Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar

A Rahmatulloh[*1], R Gunawan[2]

[1,2]Department of Informatics, Faculty of Engineering, Universitas Siliwangi

E-mail: alam@unsil.ac.id[1], rohmatgunawan@unsil.ac.id[2]

Submitted: 30 December 2019, revised: 3 February 2020, accepted: 13 February 2020

Abstrak. Google Scholar adalah layanan berbasis web untuk mencari literatur akademik. Berbagai jenis referensi yang dapat diakses contohnya adalah: makalah *peer-review*, tesis, buku, abstrak dan artikel dari penerbit akademik, komunitas profesional, pusat data pra-cetak, universitas dan organisasi akademik lainnya. Google Scholar menyediakan fitur pembuatan profil untuk setiap peneliti, pakar, dan dosen. Jumlah publikasi dari lembaga akademis bersama dengan data yang detail tentang publikasi artikel ilmiah dapat diakses melalui Google Scholar. Rekapitulasi publikasi artikel ilmiah dari masing-masing peneliti di suatu lembaga diperlukan untuk menentukan kinerja penelitian secara kolektif. Namun, hal ini masih menyisakan masalah, yaitu belum tersedianya layanan rekapitulasi publikasi artikel ilmiah untuk setiap peneliti di suatu lembaga. Oleh karena itu, penelitian ini berupaya melakukan rekapitulasi publikasi artikel ilmiah. Pengumpulan data dari Google Scholar dilakukan dengan menerapkan teknologi *web scrapping*. Eksperimen *web scrapping* dari Google Scholar dalam penelitian ini telah berhasil mengambil 238 data peneliti dan 2.523 file artikel. Data yang telah diunduh disimpan ke dalam basis data, kemudian digunakan untuk rekapitulasi publikasi artikel ilmiah, yang dapat menampilkan: daftar profil peneliti, daftar afiliasi, daftar kutipan, dan daftar judul artikel yang dapat dicetak dalam bentuk *.pdf atau *.xlsx dan dilengkapi dengan pencarian data dan fitur penyortiran.

Kata kunci: koleksi data; google scholar; *web scraping*

Abstract. Google Scholar is a web-based service for searching broad academic literature. Various types of references can be accessed such as: peer-reviewed papers, theses, books, abstracts and articles from academic publishers, professional communities, pre-printed data center, universities and other academic organizations. Google Scholar provides the profile creation feature of every researcher, expert, and lecturer. The quantity of publication from an academic institution along with detailed data on the publication of scientific articles can be accessed through Google Scholar. Recapitulation of the publication of scientific articles of each researcher in an institution is needed to determine the research performance collectively. However, it still leaves a problem, that is the unavailability of recapitulation services publication of scientific articles for each researcher in an institution. Therefore, this study attempts to make the recapitulation of scientific article publications. Data collection from Google Scholar was carried out by applying web scraping technology. The scraping experiment from Google Scholar in this study has succeeded in retrieving 238 researchers' data and 2,523 article files. The data that had been downloaded was stored in a database, then used to recapitulate the publication of scientific articles, which can display: a list of researcher profiles, a list of affiliations, a list of citation and a list of article titles that can be printed in the form of *. Pdf or * .xlsx and is equipped with features data search and sorting.

Keywords: data collection; google scholar; web scraping

1. Introduction

For scientists or researchers, publishing research results is an obligation. Forms of research publications include: books, Intellectual Property Rights (IPR) and scientific articles. The publications for the academic community, especially universities have a significant impact on the awareness of lecturers on the importance of conducting studies, research and writing scientific papers [1]. Google Scholar is a web-based service from Google Incorporation to search broad academic literature. One can search in all fields of science and references, such as: peer-reviewed papers, theses, books, abstracts, and articles from academic publishers, professional communities, pre-printed data centers, universities, and other academic organizations. Google Scholar is designed to arrange articles written by researchers into an account that can display the frequency of article citations and later increase the citation of the articles in other academic literature [2].

Google Scholar provides a profile creation feature for every researcher, expert or lecturer. The quantity of publications from an academic institution can be accessed through Google Scholar. Researchers' profiles and scientific article publication data can also be accessed through Google Scholar. Every scientific article that has been published in an online journal, only requires a short amount of time to be indexed by Google Scholar. A recap of the publication of scientific articles of each researcher in an institution or organization is needed to determine the research performance collectively. However, a problems then emerges, that is the unavailability of services to recap the publication of scientific articles for each researcher in an institution or organization. It requires time and efforts to obtain collective data or recapitulation of the publication of all researchers or lecturers from an institution or college. As the result, the publication data of scientific articles can be utilized by academic institutions or organizations. This research obtained data from Google Scholar to recapitulate scientific article publication data by applying web scraping technology. Web scraping is a technology that allows the taking of resources from the web and the results can be utilized again by other systems. The process of retrieving data or information from sites on the internet is called web scraping [3], [4], [5]; web extraction [6], [7], [8]; web mining [9], [10]; and web harvesting [11], [12].

Several studies related to the implementation of web scraping of scientific article or literature from the internet have been carried out beforehand including: web scraping for Indonesian - English parallel corpus using HTML DOM method [4], web-scraping software in searching for gray literature [5], application of web scraping techniques in scientific article search engines [13], the application of web scraping and winnowing web for the detection of plagiarism in the final project title [14], [15]. There are several algorithms that can be used in web scraping such as: regular expressions, HTML DOM, and Xpath [16]. Each algorithm has its own characteristics, so it needs a good understanding before applying it. The regular expression algorithm requires less memory compared to the HTML DOM, and Xpath methods, and HTML DOM takes the least amount of time and uses the least data compared to regular expressions and Xpath [15]. In this study, web scraping using the HTML DOM method was used to download scientific article publication data from Google Scholar based on the Id or class contained in the Google Scholar web source code. The data that were successfully downloaded was stored in a database, then used to recapitulate the publication of scientific articles, which are designed to display: a list of researcher profiles, a list of affiliations, a list of citation and a list of article titles that can be printed in the form of * .pdf or * .xlsx and completed with data search and sorting feature.

