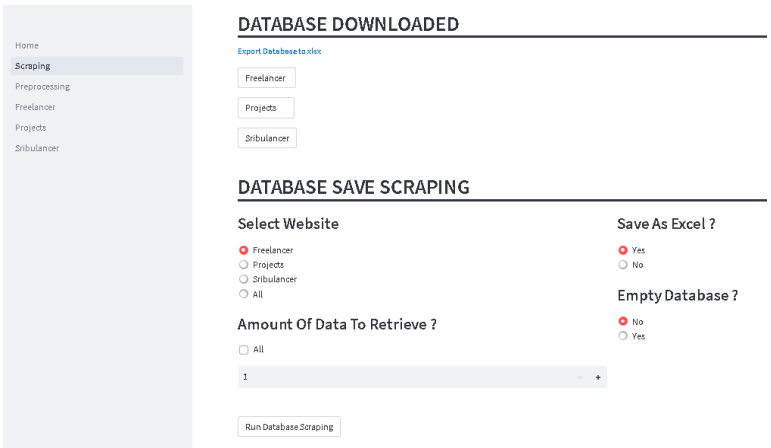

Figure 4. Scraping data menu

b) The Preprocessing Data menu functions to do preprocessing and view and download preprocessing result data that has been exported to excel. Data from web scraping must go through the preprocessing stage to search data. The data preprocessing menu is shown in Figure 5.

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 2, September 2023, pp. 521-530
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

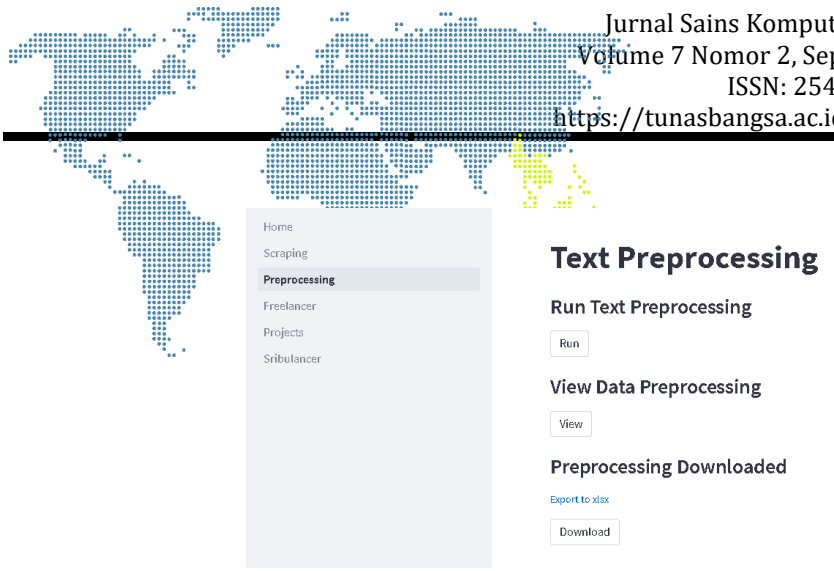

Figure 5. Preprocessing data menu

c) The display of the Web Scraping Results Menu in Figure 6 is an example of scraping results from the Sribulancer website. This menu also provides information in the form of wage graphs based on each job category.

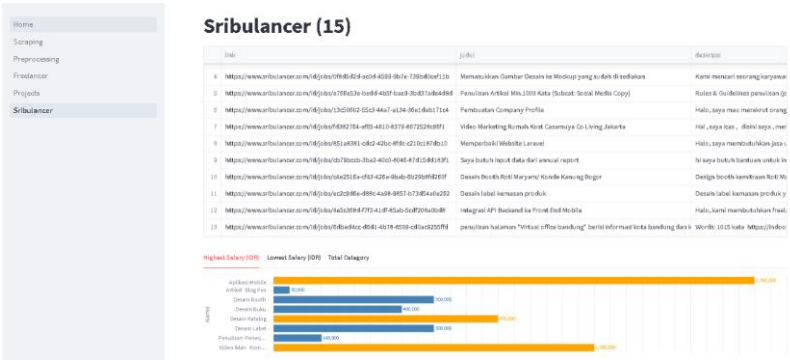

Figure 6. Web scraping result menu

d) The Data Search menu functions to display search results for freelance jobs by entering a query, and users can also choose how much data they want to display. The search data results are sorted based on the value of cosine similarities from the highest to the lowest. Data search menu can be seen in Figure 7.

