*Telematika: Jurnal Informatika dan Teknologi Informasi*
*Vol. 20, No. 2, Juni 2023, pp.139-152*
*ISSN: 1829-667X / E-ISSN: 2460-9021*
*DOI:10.31515/telematika.v20i2.9575*

# *Implementation of Web Scraping on Google Search Engine for Text Collection Into Structured 2D List*

Implementasi Web Scraping pada Google Search Engine untuk Pengumpulan Teks ke Dalam List 2D Terstruktur

**Tresna Maulana Fahrudin[1], Prismahardi Aji Riyantoko[2], Kartika Maulida Hindrayani[3]**

[1,2,3] Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia

[1*]tresna.maulana.ds@upnjatim.ac.id, [2]prismahardi.aji.ds@upnjatim.ac.id,
[3]kartika.maulida.ds@upnjatim.ac.id
**: Penulis korenspondensi* (*corresponding author*)

***Informasi Artikel***
*Received: April 2023*
*Revised: May 2023*
*Accepted: June 2023*
*Published: June 2023*

***Abstract***

*Purpose: This research proposes the implementation of web scraping on Google Search Engine to collect text into a structured 2D list.*
*Design/methodology/approach: Implementing two important stages in the process of collecting data through web scraping, namely the HTML parsing process to extract links (URL) on Google Search Engine pages, and HTML parsing process to extract the body text from website pages on each link that has been collected.*
*Findings/result: The inputted query is adjusted to the latest issues and news in Indonesia, for example the President's important figures, the month of Ramadan and Idul Fitri, riots tragedy (stadium) and natural disasters, rising prices of basic commodities, oil and gold, as well as other news. The least number of links obtained was 56 links and the most was 151 links, while the processing time to obtain links for each of the fastest queries was 1 minute 6.3 seconds and the longest was 2 minutes 49.1 seconds. The results of scraping links from these queries were obtained from Wikipedia, Detik, Kompas, the Election Supervisory Body (Bawaslu), CNN Indonesia, the General Election Commission (KPU), Pikiran Rakyat, and others.*
*Originality/value/state of the art: Based on previous research, this study provides an alternative to produce optimal collection of links and text from web scraping results in the form of a 2D list structure. Lists in the Python programming language can store character sequences in*

*the form of strings and can be accessed using index keys, and manipulate text efficiently.*

*Keywords:web scraping, google search engine, text collection, structure of 2D list, parsing HTML*
Kata kunci: web scraping, google search engine, pengumpulan teks, struktur list 2D, parsing HTML

**Abstrak**

Tujuan: Penelitian ini mengusulkan implementasi web scraping pada Google Search Engine untuk mengumpulkan teks ke dalam bentuk list 2D terstruktur.
Perancangan/metode/pendekatan: Menerapkan dua tahap penting dalam proses pengumpulan data melalui web scraping yakni proses *parsing* HTML untuk mengekstrak tautan (URL) pada halaman Google Search Engine, dan proses *parsing* HTML untuk mengekstrak *body* teks dari halaman website pada masing-masing tautan yang telah dikoleksi.
Hasil: Kueri yang diinputkan disesuaikan dengan isu dan berita terkini di Indonesia misalnya tokoh penting Presiden, bulan Ramadhan dan Idul Fitri, tragedi kerusuhan dan bencana alam, kenaikan harga bahan pokok, minyak, dan emas, serta berita lainnya. Jumlah tautan yang diperoleh yang paling sedikit adalah 56 tautan dan yang paling banyak adalah 151 tautan, sedangkan waktu pemrosesan untuk memperoleh tautan pada masing-masing kueri yang paling cepat adalah 1 menit 6.3 detik dan yang paling lama adalah 2 menit 49.1 detik. Hasil *scraping* tautan dari kueri tersebut di antaranya diperoleh dari Wikipedia, Detik, Kompas, Badan Pengawas Pemilihan Umum (Bawaslu), CNN Indonesia, Komisi Pemilihan Umum (KPU), Pikiran Rakyat, dan lainnya.
Keaslian/ *state of the art*: Berdasarkan penelitian terdahulu, penelitian ini memberikan alternatif untuk menghasilkan pengumpulan tautan dan teks dari hasil web scraping yang optimal dalam bentuk struktur list 2D. List dalam bahasa pemrograman Python dapat menyimpan urutan karakter dalam bentuk string dan dapat diakses menggunakan indeks kunci, dan memanipulasi teks secara efisien.