2. Literature Review

The parallel corpus is two interconnected text documents. The first text document contains a collection of source sentences, while the second document contains a collection of translated sentences. The parallel corpus serves as the main source in developing statistical translation machines. Collecting parallel corpus manually requires a long time and cost. The research conducted by [4], tried to implement web scraping with HTML DOM method to collect parallel corpus in Indonesian and English. Experiments on his research have been able to produce 38,712 pairs of parallel corpus from the bilingual news website http://www.berita2bahasa.com/ as well as Indonesian news collection documents as a source and English as the translation. The research conducted by [5], suggests a variety of tools that can be used to search for references and scraping gray literature. There are about 15 platforms that can be used for scraping data are presented and are equipped with descriptions, prices, and URLs to access them. The results of his research have provided information about the availability of a variety of free and low-cost web scraping software and provide opportunities for those who have limited resources, especially researchers who work alone or work in small organizations.

The research conducted by [13], tried to apply web scraping technology to retrieve data from several scientific article search engines. Three scientific article search engine webs were chosen for his research, including: Digital Referral Garama (Garuda) http://garuda.ristekdikti.go.id/, Indonesian Scientific Journal Database (ISJD) http: //isjd.pdii.lipi. go.id/ and Google Scholar https://scholar.google.com/. His research has succeeded in applying web scraping techniques and downloading some data from selected scientific article search engines. The downloaded data then stored in a database table consisting of 6 attributes: id, website, keywords, results, file_download, and date_time_update. The research conducted by [14], applied winnowing algorithm to find the level of similarity in the publication of scientific article titles. Google Scholar was used to obtain research title data that had been previously available as a comparison with the research title entered. Web scraping with CURL (URL Client) and Hypertext Markup Language-Document Object Model (HTML DOM) parser were used to retrieve the title data from Google Scholar. Experiments in his research, have succeeded in presenting a percentage level of similarity in percent with the category of low, middle or high plagiarism.

The creation of a recapitulation service for publishing scientific articles collectively for an institution is the main focus of this research. Therefore, the recapitulation of scientific article publications can be done easily. In this study, the data were scraped using the HTML DOM method based on the Id or class contained in the Google Scholar web source code. Data scraping was done based on each researcher's Google Scholar Id. The downloaded data were then stored in a database to recapitulate the publication of scientific articles.

3. Methodology

3.1. Data Collection

There are three main steps involved in the process of retrieving data from the Web Scholar. Activities undertaken at this stage consist of: mapping google scholar web pages, developing web scraping source code, save the scraped data to the database, and reporting as shown in Figure 1.

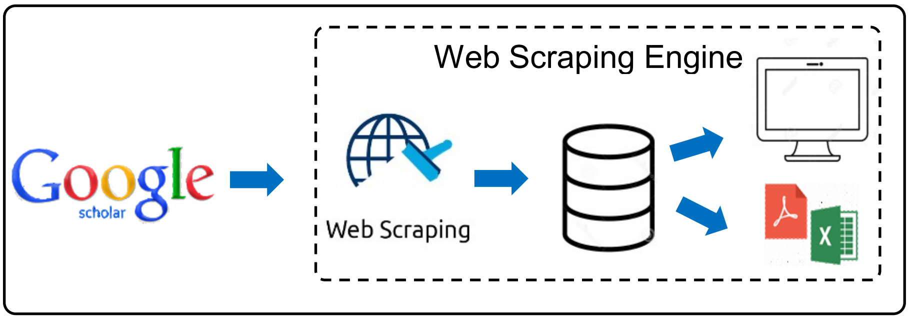
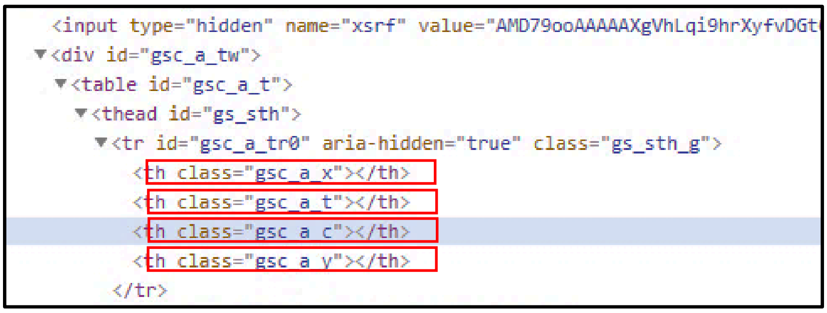


Figure 1. System Architecture Web Scraping Google Scholar

3.1.1. Mapping Google Scholar web pages

It was done by displaying the source code of web pages through a web browser and identifying each id or class on the web page element. The identification results of the id or class were chosen according to the data attributes to be scrapped. Figure 2 shows an example of some of id or class identified on the target website.



```
   <input type="hidden" name="xsrf" value="AMD79ooAAAAAXgVhLqi9hrXyfvDGt
 ▼<div id="gsc_a_tw">
   ▼<table id="gsc_a_t">
     ▼<thead id="gs_sth">
       ▼<tr id="gsc_a_tr0" aria-hidden="true" class="gs_sth_g">
           <th class="gsc_a_x"></th>
           <th class="gsc_a_t"></th>
           <th class="gsc_a_c"></th>
           <th class="gsc_a_y"></th>
         </tr>
```

Figure 2. The source code of the Google Scholar

web page marked on the class element

In Figure 2, some examples of classes are displayed in the source code of Google Scholar web pages, such as: gsc_a_x, gsc_a_t, gsc_a_c, gsc_a_y. The id or class was then identified and adjusted to the attributes of the data to be downloaded.

3.1.2. Developing web scraping source code

It was done by using PHP version 5, Apache Web Server, and MySQL Database. Some functions inserted in the PHP source code were designed to pass data scraping based on the id or class on the Google Scholar web page. The source code snippet for Google scholar scraping is shown in Figure 3.

```
<?php
include 'simple_html_dom.php';
class request_paper {
      public $id;
      function __construct($id)
      public function url()
      public function url_request($url)
      public function foto()
      public function info()
      public function citation()
      public function paper()
      public function year()
      public function graf()
     }
?>
```

Figure 3. Pseudocode for Scraping Data Google Scholar

Figure 3 shows a pseudocode designed to scrap data from Google Scholar. There is one class named request_paper, which has 10 methods.