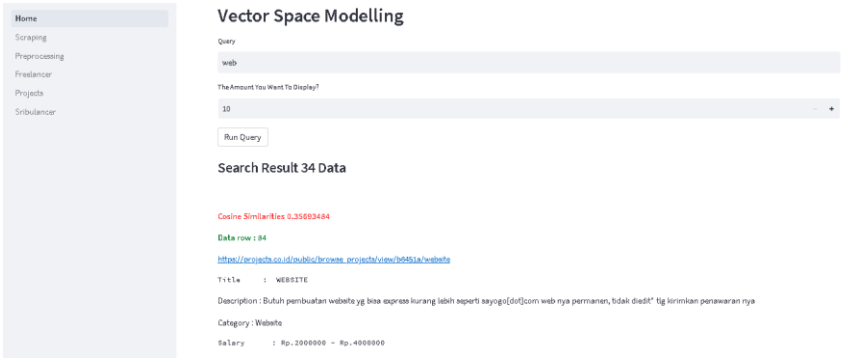

Figure 7. Data search menu using VSM

3.3. Evaluation of Visualization Model

Document search testing using the vector space model is carried out by calculating recall and precision values to determine the method's performance in searching for relevant documents. The test uses 31 documents by entering five keywords related to documents. The test results for calculating recall and precision values can be seen in Table 2.

Table 2. Evaluation result

Keyword	Result			Recall	Precision
	Relevant	Irrelevant	Not Found		
Program web	5	1	0	100%	83%
Desain grafis	11	6	0	100%	65%
Penulisan artikel	7	7	0	100%	50%
Edit video	5	5	0	100%	50%
Terjemah bahasa inggris	4	9	0	100%	31%
Average				100%	56%

Based on the test results in Table 2, an average recall value of 100% is obtained, while the average precision value is 56%. This relates to the number of documents that were successfully indexed. The greater the number of documents successfully indexed, the greater the number of relevant documents will affect the recall value. However, the greater the amount of noise (or irrelevant documents), the lower the precision level. In addition, because the author uses three parameters for data retrieval based on the title, description, and category, it greatly affects the value of recall and precision. So the possibility of finding data will, of course, be even greater because it uses three parameters, but this greatly affects the level of precision because if an irrelevant document contains data in the query, the document will be retrieved even though the words in the document are irrelevant in the context intended by the user.

In the case of searching for data with the keyword "web program", we managed to find all relevant documents because all of these documents contained the word "program" or "web" from the three parameters used. Meanwhile, a document titled "Dicari Blog/web Indo ataupun bule untuk di beli" was found for irrelevant documents. This document can be displayed when searching for data because the document's title contains the word "web", even though the document contains website purchases and is not what the user intended.

4. CONCLUSION

This study successfully used web scraping to collect data from 3 freelancer web. However, carrying out the web scraping process on the Sribulancer site requires a good connection because the data is displayed dynamically using JavaScript. In the appearance of data, the Streamlit framework is used for presenting data and helping users to navigate the process of collecting and processing data. Searching for data using the Vector Space Model method on

scraping results produces an average perfect recall value of 1 or 100%, while the average precision value is 56%. The data retrieval process uses three parameters, namely: title, description, and category of freelance vacancies data, to increase recall results in data search, but this significantly affects the precision value because the possibility to retrieve irrelevant data is also more significant if the document contains a word on query user even if the context doesn't match. The research produced a web application that can carry out the web scraping process, search for freelance job vacancy information from several websites simultaneously and display the results of dataset processing in a more concise and helpful form.

