Ilkogretim Online - Elementary Education Online, 2021; 20 (3): pp. 432-441
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.43

Implementation of Web Scraping on News Sites Using the Supervised Learning Method

Dedy Rahman Prehanto , *University of Surabaya, Dept. of Information System – Indonesia*
Aries Dwi Indriyanti, *University of Surabaya, Dept. of Information System – Indonesia*
I Gusti Lanang Eka Prismana, *University of Surabaya, Dept. of Information Technology – Indonesia*
Ginanjar Setyo Permadi, *University of Hasyim Asy'ari, Dept. of Informatics Management – Indonesia*
Edwin Hari Agus Prastyo, *University of Hasyim Asy'ari, Dept. of Information Technology – Indonesia*

Abstract. Indonesia is one of the highest internet users in the world, including in the penetration of information on the internet, online news media. But in general news sites not only display news information, but most sites also display other information such as advertisements and also forms of navigation that interfere with news site readers and interfere with reader's comfort, from these problems this study aims to implement web scraping techniques with supervised learning methods and analyzing the form of DOM tree and XPath news sites. The supervised learning approach method is the method used in this study, which is one of the methods of machine learning. By combining these web scraping techniques with supervised learning, the aim is to be able to implement and optimize web scraping techniques to gather news information from various sites. To do basic web scraping namely knowing DOM patterns, XPath structure as a data model or selector at each site. The results of research in the form of a web scrap application that can retrieve news site content without copy paste and the data is stored in a database and displayed to the user application form for the reader without any ads and navigation that disturb the reader.

Keywords: web scraping, supervised learning, XPath, DOM tree.

Received: 05.12.2020	Accepted: 10.01.2021	Published: 05.02.2021

INTRODUCTION

Indonesia is one of the highest internet users in the world. Based on the infographic survey agency for penetration & behavior of Indonesian internet users, it reaches 68% of the 264.18 million Indonesian population, many users penetrate information via the internet. Information is the most important thing in life, sources of information can be through news sites(Permadi et al., 2018). A news site is a website that provides various information from various news sources to be displayed to users(Permadi et al., 2019).
However, in general, sites do not only display information from outside sources, sometimes they also display information on their own website such as advertisements and navigation that disturbs news site readers and disturbs the convenience of the readers, and also some people may still collect data on the website with copying one by one on the website, but if the website that you manage is a large site with thousands of data, of course the work takes a long time, so to solve this problem the researchers conducted one of the web content extraction techniques using machine learning(Dedy Rahman Prehanto et al., 2019).

In classification, supervised learning is carried out by conducting training (training) to form a model(D. R. Prehanto et al., 2019). Classifier (classification algorithm) will form a model that adapts according to the features that exist in the data(Mashuri et al., 2019). The purpose of applying the supervised learning method is to implement and optimize web scraping techniques, as well as make it easier for users to process information from the results of web scraping techniques used to gather news information from various sites(Xia et al., 2017). To do web scraping, the basic thing is to know the DOM tree structure on news sites(Iacus, 2015).

The researcher analyzed the DOM tree and Xpath patterns which were later used as materials in the training data(Niculescu-Mizil & Caruana, 2005). DOM tree is a standard html document structure to form a web page, while Xpath is a query language from the representation of the DOM resulting in a particular node(Hardt et al., 2016). Rizaldi's research states that the use of Xpath is better than the CSS selector, therefore this study uses Xpatht.

So this study takes the title Implementation of Web Scraping on News Sites Using the Supervised Learning Method. Below are some definitions related to this research.

.

METHODOLOGY

Web scraping

Web scraping is a process of data collecting through humans using a web browser, to handle the web that does not provide an API, web scraping requests data in HTML tags and then parses that data to extract the required information(Nylen & Wallisch, 2017). In the implementation, the use of web scraping includes programming techniques and technologies, such as data analysis and information security. Web scraping retrieves HTML data from a domain name, parses that data for target information, optionally stores target information, and moves to another page to repeat the scraping process(S.C.M. de S Sirisuriya, 2015).