3.1.3. Save the scraping data to the database

It was done after the scraping process is complete. MySQL Server was used in this study as a tool to store data, which was connected with PHP based applications. Several tables were designed to store data from scarping results, including: Researcher profile table, title table, citation table, affiliation table and others.

3.1.4. Reporting

It was done by accessing data that had been stored in a database. The data was then displayed to recapitulate the publication of scientific articles, which are designed to display: a list of researcher profiles, a list of affiliates, a list of citations, and a list of article titles that can be printed in the form of * .pdf or * .xlsx.

3.2. Data Requirement

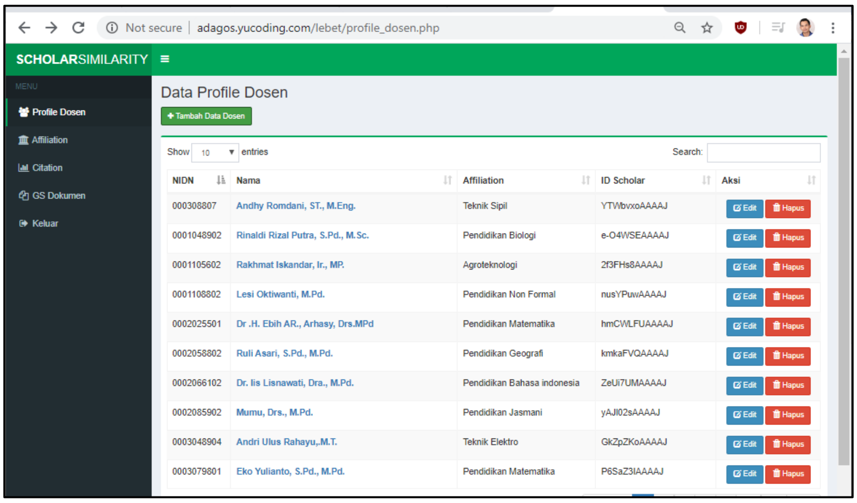
There are various data attributes available on the Web Scholar. Some data attributes needed to produce a recapitulation of scientific article publications in this study are shown in Table 1.

Table 1. Google Scholar Data Object

No	ID	Name	Description
1	#gsc_a_b	Article Table	Article Publication
2	#a	Article Title	Article Title
3	#gsc_a_c	Citation	Article Citation
4	#gsc_a_y	Year	Article Year
5	#gs_gray	Author/Publisher	Article Writer
6	#gsc_prf_pu .gs_rimg	Image	Author Photo
7	#gsc_prf_inta	Information	Information
8	#gsc_rsb_st	Citation_recap_	Citation
9	#gsc_md_hist_b	Graph	Graph

4. Results and Analysis

As previously designed, in this stage the web scrap can be implemented in the web-based application that can be accessed via the URL http://adagos.yucoding.com. There are two levels of access to the application developed, namely: admin and public. The main page for users with admin privileges after successfully logging in is shown in Figure 4.



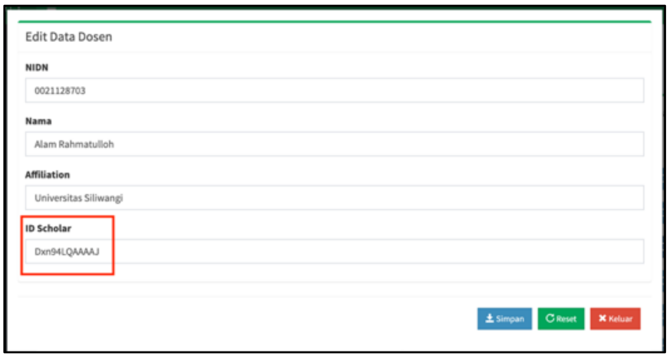
Data Profile Dosen

Show 10 entries Search:

NIDN	Nama	Affiliation	ID Scholar	Aksi
000308807	Andhy Romdani, ST., M.Eng.	Teknik Sipil	YTWbvxoAAAAJ	Edit Hapus
0001048902	Rinaldi Rizal Putra, S.Pd., M.Sc.	Pendidikan Biologi	e-O4WSEAAAAJ	Edit Hapus
0001105602	Rakhmat Iskandar, Ir., MP.	Agroteknologi	2f3FHs8AAAAJ	Edit Hapus
0001108802	Lesi Oktiwanti, M.Pd.	Pendidikan Non Formal	nusYPuwAAAAJ	Edit Hapus
0002025501	Dr .H. Ebih AR., Arhasy, Drs.MPd	Pendidikan Matematika	hmCWLFUAAAAJ	Edit Hapus
0002058802	Ruli Asari, S.Pd., M.Pd.	Pendidikan Geografi	kmkaFVQAAAAJ	Edit Hapus
0002066102	Dr. Iis Lisnawati, Dra., M.Pd.	Pendidikan Bahasa indonesia	ZeUi7UMAAAAJ	Edit Hapus
0002085902	Mumu, Drs., M.Pd.	Pendidikan Jasmani	yAJI02sAAAAJ	Edit Hapus
0003048904	Andri Ulus Rahayu,.M.T.	Teknik Elektro	GkZpZKoAAAAJ	Edit Hapus
0003079801	Eko Yulianto, S.Pd., M.Pd.	Pendidikan Matematika	P6SaZ3IAAAAJ	Edit Hapus

Figure 4. The user's main page with admin access level

In Figure 4, the section displays a menu that can be accessed by the admin and the right part is a record of each item that has been input. For example, "Lecturer Profile" menu on the left is selected, then the right list of lecture profiles that have been successfully input and stored in a database are shown on the right. Users with admin access rights in addition to accessing data can also manipulate data. In this study, lecturer data as a researcher in a university was chosen as a sample data for the experiment. The lecturer data with attributes of NIDN, Name, Affiliation, and Google Scholar ID was input at an early stage before scraping. The lecturer data input form display is shown in Figure 5.