## 1. Pendahuluan

Menurut penelitian dan survei yang dilakukan oleh berbagai lembaga, penggunaan data tidak terstruktur terus meningkat dari tahun ke tahun. Estimasi persentase penggunaan data tidak terstruktur di beberapa industri perbankan dan keuangan mencapai sekitar 80%, industri kesehatan 80%, industri retail dan e-*commerce* 90%, industri manufaktur 70%, dan industri transportasi 70% [2]. Estimasi ini bersifat umum dan dapat berbeda-beda tergantung pada sumber dan jenis industri. Penggunaan data tidak terstruktur sangat penting di berbagai industri dan akan terus meningkat seiring dengan berkembangnya teknologi dan bisnis. Pertumbuhan

*ISSN: 1829-667X / E-ISSN: 2460-9021*
*DOI:10.31515/telematika.v20i2.9575*

data tidak terstruktur juga dipengaruhi masifnya penggunaan media sosial dan unggahan blog pada halaman website. Data yang semakin meningkat menjadi peluang yang baik untuk dikumpulkan dan dianalisis lebih lanjut menggunakan metode dan perangkat lunak terkini guna menghasilkan suatu wawasan dan kebijakan bagi pemangku kepentingan.

Google Search Engine merupakan salah satu mesin pencari paling populer dan banyak digunakan di dunia untuk pencarian data terstukrtur maupun tidak terstruktur. Menurut laporan StatCounter Global Stats pada Februari 2023, Google Search Engine memiliki pangsa pasar sebesar 81.62% di seluruh dunia [3]. Hal ini menunjukkan bahwa Google Search Engine masih menjadi pilihan utama bagi pengguna internet di seluruh dunia untuk mencari informasi. Google Search Engine juga merupakan salah satu mesin pencari terbesar dan paling populer di dunia yang dapat digunakan untuk mencari sumber data tidak terstruktur. Google menggunakan teknologi pemrosesan bahasa alami dan mesin pembelajaran untuk mengindeks dan memahami konten dari berbagai jenis sumber data, termasuk teks, gambar, audio, dan video. Algoritma Google akan mencocokkan kata kunci atau pertanyaan pengguna dengan konten di indeks Google yang terdiri dari miliaran halaman web dan sumber data tidak terstruktur lainnya, lalu menampilkan hasil yang paling relevan dan bermanfaat bagi pengguna [4].

Sumber data tidak terstruktur merupakan sumber data yang tidak memiliki format atau struktur yang teratur dan dapat diakses dengan mudah. Sumber data ini dapat berasal dari berbagai sumber, termasuk teks, audio, video, gambar, dan dokumen yang tidak memiliki format standar. Dalam lingkungan digital saat ini, sekitar 90% dari data yang terus bertambah merupakan data yang tidak terstruktur. Jenis data ini tidak sesuai untuk disimpan dalam database relasional sehingga dikembangkanlah skenario dengan memanfaatkan database NoSQL. Saat ini, terdapat empat keluarga database NoSQL yang meliputi *key-value*, orientasi kolom, orientasi grafik, dan orientasi dokumen. Banyak perusahaan terkemuka seperti Amazon, LinkedIn, Facebook, Google, dan YouTube menggunakan data NoSQL [1] dan mengganti database konvensional mereka dengan database NoSQL. Namun, database NoSQL biasanya didesain dengan fokus pada kecepatan dan skalabilitas dibandingkan konsistensi data, keterbatasan operasi kompleks seperti pengindeksan yang membutuhkan struktur data tambahan yang mengindeks nilai-nilai tertentu dalam database, konfigurasi server, instalasi driver, dan perawatan.

Terdapat struktur data yang dapat digunakan untuk memproses data tidak terstruktur seperti teks, terutama dalam hal mengakses dan memanipulasi teks secara efisien yakni menggunakan list. List dapat menyimpan urutan karakter dalam bentuk string dan dapat diakses menggunakan indeks kunci. List adalah struktur data bawaan dalam bahasa pemrograman Python sehingga tidak memerlukan instalasi dan konfigurasi tambahan. List memungkinkan untuk mudah membuat, mengakses, dan mengubah data dengan sintaksis Python yang sederhana. Namun, list juga memiliki keterbatasan dari sisi skalabilitas untuk menyimpan data yang sangat besar dan memerlukan pertumbuhan yang sangat cepat, keterbatasan kueri kompleks, dan keterbatasan persistensi untuk menyimpan data yang tahan lama.