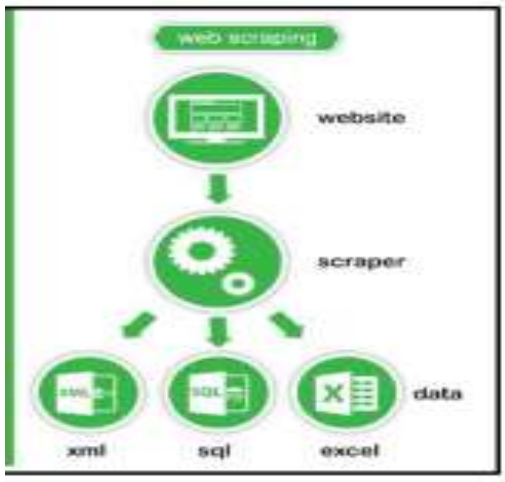


Figure 1

Web Scraping

Web scraping describes the use of a program for extracting data from HTML files that exist on a particular site on the internet. Web scraping is the term used to describe the data extraction process on electronic files using a computer program(Ulbricht, 2020).

The next research was carried out by Rizaldi, explaining the crawling technique using Scrapy and Xpath, in this research XPath managed to meet the expected target. This is indicated by the production of a news corpus that has been divided into 3 news categories namely entertainment, community and culinary news(Khalil & Fakir, 2017).

Web scraping has several stages, the first step is to make scraping templates, explore the site navigation, automate navigation and extract information and save it(Mitchell, 2015).

Illustrating the steps for performing web scraping, starting with determining which sites will be used as scraping objects, then analyzing the page structure of the site to create a template(Aries Dwi Indriyanti et al., 2019). Then carry out exploration and navigation, the results of this process are then automated for extraction(A. D. Indriyanti et al., 2019). The final step is to save the scraping that has been done. The process of extracting information before it is saved can be done by utilizing the XPath method to find the desired information with a pattern according to a predefined template(Kistofer et al., 2019).

Figure 2

Web scraping steps

Content Extraction using Machine Learning

Content extraction using Machine Learning which is part of the branch of artificial intelligence. This includes the construction and study of training data systems as an introductory model, the essence of machine learning is related with representation and generalization. Generalization is a trait that the system will perform well on unseen data(Mutasim et al., 2018).

Supervised Learning

Supervised learning is a method for classifying each object in the data into several classes(Iacus, 2015). In supervised learning, each object in a data has features, namely the features that exist in each object. Each object in a data has the same number of features(Huber & Stuckenschmidt, 2020). Features are used as input to define a class on an object. In supervised learning, the class of each object is known(Huber & Stuckenschmidt, 2020). Therefore, the problem faced in supervised learning is how to map the objects into the right class using the features that each object has(Grandvalet & Bengio, 2005).

Classification in supervised learning is done by conducting training (training) to form a model. Classifier (classification algorithm) will form a model that adapts according to the features that exist in the data(Laine & Aila, 2017). The resulting model can be in the form of a tree, rule, or a function that can predict a class based on the features the data has(You et al., 2018)

Python

Python is a desktop or web-based programming language, a popular programming language used by many developers(Chollet, 2018). Python is one of the popular programming languages used by many developers. A survey according to the programming language website via www.tiobe.com, Python was ranked 4th in 2019(GCM et al., 2016). Python can also be used for enterprises. Pyhton is included in the programming language level, including high level language(Buduma & Locascio, 2017). Python is a programming language that can be used to build applications, be it desktop-based, web-based or mobile-based.

In python, there is a moduleused for the data extraction process so that the results obtained are faster and more practical. Python modules used in this study are(PF et al., 2016):

a. Urllib is a module for handling request urls, the url you want toextract will be handled by urllib.
b. BeautifulSoup is a module for parsing HTML. HTML parsing is a technique for separating text from HTML code tags on a website page to produce specific data

RESULTS AND DISSCUSION

In this study, using processing data derived from online news sites totaling 100 pages of news sites. List of news sites that extract web content in the table.

Table 1. *list of news sites*

No	Name Web	Url
1	Media Indonesia	https://mediaindonesia.com
2	Kompas	https://www.kompas.com
3	Bisnis Indonesia	https://teknologi.bisnis.com
4	Pikiran Rakyat	https://www.pikiran-rakyat.com
5	Cek & Ricek	https://ceknricek.com
6	Siwalima	https://siwalimanews.com
7	Waspada	http://waspada.co.id
8	Analisa	https://analisadaily.com
9	Tribun Timur	https://makassar.tribunnews.com
10	Kedaulatan Rakyat	https://krjogja.com
11	Harian Jogja	https://jogjapolitan.harianjogja.com
12	Suara Merdeka	https://www.suaramerdeka.com
13	Solo Pos	https://www.solopos.com
14	Koran Sindo	https://jateng.sindonews.com
15	Sindo Weekly	https://sumeks.co
16	Sumatera Ekspres	http://www.sindoweekly.com