Edit Data Dosen

NIDN
0021128703

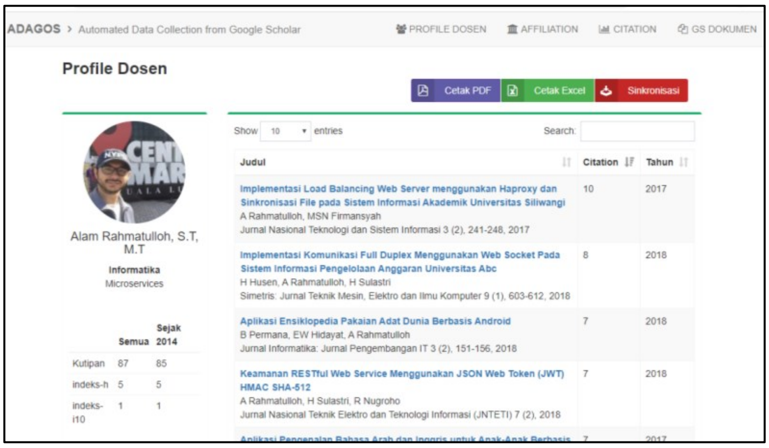
Nama
Alam Rahmatulloh

Affiliation
Universitas Siliwangi

ID Scholar
Dxn94LQAAAAJ

Figure 5. Lecturer data input form

Figure 5 shows the lecturer data input page. Each lecturer whose data will be input into the system must have a Google scholar ID. If the lecturers have not got a Google Scholar ID, they are required to create a research profile, especially through https://scholar.google.com/. Experiments in this study have succeeded in inputting data from 238 lecturers who are members of 10 affiliates.



Alam Rahmatulloh, S.T, M.T
Informatika
Microservices

	Semua	Sejak 2014
Kutipan	87	85
indeks-h	5	5
indeks-i10	1	1

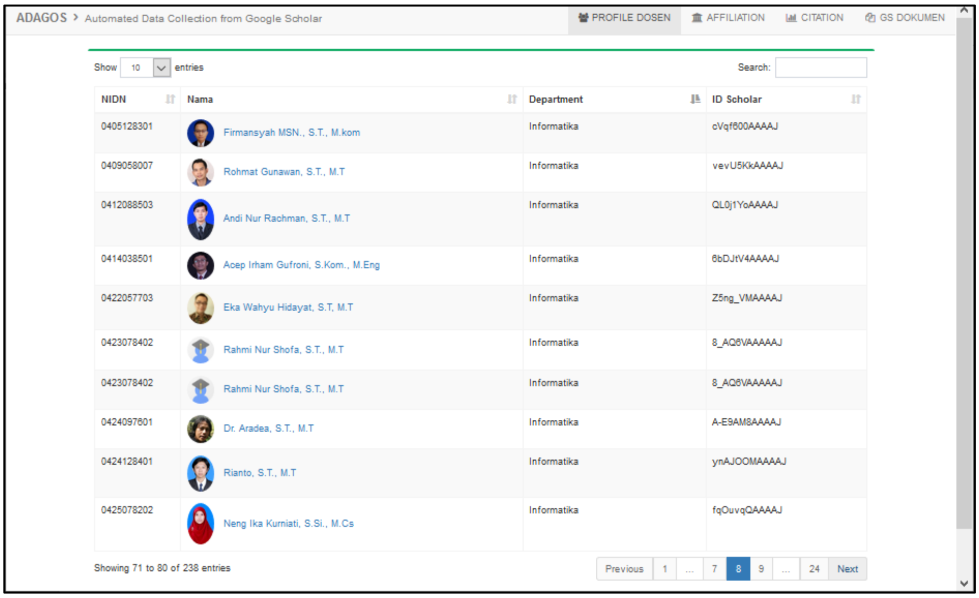
Show 10 entries Search:

Judul	Citation	Tahun
Implementasi Load Balancing Web Server menggunakan Haproxy dan Sinkronisasi File pada Sistem Informasi Akademik Universitas Siliwangi A Rahmatulloh, MSN Firmansyah Jurnal Nasional Teknologi dan Sistem Informasi 3 (2), 241-248, 2017	10	2017
Implementasi Komunikasi Full Duplex Menggunakan Web Socket Pada Sistem Informasi Pengelolaan Anggaran Universitas Abc H Husen, A Rahmatulloh, H Sulastri Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer 9 (1), 603-612, 2018	8	2018
Aplikasi Ensiklopedia Pakaian Adat Dunia Berbasis Android B Permana, EW Hidayat, A Rahmatulloh Jurnal Informatika: Jurnal Pengembangan IT 3 (2), 151-156, 2018	7	2018
Keamanan RESTful Web Service Menggunakan JSON Web Token (JWT) HMAC SHA-512 A Rahmatulloh, H Sulastri, R Nugroho Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI) 7 (2), 2018	7	2018
Aplikasi Pengenalan Bahasa Arab dan Inggris untuk Anak-Anak Berbasis	7	2017

Figure 6. Display of researchers web scraping results from google scholar

Data scraping can be done after the lecturer profile is added and stored in a database. The scraping process begins with selecting one of the lecturer profiles, then selecting the "Syncronization" menu as shown in Figure 6. Scraping is done based on the Google Scholar ID and id class selected according to the required data attributes. After the scraping process is finished, it will display the article data that has been successfully downloaded as shown in Figure 6.