REFERENCES

[1] W. Wahyu Agung Firrezqi, "Peran Situs Freelance Project. co. id Dalam Membantu Masalah Perekonomian di Indonesia," *Analisis Peran Situs Freelance Project. co. id Dalam Membantu Masalah Perekonomian di Indonesia*, vol. 2, no. 2, pp. 1–8, 2020.

[2] M. Ayoobzadeh, "Freelance job search during times of uncertainty: protean career orientation, career competencies and job search," *Personnel review*, vol. 51, no. 1, pp. 40–56, 2022.

[3] M. Mustofa, "Pekerja Lepas (Freelancer) dalam Dunia Bisnis," *Jurnal MoZaiK*, vol. 10, no. 1, pp. 19–25, 2018.

[4] S. Kadam, S. Shinde, A. Sharma, S. Mali, and B. E. Student, "Price comparison of computer parts using web scraping," *Int. J. Eng. Sci*, 2018.

[5] K. Henrys, "Importance of web scraping in e-commerce and e-marketing," *Available at SSRN 3769593*, 2021.

[6] R. Ridwan and T. A. Hermawan, "Penerapan mesin pencari informasi dengan menggunakan metode Vector Space Model," *Jurnal Teknik Informatika (JUTEKIN)*, vol. 7, no. 2, 2019.

[7] M. Eminagaoglu, "A new similarity measure for vector space models in text classification and information retrieval," *J Inf Sci*, vol. 48, no. 4, pp. 463–476, 2022.

[8] S. Han and C. K. Anderson, "Web scraping for hospitality research: Overview, opportunities, and implications," *Cornell Hospitality Quarterly*, vol. 62, no. 1, pp. 89–104, 2021.

[9] J. N. Semendawai, I. Febiola, B. Pamungkas, and M. D. Ruliansyah, "Perancangan Aplikasi Otomatisasi Menggunakan Bahasa Pemrograman Python Pada Aktivitas Monitoring Pemakaian Data Harian Kartu Internet Of Things," *Jurnal Rekayasa Elektro Sriwijaya*, vol. 3, no. 1, pp. 193–198, 2021.

[10] A. Anna and A. Hendini, "Implementasi vector space model pada sistem pencarian mesin karaoke," *Evolusi : Jurnal Sains dan Manajemen*, vol. 6, no. 1, Mar. 2018, doi: 10.31294/evolusi.v6i1.3535.

[11] Y. Julianto, D. H. Setiabudi, and S. Rostianingsih, "Analisis Sentimen Ulasan Restoran Menggunakan Metode Support Vector Machine," *Jurnal Infra*, vol. 10, no. 1, pp. 1–7, 2022.

[12] A. Nurkholis, D. Alita, and A. Munandar, "Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 2, Apr. 2022.

[13] A. Nurkholis, Z. Abidin, and H. Sulistiani, "Optimasi Parameter Support Vector Machine Berbasis Algoritma Firefly Pada Data Opini Film," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 5, no. 5, pp. 904–910, 2021.

[14] F. Amin, "Sistem Temu Kembali Informasi dengan Pemeringkatan Metode Vector Space Model," *Dinamik*, vol. 18, no. 2, 2013.

[15] G. Sidorov and G. Sidorov, "Vector Space Model for Texts and the tf-idf Measure," *Syntactic n-grams in Computational Linguistics*, pp. 11–15, 2019.

[16] B. P. Zen, I. Susanto, and D. Finaliamartha, "TF-IDF Method and Vector Space Model Regarding the Covid-19 Vaccine on Online News," *Sinkron: jurnal dan penelitian teknik informatika*, vol. 6, no. 1, pp. 69–79, 2021.

[17] M. A. Azis, A. Hamid, A. Fauzi, E. Yulianto, and V. Riyanto, "Information retrieval system in text-based skripsi document search file using vector space model method," in *Journal of Physics: Conference Series*, 2019, vol. 1367, no. 1, p. 012016.