17	Radar Palembang	http://www.radar-palembang.com
18	Tribul Sumsel	https://sumsel.tribunnews.com
19	Palempang Ekspres	https://palembang.tribunnews.com
20	Republika	https://republika.co.id
21	Antara	https://www.antaranews.com
22	Okezone	https://www.okezone.com/
23	merdeka	https://www.merdeka.com/
24	detik	https://www.detik.com/
25	liputan	https://www.liputan6.com/

From all the list of sites in the research object, only certain data will be extracted and saved into several variables. The first variable is a link that contains a link to a web page that contains news. second is the headline, which contains news headlines, next is the content of news content, and the fourth is the XPath Selector variable.

DOM tree and Xpath tree structure analysis

In this analysis manual tag classification and after observing the html structure and tags, an example of an HTML document modeled as a tree is shown in Figure.

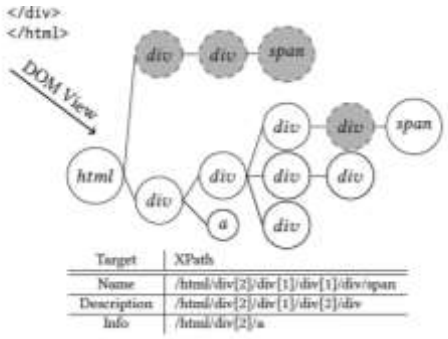


Figure 3 . *HMTL document and model tree*

In this observation, the researcher conducted modeling from thehtml document for each website page and the tree is modeled to get the desired tags in the 70 news website web pages. As an example of Figure 4 scraping hml tags.

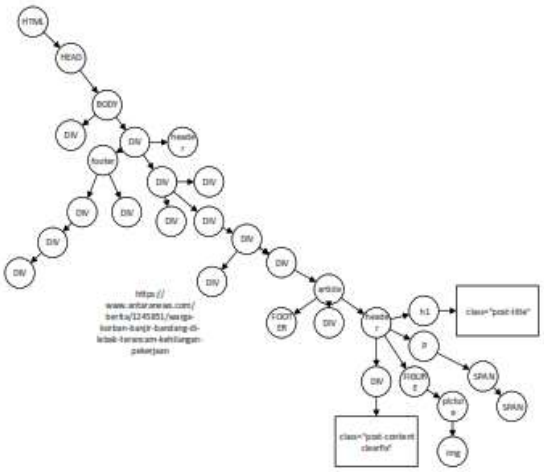


Figure 4 . *DOM tree structure*

The results of research on the implementation of web scraping techniques on news sites using the supervised learning method are as follows.
a. Knowing the DOM tree and XPath patterns contained in a news site.
b. Admins can manage news sites and manage news categories.
c. Admins can manage the news site's XPath model selector.

d. Admins can manage the target page of the news site to be scraped or extracted from the headlines and content.
e. Users can read scraping results containing news headlines and news content in the newsscrapp mobile web application.
f. Users can read news according to the categories provided and search for news according to the headline.

The results of the analysis of the DOM tree structure and site XPath

In retrieval of news content two XPath sections are taken the first XPath the title and the second XPath the contents, the XPath laying titles before content on the website has a different structure forms so that this analysis is carried out.

a. Media Indonesia

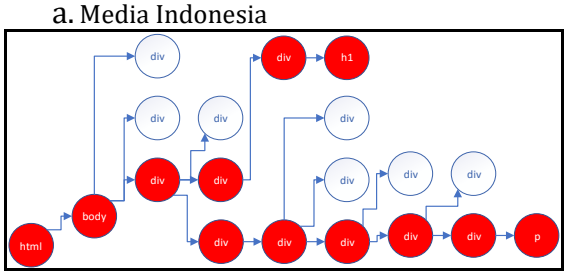
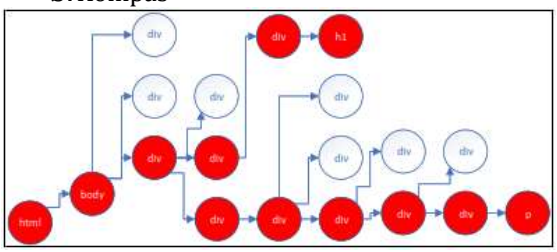