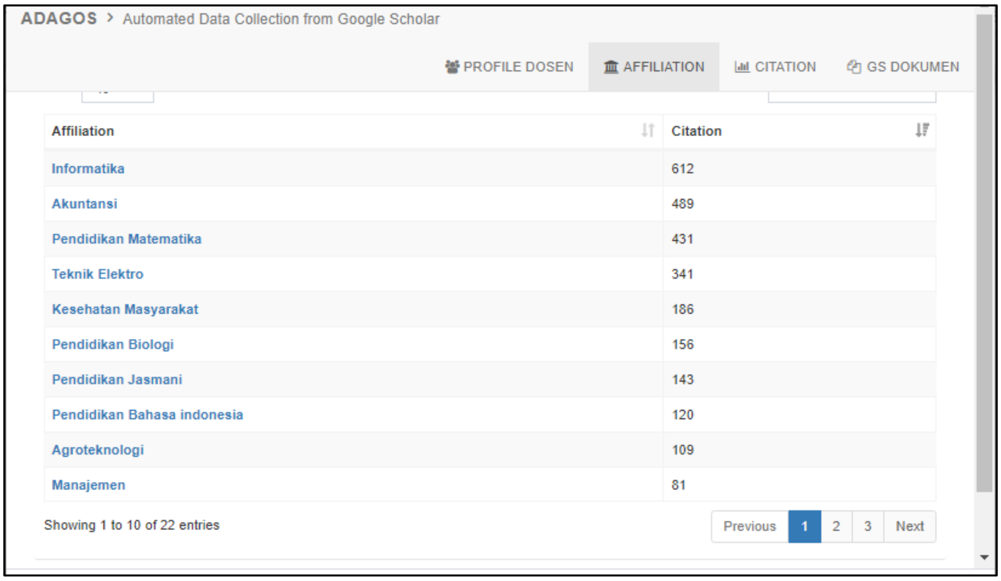
The display in Figure 6 is similar to the profile display on Google Scholar, because all data used is the result of the scraping process from Google Scholar. The recapitulation process is automatically created after the scraping is done. Lecturer list display can be sorted by NIDN, Name or department as shown in Figure 7.



NIDN	Nama	Department	ID Scholar
0405128301	Firmansyah MSN., S.T., M.kom	Informatika	cVqf600AAAAJ
0409058007	Rohmat Gunawan, S.T., M.T	Informatika	vevU5KkAAAAJ
0412088503	Andi Nur Rachman, S.T., M.T	Informatika	QL0j1YoAAAAJ
0414038501	Acep Irham Gufroni, S.Kom., M.Eng	Informatika	6bDJtV4AAAAJ
0422057703	Eka Wahyu Hidayat, S.T, M.T	Informatika	Z5ng_VMAAAAJ
0423078402	Rahmi Nur Shofa, S.T., M.T	Informatika	8_AQ6VAAAAAJ
0423078402	Rahmi Nur Shofa, S.T., M.T	Informatika	8_AQ6VAAAAAJ
0424097601	Dr. Aradea, S.T., M.T	Informatika	A-E9AM8AAAAJ
0424128401	Rianto, S.T., M.T	Informatika	ynAJOOMAAAAJ
0425078202	Neng Ika Kurniati, S.Si., M.Cs	Informatika	fqOuvqQAAAAJ

Showing 71 to 80 of 238 entries

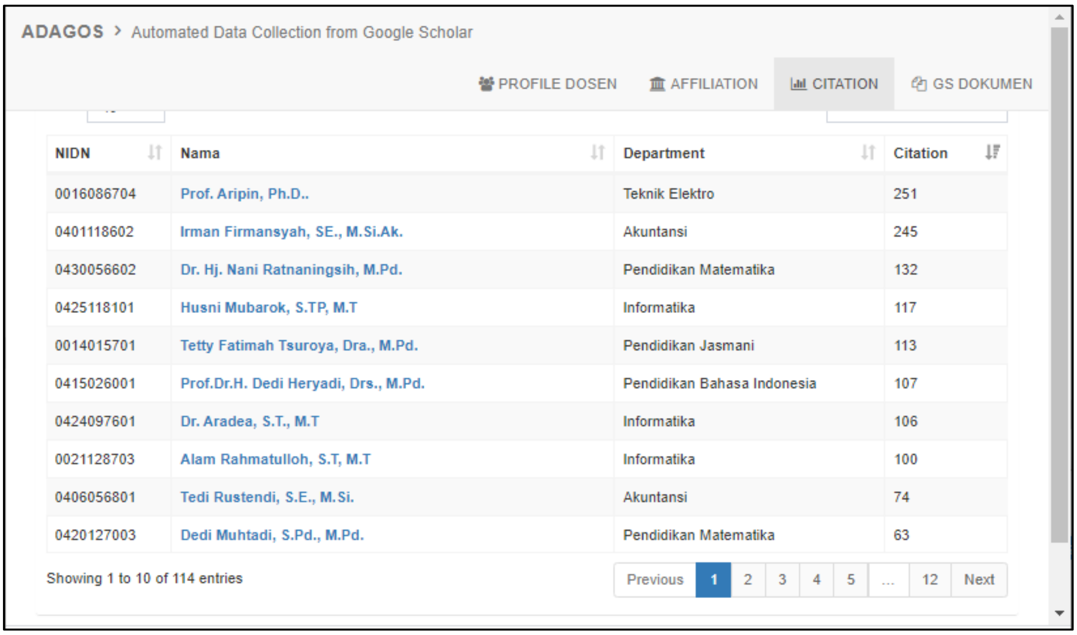
Figure 7. List of researchers



Affiliation	Citation
Informatika	612
Akuntansi	489
Pendidikan Matematika	431
Teknik Elektro	341
Kesehatan Masyarakat	186
Pendidikan Biologi	156
Pendidikan Jasmani	143
Pendidikan Bahasa indonesia	120
Agroteknologi	109
Manajemen	81

