

ABSTRAK

Anderson Montella Karel

Skripsi

Web Scraping dan Information Retrieval dari Google Scholar, Sinta, ScimagoJr Untuk Pencatatan Otomatis Publikasi Dosen UKP

Setiap publikasi yang dikeluarkan oleh dosen perlu dicatat sebagai kinerja perguruan tinggi, tetapi pencatatan ini masih dilakukan secara manual oleh dosen yang mengeluarkan publikasi itu sendiri. Pencatatan manual ini membuat tugas administrasi dosen bertambah. Selain itu data sitasi yang dicatat juga sering berubah, yang membuat pencatatan secara manual tidak memungkinkan.

Teknologi web scraping dan headless web browser dapat digunakan untuk mengotomasi proses pencatatan publikasi, sehingga tugas dosen berkurang. Script web scraping akan dibuat sebagai API yang bisa dipanggil untuk melakukan pencatatan secara otomatis. Data yang didapatkan melalui proses scraping akan disimpan pada database PostgreSQL. Data publikasi yang disimpan dalam database dapat dilihat melalui sebuah aplikasi web sederhana yang dibuat dengan framework Laravel.

Dari pengujian data, sistem ini bisa menemukan dan melakukan pencatatan otomatis pada publikasi dosen yang tercatat pada website sumber, ini ditandai pada kesamaan data abstrak 100%. Tetapi sistem masih mempunyai masalah untuk memberikan kategori yang akurat, ini ditandai dengan kesamaan data yang sangat rendah yaitu 7%. Sistem juga memerlukan id profile dosen pada website sumber untuk bisa memulai proses scraping nya.

Keywords: *web scraping, information retrieval, google scholar, sinta, scimagojr*

ABSTRACT

Anderson Montella Karel

Undergraduate Thesis

Web Scraping and Information Retrieval from Google Scholar, Sinta, ScimagoJr for the Automatic Recording of PCU's Lecturer's Publication

Every publication published by a lecturer needs to be recorded. Right now, the recording process is done manually and done by the lecturer who published the publication. The manual recording adds administrative tasks for the lecturer. The data recorded can also change, like the number of citations a publication has which makes manual recording impossible.

Web scraping and headless web browser technology can be used to automate the publication recording process, reducing lecturer's task. The web scraping script will be made as an API that can be called to do the recording process automatically. The data got from the scraping process will be stored in a PostgreSQL database. The publication data stored in the database can be viewed in a simple web application made with the Laravel framework.

From testing the data, the system can find and record the publication automatically, indicated by the 100% abstract data match. But the system still have problem getting the ISSN, and the publication date, and the system cannot give an accurate category indicated by the low data match, which is 7%. The system also needs the id profile of the lecturer on the resource website to be able to start the scraping process.

Keywords: *Web Scraping, Information Retrieval*, Google Scholar, SINTA, ScimagoJr

DAFTAR ISI

HALAMAN JUDUL	i
LEMBAR PENGESAHAN.....	ii
LEMBARAN PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS	iii
KATA PENGANTAR.....	iv
ABSTRAK	v
ABSTRACT	vi
DAFTAR ISI	vii
DAFTAR TABEL.....	ix
DAFTAR GAMBAR.....	x
DAFTAR SEGMENT PROGRAM	xii
1. PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah	2
1.3 Tujuan Penelitian.....	2
1.4 Manfaat Penelitian.....	2
1.5 Ruang Lingkup	2
1.6 Metodologi Penelitian	4
1.7 Sistematika Penulisan.....	4
2. LANDASAN TEORI	6
2.1 Tinjauan Pustaka	6
2.1.1 Publikasi Ilmiah.....	6
2.1.2 Web Scraping.....	7
2.1.3 Information Retrieval	7
2.1.4 Regular Expression	8
2.1.5 Automated Web Browser.....	8
2.1.6 Application Programming Interface	9
2.2 Tinjauan Studi.....	9
2.2.1 Social Media Web Scraping using Social Media Developers API and Regex (Dewi et al., 2019)	9
2.2.2 Optimization and Security in Information Retrieval, Extraction, Processing, and Presentation on a Cloud Platform (Alexandrescu, 2019).....	10
2.2.3 Web Scraping Techniques to Collect Weather Data in South Sumatera (Fatmasari et al., 2018).....	11

2.2.4	Implementation of Web Scraping for Journal Data Collection on the SINTA Website (Adila, 2022)	11
3.	ANALISIS DAN DESAIN SISTEM	12
3.1	Analisis Permasalahan	12
3.2	Analisis Kebutuhan	12
3.3	Flowchart Program	12
3.4	Entity Relationship Diagram (ERD)	17
3.4.1	Tabel Pegawai	18
3.4.2	Tabel Mitra	18
3.4.3	Tabel Mahasiswa	19
3.4.4	Tabel Kategori	19
3.4.5	Tabel Publikasi	20
3.4.6	Tabel Penulis	20
3.4.7	Tabel Authors	21
3.5	Desain User Interface	21
3.5.1	Login Page	21
3.5.2	List Pegawai	22
3.5.3	List Publikasi	23
3.5.4	Detail Publikasi	23
4.	IMPLEMENTASI SISTEM	25
4.1	Aplikasi Pemrograman	25
4.2	Implementasi Program	25
4.3	Implementasi Pengujian Data	37
5.	PENGUJIAN SISTEM	43
5.1	Pengujian Program	43
5.2	Pengujian API	46
5.2.1	Google Scholar	46
5.2.2	SINTA	49
5.2.3	ScimagoJr	58
5.3	Pengujian data	60
5.3.1	Pengambilan data hasil Scraping	60
6.	PENUTUP	66
6.1	Kesimpulan	66
6.2	Saran	66
7.	DAFTAR PUSTAKA	67