Figure 5
Dom tree media Indonesia

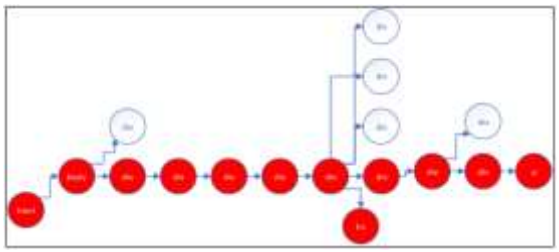
No	Target	XPath
1	Headline	/html/body/div[3]/div[2]/div/h1
2	Content	/html/body/div[3]/div[3]/div[1]/ div[3]/div[2]/div[2]/p

b. Kompas



No	Target	XPath
1	Headline	/html/body/div[3]/div[2]/div/h1
2	Content	/html/body/div[3]/div[3]/div[1]/ div[3]/div[2]/div[2]/p

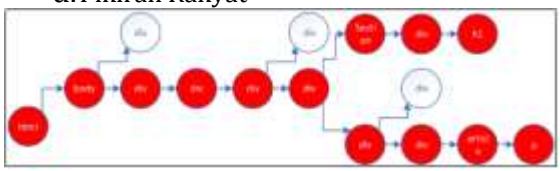
c. Bisnis Indonesia



No	Target	XPath

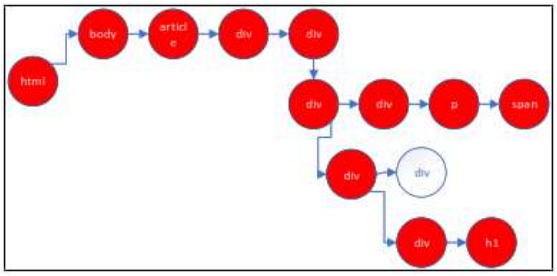
1	Headline	/html/body/div[2]/div/div/div[1]/h1
2	Content	/html/body/div[3]/div[3]/div[1]/div[3]/div[2]/div[2]/p

d. Pikiran Rakyat



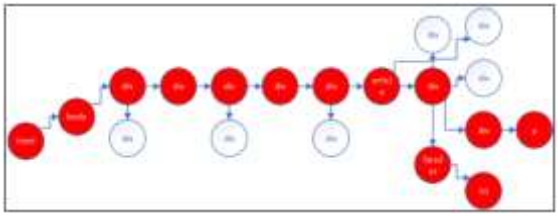
No	Target	XPath
1	Headline	/html/body/div[2]/div/div/div[2]/section/div/h1
2	Content	/html/body/div[2]/div/div/div[2]/div[1]/div[2]/article/p

e. Cek & Ricek



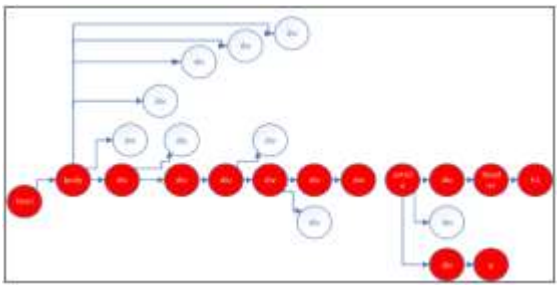
No	Target	XPath
1	Headline	/html/body/article/div/div[1]/div[1]/div[2]/h1
2	Content	/html/body/article/div/div[1]/div[1]/div/p/span/text()

f. Siwalima news



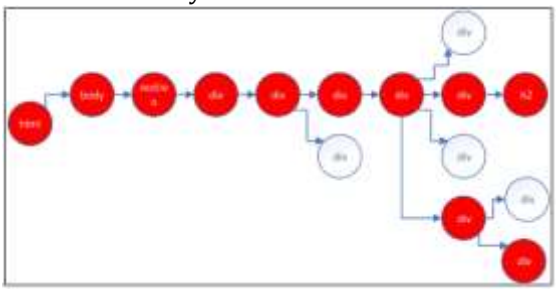
No	Target	XPath
1	Headline	/html/body/div[1]/div/div/div[1]/div[1]/article/div[2]/header/h1
2	Content	/html/body/div[1]/div/div/div[1]/div[1]/article/div[2]/div[3]/p