Showing 1 to 10 of 22 entries

Figure 8. List of citations recapitulation based on affiliation or department

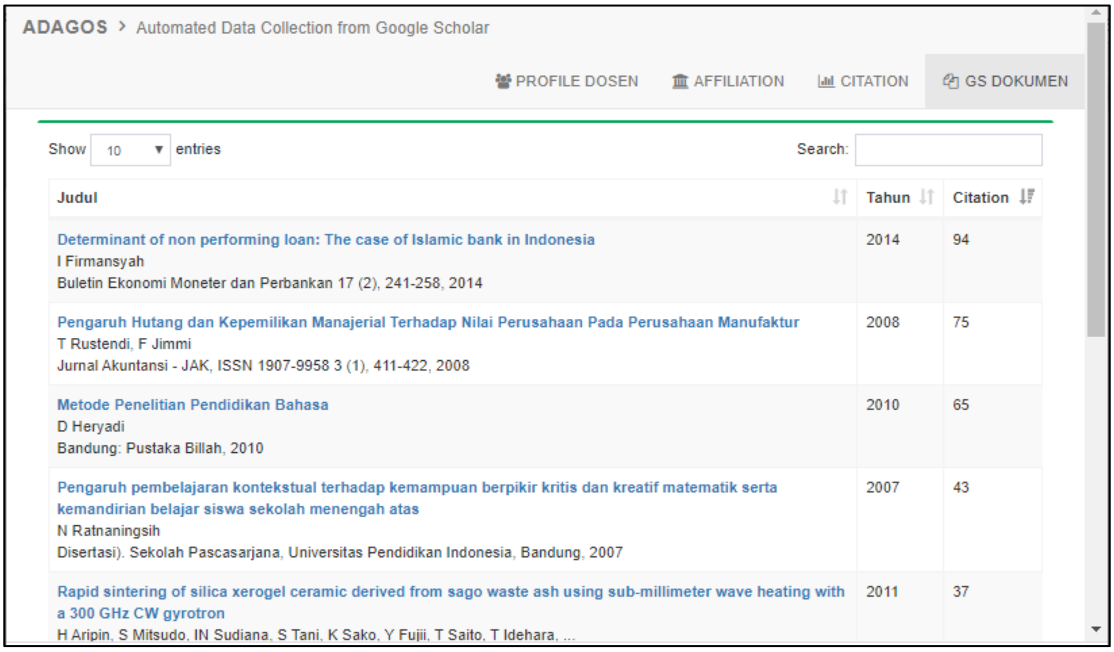


NIDN	Nama	Department	Citation
0016086704	Prof. Aripin, Ph.D..	Teknik Elektro	251
0401118602	Irman Firmansyah, SE., M.Si.Ak.	Akuntansi	245
0430056602	Dr. Hj. Nani Ratnaningsih, M.Pd.	Pendidikan Matematika	132
0425118101	Husni Mubarok, S.TP, M.T	Informatika	117
0014015701	Tetty Fatimah Tsuroya, Dra., M.Pd.	Pendidikan Jasmani	113
0415026001	Prof.Dr.H. Dedi Heryadi, Drs., M.Pd.	Pendidikan Bahasa Indonesia	107
0424097601	Dr. Aradea, S.T., M.T	Informatika	106
0021128703	Alam Rahmatulloh, S.T, M.T	Informatika	100
0406056801	Tedi Rustendi, S.E., M.Si.	Akuntansi	74
0420127003	Dedi Muhtadi, S.Pd., M.Pd.	Pendidikan Matematika	63

Showing 1 to 10 of 114 entries

Figure 9. List of lecturer profile recapitulations based on the most cited articles

The difference of the web scrapping from Google Scholar can be seen in Figure 8-10. The scraped data can be utilized and processed in other forms, for example automatic data collection, recapitulation of the number of studies, the number of citations, and so forth.



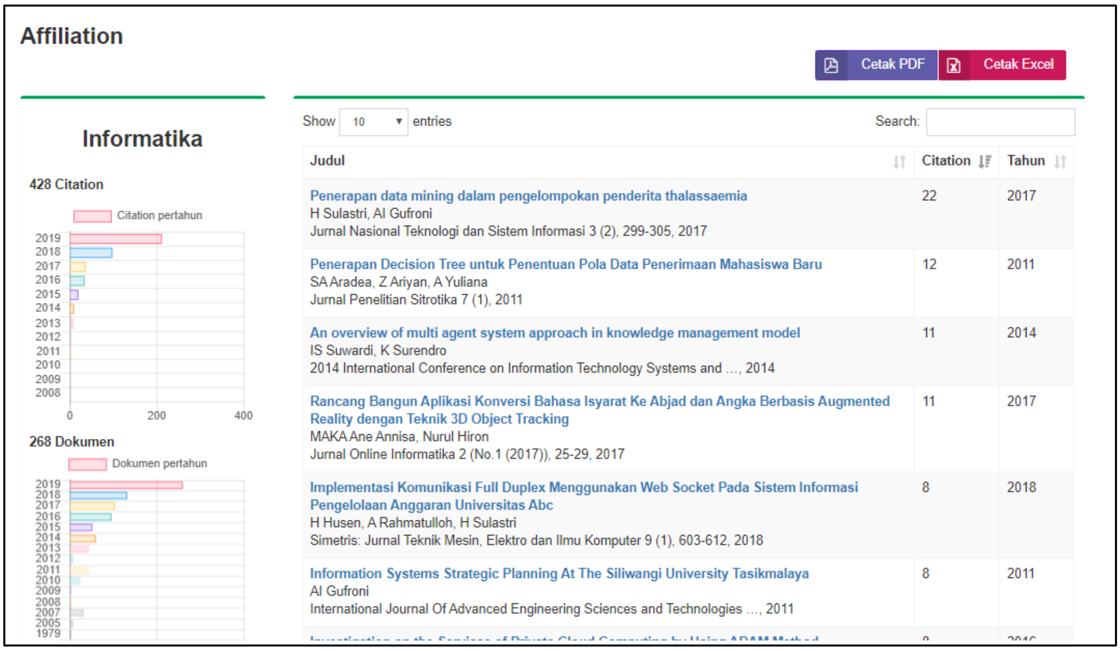
Show 10 entries

Search:

Judul	Tahun	Citation
Determinant of non performing loan: The case of Islamic bank in Indonesia I Firmansyah Buletin Ekonomi Moneter dan Perbankan 17 (2), 241-258, 2014	2014	94
Pengaruh Hutang dan Kepemilikan Manajerial Terhadap Nilai Perusahaan Pada Perusahaan Manufaktur T Rustendi, F Jimmi Jurnal Akuntansi - JAK, ISSN 1907-9958 3 (1), 411-422, 2008	2008	75
Metode Penelitian Pendidikan Bahasa D Heryadi Bandung: Pustaka Billah, 2010	2010	65
Pengaruh pembelajaran kontekstual terhadap kemampuan berpikir kritis dan kreatif matematik serta kemandirian belajar siswa sekolah menengah atas N Ratnaningsih Disertasi). Sekolah Pascasarjana, Universitas Pendidikan Indonesia, Bandung, 2007	2007	43
Rapid sintering of silica xerogel ceramic derived from sago waste ash using sub-millimeter wave heating with a 300 GHz CW gyrotron H Aripin, S Mitsudo, IN Sudiana, S Tani, K Sako, Y Fujii, T Saito, T Idehara, ...	2011	37

Figure 10. List of article data recapitulation based on the number of articles cited

Each scientific article that is in an affiliation can be displayed in order, for example based on the number of citations as shown in Figure 11.