Web scraping menjadi salah satu teknik untuk memperoleh data tidak terstruktur, teknik ini digunakan untuk mengambil data dari situs web secara otomatis menggunakan bot atau program komputer. Teknik ini dapat digunakan untuk mengumpulkan informasi dari berbagai sumber, termasuk mesin pencari seperti Google Search Engine. Implementasi web scraping pada Google Search Engine dapat membantu pengguna untuk mengumpulkan teks dari halaman pencarian Google dan mengubahnya menjadi struktur data, seperti list, tabel, dan spreadsheet. Penerapan

web scraping pada Google Search Engine dapat dilakukan dengan menggunakan berbagai teknologi, misalnya menggunakan bahasa pemrograman Python dengan dukungan pustaka BeautifulSoup dan Selenium [5]. Dalam prosesnya, program akan melakukan pencarian pada halaman Google dan mengekstrak teks dari hasil pencarian. Setelah itu, teks akan diolah menjadi struktur data yang lebih mudah diolah dan diatur.

Beberapa penelitian terkait implementasi web scraping pada Google Search Engine yang telah diteliti oleh penelitian lainnya diantaranya Rahmatulloh dan Gunawan mengusulkan penelitian tentang "Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar". Penelitian tersebut mengusulkan penggunaan teknik web scraping menggunakan metode HTML DOM (*Hypertext Markup Language-Document Object Model)* untuk mengumpulkan data artikel ilmiah dari Google Scholar [6]. Penelitiannya menjelaskan langkah-langkah untuk melakukan web scraping, mulai dari memahami struktur HTML hingga mengekstrak data yang diperlukan. Selain itu, juga memberikan contoh penggunaan teknik ini untuk mengumpulkan data artikel ilmiah dari Google Scholar. Hasil penelitian menunjukkan bahwa teknik web scraping dengan metode HTML DOM dapat efektif digunakan untuk mengumpulkan data artikel ilmiah dari Google Scholar dengan tingkat akurasi yang tinggi. Penelitian tersebut dapat bermanfaat bagi peneliti dan praktisi dalam mengumpulkan data dari Google Scholar secara efektif dan efisien.

Flores dkk, mengusung penelitian tentang "Penerapan Web Scraping Sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan Keyword". Penelitian tersebut membahas tentang penerapan teknik web scraping sebagai media pencarian dan penyimpanan artikel ilmiah secara otomatis berdasarkan keyword tertentu [7]. Metode web scraping yang digunakan adalah BeautifulSoup dan Scrapy. Eksperimen yang dilakukan dengan memasukkan kata kunci tertentu pada mesin pencarian Google Scholar dan menggunakan teknik web scraping untuk mengambil data dari setiap artikel ilmiah yang ditemukan. Kemudian, data tersebut disimpan dalam format CSV dan database MySQL. Hasil eksperimen menunjukkan bahwa teknik web scraping dapat digunakan untuk memperoleh data artikel ilmiah dengan mudah dan efektif. Teknik ini dapat membantu peneliti untuk menghemat waktu dan usaha dalam melakukan pencarian artikel ilmiah yang relevan dengan topik penelitian mereka.