g. Waspada



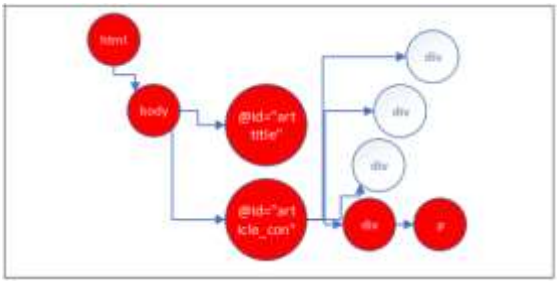
No	Target	XPath
1	Headline	/html/body/div[6]/div[2]/div/div[2]/div[1]/div/article/div[1]/hea der/h1
2	Content	/html/body/div[6]/div[2]/div/div[2]/div[1]/div/article/div[3]/p

h. Analisaday



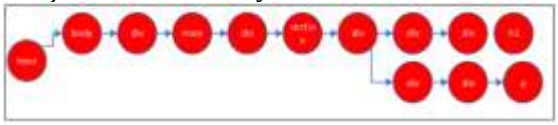
No	Target	XPath
1	Headline	/html/body/div[1]/main/div/section/div/div[1]/div[1]/h1
2	Content	/html/body/div[1]/main/div/section/div/div[1]/div[2]/div/p

i. Makasar Trimbun



No	Target	XPath
1	Headline	/html/body/div[1]/main/div/section/div/div[1]/div[1]/h1
2	Content	/html/body/div[1]/main/div/section/div/div[1]/div[2]/div/p

j. Kedaulatan Rakyat



No	Target	XPath
1	Headline	/html/body/div[1]/main/div/section/div/div[1]/div[1]/h1
2	Content	/html/body/div[1]/main/div/section/div/div[1]/div[2]/div/p

This target data page contains a list of site links that will be scrapped or retrieved from the news content through the XPath selector (training data) that has been provided previously. The admin clicks the button, the process will perform the html feature extraction and analyze the XPath matched with the XPath selector or the previous XPath model that has been provided if the same then the headline content and the news content will be saved if the XPath selector model is not the same as the news site that will be scraping the food datacan retrieve the news content and the system will display an error such as Figure.

Figure 6. *Target Site*

This user page displays all selected news contents by pressing the read-further button and news search pages based on news categories as shown in the Figure 7.

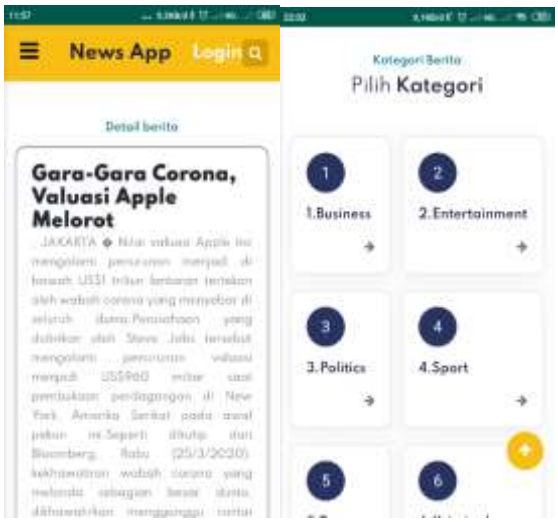


Figure 6. *read the news and choose a category*

CONCLUSION

Based on the results and previous discussion, the results of this study are in the form of DOM tree and XPath pattern analysis of the news sites studied, and implemented in the form of a web-based scrap news web application using the python programming language with the flask framework. The results obtained are knowing the pattern of DOM and XPath news sites. The XPath pattern always differs in a news site among others. The results of the extraction data using XPath were more complete. By using this supervised learning method, the more data on the XPath training model the better this application will be.

So when conducting web scraping, the site is no longer analyzed because it's already in the XPath model data.

REFERENCES

Buduma, N., & Locascio, N. (2017). Fundamentals of deep learning: Designing next-generation machine intelligence algorithms. / Nikhil Buduma ; with contributions by Nicholas Locascio. In Designing next-generation machine intelligence algorithms.

Chollet, F. (2018). Deep Learning with Phyton. In Manning.

GCM, G., Laterra, P., Aparicio, V., & Costa, J. (2016). Glyphosate retention in grassland riparian areas is reduced by the invasion of exotic trees. Phyton. https://doi.org/10.32604/phyton.2016.85.108

Grandvalet, Y., & Bengio, Y. (2005). Semi-supervised learning by entropy minimization. Advances in Neural Information Processing Systems.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems.