Affiliation

Informatika

428 Citation

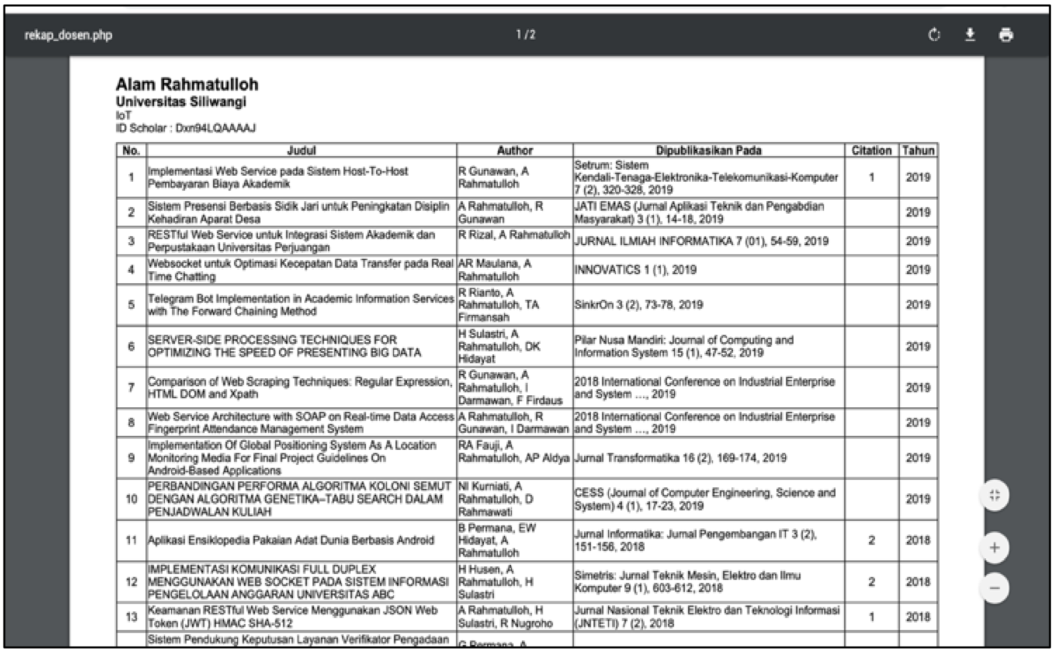
268 Dokumen

Show 10 entries   Search:

Judul	Citation	Tahun
Penerapan data mining dalam pengelompokan penderita thalassaemia H Sulastri, AI Gufroni Jurnal Nasional Teknologi dan Sistem Informasi 3 (2), 299-305, 2017	22	2017
Penerapan Decision Tree untuk Penentuan Pola Data Penerimaan Mahasiswa Baru SA Aradea, Z Ariyan, A Yuliana Jurnal Penelitian Sitrotika 7 (1), 2011	12	2011
An overview of multi agent system approach in knowledge management model IS Suwardi, K Surendro 2014 International Conference on Information Technology Systems and ..., 2014	11	2014
Rancang Bangun Aplikasi Konversi Bahasa Isyarat Ke Abjad dan Angka Berbasis Augmented Reality dengan Teknik 3D Object Tracking MAKA Ane Annisa, Nurul Hiron Jurnal Online Informatika 2 (No.1 (2017)), 25-29, 2017	11	2017
Implementasi Komunikasi Full Duplex Menggunakan Web Socket Pada Sistem Informasi Pengelolaan Anggaran Universitas Abc H Husen, A Rahmatulloh, H Sulastri Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer 9 (1), 603-612, 2018	8	2018
Information Systems Strategic Planning At The Siliwangi University Tasikmalaya AI Gufroni International Journal Of Advanced Engineering Sciences and Technologies ..., 2011	8	2011
Investigation on the Services of Private Cloud Computing by Using ADAM Method	8	2016

Figure 11. Display the title of the article sorted by the number of citations in an affiliate

Examples of displaying scraping scientific report data reports that are poured into *.pdf format can be seen in Figure 12. Besides being in PDF format, reports can also be downloaded in the form of Excel.



Alam Rahmatulloh
Universitas Siliwangi
IoT
ID Scholar : Dxn94LQAAAAJ

No.	Judul	Author	Dipublikasikan Pada	Citation	Tahun
1	Implementasi Web Service pada Sistem Host-To-Host Pembayaran Biaya Akademik	R Gunawan, A Rahmatulloh	Setrum: Sistem Kendali-Tenaga-Elektronika-Telekomunikasi-Komputer 7 (2), 320-328, 2019	1	2019
2	Sistem Presensi Berbasis Sidik Jari untuk Peningkatan Disiplin Kehadiran Aparat Desa	A Rahmatulloh, R Gunawan	JATI EMAS (Jurnal Aplikasi Teknik dan Pengabdian Masyarakat) 3 (1), 14-18, 2019		2019
3	RESTful Web Service untuk Integrasi Sistem Akademik dan Perpustakaan Universitas Perjuangan	R Rizal, A Rahmatulloh	JURNAL ILMIAH INFORMATIKA 7 (01), 54-59, 2019		2019
4	Websocket untuk Optimasi Kecepatan Data Transfer pada Real Time Chatting	AR Maulana, A Rahmatulloh	INNOVATICS 1 (1), 2019		2019
5	Telegram Bot Implementation in Academic Information Services with The Forward Chaining Method	R Rianto, A Rahmatulloh, TA Firmansah	SinkrOn 3 (2), 73-78, 2019		2019
6	SERVER-SIDE PROCESSING TECHNIQUES FOR OPTIMIZING THE SPEED OF PRESENTING BIG DATA	H Sulastri, A Rahmatulloh, DK Hidayat	Pilar Nusa Mandiri: Journal of Computing and Information System 15 (1), 47-52, 2019		2019
7	Comparison of Web Scraping Techniques: Regular Expression, HTML DOM and Xpath	R Gunawan, A Rahmatulloh, I Darmawan, F Firdaus	2018 International Conference on Industrial Enterprise and System ..., 2019		2019
8	Web Service Architecture with SOAP on Real-time Data Access Fingerprint Attendance Management System	A Rahmatulloh, R Gunawan, I Darmawan	2018 International Conference on Industrial Enterprise and System ..., 2019		2019
9	Implementation Of Global Positioning System As A Location Monitoring Media For Final Project Guidelines On Android-Based Applications	RA Fauji, A Rahmatulloh, AP Aldya	Jurnal Transformatika 16 (2), 169-174, 2019		2019
10	PERBANDINGAN PERFORMA ALGORITMA KOLONI SEMUT DENGAN ALGORITMA GENETIKA–TABU SEARCH DALAM PENJADWALAN KULIAH	NI Kurniati, A Rahmatulloh, D Rahmawati	CESS (Journal of Computer Engineering, Science and System) 4 (1), 17-23, 2019		2019
11	Aplikasi Ensiklopedia Pakaian Adat Dunia Berbasis Android	B Permana, EW Hidayat, A Rahmatulloh	Jurnal Informatika: Jurnal Pengembangan IT 3 (2), 151-156, 2018	2	2018
12	IMPLEMENTASI KOMUNIKASI FULL DUPLEX MENGGUNAKAN WEB SOCKET PADA SISTEM INFORMASI PENGELOLAAN ANGGARAN UNIVERSITAS ABC	H Husen, A Rahmatulloh, H Sulastri	Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer 9 (1), 603-612, 2018	2	2018
13	Keamanan RESTful Web Service Menggunakan JSON Web Token (JWT) HMAC SHA-512	A Rahmatulloh, H Sulastri, R Nugroho	Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI) 7 (2), 2018	1	2018
	Sistem Pendukung Keputusan Layanan Verifikator Pengadaan	G Permana, A			