DAFTAR TABEL

Tabel 3.1 Tabel Pegawai.....	18
Tabel 3.2 Tabel Mitra	19
Tabel 3.3 Tabel Mahasiswa	19
Tabel 3.4 Tabel Kategori.....	19
Tabel 3.5 Tabel Publikasi	20
Tabel 3.6 Tabel Publikasi	20
Tabel 3.7 Tabel Authors	21
Tabel 5.1 Pengecekan kesamaan judul data <i>scraping</i> dengan data IKP2M	61
Tabel 5.2 Rata - rata kesamaan setiap field	61
Tabel 5.3 Masalah saat pengujian.....	63

DAFTAR GAMBAR

Gambar 2.1 Flowchart alur yang diusulkan.....	10
Gambar 3.1 Flowchart scraping Google Scholar	13
Gambar 3.2 Flowchart scraping SINTA.....	14
Gambar 3.3 Flowchart detail scraping SINTA.....	15
Gambar 3.4 Flowchart pengambilan Q rank dari Scimago	16
Gambar 3.5 Flowchart memasukkan data ke database.....	17
Gambar 3.6 Entity Relationship Diagram.....	17
Gambar 3.7 Desain Halaman Login	22
Gambar 3.8 Desain Halaman List Pegawai.....	22
Gambar 3.9 Desain Halaman List Publikasi	23
Gambar 3.10 Desain Halaman Detail Publikasi	24
Gambar 5.1 Halaman Login Aplikasi.....	43
Gambar 5.2 Halaman List Publikasi	44
Gambar 5.3 Detail Publikasi	44
Gambar 5.4 List Pegawai	45
Gambar 5.5 Registrasi User Baru	45
Gambar 5.6 Data Baru akan Masuk Database	45
Gambar 5.7 Halaman Google Scholar Profile dan Button Show More	46
Gambar 5.8 Respon Penghitungan Total Publikasi	47
Gambar 5.9 Halaman Detail Publikasi Pada Google Scholar.....	47
Gambar 5.10 Detail Tambahan dari Original Link	48
Gambar 5.11 Respon Pengambilan data Sumber Google Scholar	48
Gambar 5.12 List Bertambah Setelah Data Dimasukkan Database	49
Gambar 5.13 Contoh Halaman Profile Sinta	49
Gambar 5.14 Respon Pengambilan Data Publikasi dari SINTA View Books	50
Gambar 5.15 List Setelah Data View Buku Masuk ke Database	50
Gambar 5.16 Halaman SINTA view Google Scholar	51
Gambar 5.17 Respon Pengambilan Data view Google Scholar.....	52
Gambar 5.18 List Setelah Data View Google Scholar Dimasukkan ke Database	52
Gambar 5.19 Halaman SINTA View Scopus.....	53
Gambar 5.20 Respon Pengambilan Data Pada View Scopus	54
Gambar 5.21 List Setelah Data View Scopus Masuk ke Database	54
Gambar 5.22 Halaman SINTA View Web of Science	55
Gambar 5.23 Respon Pengambilan Data SINTA View Web of Science	56
Gambar 5.24 List Setelah Data View Web of Science Dimasukkan ke Database.....	56
Gambar 5.25 Halaman SINTA View Garuda	57
Gambar 5.26 Respon Pengambilan Data View Garuda.....	57
Gambar 5.27 List Setelah Data View Garuda Dimasukkan ke Database.....	58
Gambar 5.28 Landing Page ScimagoJr	58
Gambar 5.29 Program Mencari Detail Publikasi	59
Gambar 5.30 Halaman Detail Sumber Publikasi Pada ScimagoJr	59
Gambar 5.31 Respon Pengambilan Dara Q rank dari ScimagoJr	60
Gambar 5.32 Respon Pengambilan Data Berdasarkan Nama.....	60
Gambar 5.33 Id Profile Google Scholar	63
Gambar 5.34 Id Profile SINTA.....	64
Gambar 5.35 HTML yang diterima saat terkena bot detection	64

DAFTAR SEGMENT PROGRAM

Segment Program 4.1 Koneksi kepada database	25
Segment Program 4.2 Membuka website sumber Google Scholar	25
Segment Program 4.3 Memberi delay pada program	26
Segment Program 4.4 Pengambilan Data dari Google Scholar berdasarkan Index	26
Segment Program 4.5 Pengambilan data dari sumber ScimagoJr	28
Segment Program 4.6 Pengambilan Data dari Sumber SINTA View Google Scholar...	30
Segment Program 4.7 Pengambilan Data pada Sumber SINTA View Scopus	31
Segment Program 4.8 Pengambilan Data dari Sumber SINTA View Web of Science...	31
Segment Program 4.9 Pengambilan Data dari Sumber SINTA View Garuda	32
Segment Program 4.10 Pengambilan Data pada Sumber SINTA View Books	32
Segment Program 4.11 Memasukkan Data Publikasi ke Database	33
Segment Program 4.12 Melakukan Update jika Judul Sudah Ditemukan	34
Segment Program 4.13 Pengambilan Tahun dari Hasil Scraping.....	34
Segment Program 4.14 Memasukkan ISSN ke Dalam Database	35
Segment Program 4.15 Pemberian Kategori Journal.....	36
Segment Program 4.16 Memberikan Kategori Buku berdasarkan ISBN	36
Segment Program 4.17 Memasukkan Source, Keyword, dan Sitasi	37
Segment Program 4.18 Pengambilan Data dari Database Hasil Scraping	37
Segment Program 4.19 Pengambilan Data IKP2M	39
Segment Program 4.20 Pengecekan Kesamaan Field Publikasi.....	40