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2020.02.005.

Iacus, S. M. (2015). Automated Data Collection with R - A Practical Guide to Web Scraping and Text Mining . Journal of Statistical Software. https://doi.org/10.18637/jss.v068.b03.

Indriyanti, A. D., Prehanto, D. R., Prismana, I. G. L. E. P., Soeryanto, Sujatmiko, B., & Fikandda, J. (2019). Simple Additive Weighting algorithm to aid administrator decision making of the underprivileged scholarship. Journal of Physics: Conference Series. https://doi.org/10.1088/1742-6596/1402/6/066070.

Indriyanti, Aries Dwi, Prehanto, D. R., Permadi, G. S., Mashuri, C., & Vitadiar, T. Z. (2019). Using Fuzzy Time Series (FTS) and Linear Programming for Production Planning and Planting Pattern Scheduling Red Onion. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/201912523007

.Khalil, S., & Fakir, M. (2017). RCrawler: An R package for parallel web crawling and scraping. SoftwareX. https://doi.org/10.1016/j.softx.2017.04.004.

Kistofer, T., Permadi, G. S., & Vitadiar, T. Z. (2019). Development of Digital System Learning Media Using Digital Learning System. https://doi.org/10.2991/assehr.k.191217.030.

Laine, S., & Aila, T. (2017). Temporal ensembling for semi-supervised learning. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings.

Mashuri, C., Mujianto, A. H., Sucipto, H., Arsam, R. Y., & Permadi, G. S. (2019). Production Time Optimization using Campbell Dudek Smith (CDS) Algorithm for Production Scheduling. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/201912523009.

Mitchell, R. (2015). Web Scraping with Python Collecting Data From The Modern Web. In O'Reilly.

Mutasim, A. K., Tipu, R. S., Bashar, M. R., Islam, M. K., & Amin, M. A. (2018). Computational intelligence for pattern recognition in EEG signals. In Studies in Computational Intelligence. https://doi.org/10.1007/978-3-319-89629-8_11.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning. https://doi.org/10.1145/1102351.1102430.

Nylen, E. L., & Wallisch, P. (2017). Web Scraping. In Neural Data Science. https://doi.org/10.1016/b978-0-12-804043-0.00010-6.

Permadi, G. S., Adi, K., & Gernowo, R. (2018). Application Mail Tracking Using RSA Algorithm As Security Data and HOT-Fit a Model for Evaluation System. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/20183111007.

Permadi, G. S., Vitadiar, T. Z., Kistofer, T., & Mujianto, A. H. (2019). The Decision Making Trial and Evaluation Laboratory (Dematel) and Analytic Network Process (ANP) for Learning Material Evaluation System. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/201912523011.

PF, C., Scarpassa, J., Pretto-Giordano, L., Otaguiri, E., Yamada-Ogatta, S., Nakazato, G., Perugini, M., Moreira, I., & Vilas-B鬪s, G. (2016). Antibacterial activity of avocado extracts (Persea americana Mill.) against Streptococcus agalactiae. Phyton. https://doi.org/10.32604/phyton.2016.85.218.

Prehanto, D. R., Indriyanti, A. D., Nuryana, K. D., Soeryanto, S., & Mubarok, A. S. (2019). Use of Naïve Bayes classifier algorithm to detect customers' interests in buying internet token. Journal of Physics: Conference Series. https://doi.org/10.1088/1742-6596/1402/6/066069.

Prehanto, Dedy Rahman, Indriyanti, A. D., Mashuri, C., & Permadi, G. S. (2019). Soil Moisture Prediction using Fuzzy Time Series and Moisture sensor Technology on Shallot Farming. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/201912523002.

S.C.M. de S Sirisuriya. (2015). A Comparative Study on Web Scraping. 8th International Research Conference KDU.
Ulbricht, L. (2020). Scraping the demos. Digitalization, web scraping and the democratic project. Democratization. https://doi.org/10.1080/13510347.2020.1714595.
Xia, Y., Qin, T., Chen, W., Bian, J., Yu, N., & Liu, T. Y. (2017). Dual supervised learning. 34th International Conference on Machine Learning, ICML 2017.
You, Z., Raich, R., Fern, X. Z., & Kim, J. (2018). Weakly Supervised Dictionary Learning. IEEE Transactions on Signal Processing. https://doi.org/10.1109/TSP.2018.2807422.