Figure 12. Report on the results of Web Scraping in PDF file format

5. Conclusion

This research has succeeded in taking and collecting data from scientific articles from Google Scholar. Based on the experiments, the Google Scholar web scraping implementation research using the HTML DOM method, has successfully retrieved 238 researcher data from Web Scholar, with 9 attributes and 2,523 articles. The data that has been successfully obtained can be manipulated in the form of visual recap or converted to * .xlsx or * .pdf data format.

The application of automatic scheduler in Google Scholar web scraping is one of interesting topics that can be developed in subsequent research so that the data will be obtained every time change occur in the Web Scholar.

References

[1] Nurhadi, “Pentingnya Publikasi Karya Ilmiah,” 2019. [Online]. Available: https://www.uny.ac.id/?q=berita/pentingnya-publikasi-karya-ilmiah.html

[2] jurkimiaunnes, “Tentang Google Cendikia (Google Scholar).” [Online]. Available: http://kimia.unnes.ac.id/v1/2016/01/01/google-scholar/

[3] W. Scraping, “General techniques used for web scraping Wiki Guide - IGN,” pp. 1–6, 2019.

[4] V. Mitra, H. Sujaini, and A. B. P. Negara, “Rancang Bangun Aplikasi Web Scraping Untuk Korpus Paralel Indonesia - Inggris Dengan Metode HTML Dom,” *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, 2017.

[5] N. R. Haddaway, “The use of web-scraping software in searching for grey literature,” *Grey Journal*, vol. 11, no. February, pp. 186–190, 2016.

[6] N. Ibrahim, A. Hassan, and M. Nihad, “Big data analysis of web data extraction,” *International Journal of Engineering and Technology(UAE)*, vol. 7, no. 4, pp. 168–172, 2018. DOI: 10.14419/ijet.v7i4.37.24095

[7] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, “Knowledge-Based Systems Web data extraction , applications and techniques : A survey,” *Knowledge-Based Systems*, vol. 70, pp. 301–323, 2014 [Online]. DOI: 10.1016/j.knosys.2014.07.007

[8] A. Parameswaran, N. Dalvi, H. GarciaMolina, and R. Rastogi, “Optimal schemes for robust web extraction,” *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 980–991, 2011.

[9] A. Gök, A. Waterworth, and P. Shapira, “Use of web mining in studying innovation,” *Scientometrics*, vol. 102, no. 1, pp. 653–671, 2015. DOI: 10.1007/s11192-014-1434-0

[10] L. K. Joshila Grace, V. Maheswari, and D. Nagamalai, “Analysis of Web Logs And Web User In Web Mining,” *International Journal of Network Security & Its Applications*, vol. 3, no. 1, pp. 99–110, 2011. DOI: 10.5121/ijnsa.2011.3107

[11] E. Şt CHIFU, T. Şt LEŢIA, B. Budişan, and V. R. Chifu, “Web Harvesting and Sentiment Analysis of Consumer Feedback,” *ACTA TECHNICA NAPOCENSIS Electronics and Telecommunications*, vol. 56, no. 3, pp. 7–14, 2015.

[12] P. A. Johnson, R. E. Sieber, N. Magnien, and J. Ariwi, “Automated web harvesting to collect and analyse user-generated content for tourism,” *Current Issues in Tourism*, vol. 15, no. 3, pp. 293–299, 2012. DOI: 10.1080/13683500.2011.555528

[13] A. Josi, L. A. Abdillah, and Suryayusra, “Penerapan teknik web scraping pada mesin pencari artikel ilmiah,” 2014 [Online]. Available: http://arxiv.org/abs/1410.5777

[14] N. I. Kurniati, A. Rahmatulloh, and R. N. Qomar, “Web Scraping and Winnowing Algorithms for Plagiarism Detection of Final Project Titles,” *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 10, no. 2, pp. 73–83, 2019.

[15] A. Rahmatulloh, N. I. Kurniati, I. Darmawan, A. Z. Asyikin, and D. W. Jacob, “Comparison between the Stemmer Porter Effect and Nazief-Adriani on the Performance of Winnowing Algorithms for Measuring Plagiarism,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 9, no. 4, pp. 1124–1128, 2019 [Online]. Available: http://ijaseit.insightsociety.org/index.php?option=com_content&view=article&id=9&Itemid=1&article_id=8844

[16] R. Gunawan, A. Rahmatulloh, I. Darmawan, and F. Firdaus, “Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath,” 2019. DOI: 10.2991/icoiese-18.2019.50