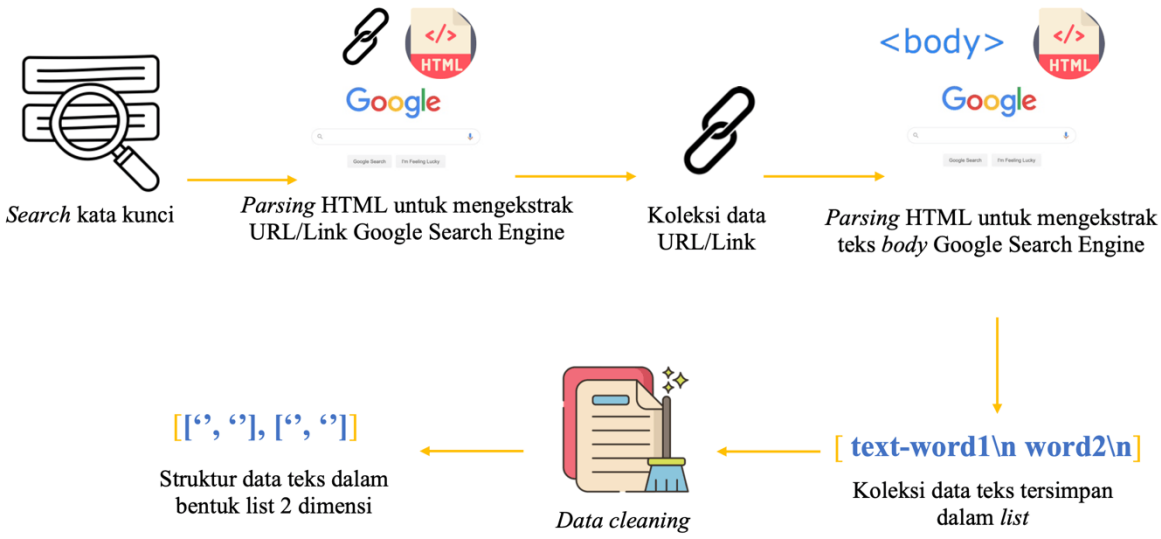
Manfaat dari implementasi web scraping pada Google Search Engine adalah pengguna dapat mengumpulkan informasi dengan cepat dan efisien. Dengan mengubah teks menjadi data yang lebih terstruktur, informasi tersebut dapat diolah dan dianalisis dengan lebih mudah. Selain itu, pengguna dapat mengumpulkan data secara otomatis sehingga tidak perlu melakukan pengumpulan menggunakan metode konvensional. Jika dibandingkan dengan penelitian terkait sebelumnya, penggunaan metode HTML DOM memiliki kekurangan yakni memerlukan *rendering* penuh halaman web sebelum dapat diakses dan diproses sehingga cukup memakan waktu terutama jika halaman web memiliki konten yang kompleks. Selain itu, penyimpanan data teks pada MySQL juga memiliki keterbatasan terkait panjang karakter yang harus disimpan. Oleh karena itu, penelitian ini mengusulkan implementasi web scraping pada Google Search Engine untuk mengumpulkan teks ke dalam bentuk list 2D terstruktur. Penelitian ini menjelaskan dua tahap penting dalam proses pengumpulan data melalui web scraping yakni a) proses *parsing* HTML untuk mengekstrak tautan (URL) pada halaman Google Search Engine, dan b) proses *parsing* HTML untuk mengekstrak *body* teks dari halaman website pada masing-masing tautan yang telah dikoleksi. Pada penelitian mencoba untuk menguji performa dari

*ISSN: 1829-667X / E-ISSN: 2460-9021*
*DOI:10.31515/telematika.v20i2.9575*

algoritma web scraping dan penyimpanan data teks menggunakan list yang dikembangkan menggunakan bahasa pemrograman Python berdasarkan input kata kunci tertentu, jumlah halaman, jumlah tautan yang diperoleh dan waktu pemrosesan. Selain itu, juga menguji keberhasilan web scraping dan penyimpanan data teks ke dalam bentuk data list 2D terstruktur.

## 2. Metode/Perancangan

Pada Gambar 1 menunjukkan desain sistem penelitian yang diusulkan, dimulai dari pencarian kata kunci (kueri) diinputkan oleh pengguna dan menentukan jumlah halaman Google Search Engine yang akan di-*scrape*. Lalu kata sistem akan melewatkan parameter tersebut ke struktur URL Google Search Engine yang terdiri dari *protocol*, *domain*, *path*, *query string*, dan tambahan jumlah *pagination.* Pustaka BeautifulSoup juga digunakan untuk melakukan *parsing* HTML agar komponen-komponen HTML menjadi rangkaian elemen yang dapat mudah dibaca. Untuk mendapatkan tautan dari berbagai sumber website sesuai kata kunci yang sudah diinputkan, Google Search Engine memiliki identifier *div class* yang bernama "yuRUbf", ini yang akan mempermudah dalam proses koleksi tautan tersebut. Setelah tautan website telah dikoleksi, tahap selanjutnya adalah mengumpulkan teks *body* halaman website dari masing-masing tautan. Dengan menggunakan metode atau fungsi *find_element_by_tag_name('body')* yang dimiliki oleh pustaka Selenium, *tag* yang digunakan adalah *'body'* yang merupakan *tag* utama pada setiap halaman web dan menandakan seluruh konten halaman web. Metode ini dapat mengakses elemen *tag* HTML *'body'* dan melakukan interaksi atau pengambilan informasi pada elemen tersebut. Jika teks *body* telah terkoleksi, dilanjutkan dengan proses pembersihan teks dan mengubah ke struktur list 2D (2 Dimensi).




**Gambar 1**. Desain Sistem Penelitian yang Diusulkan

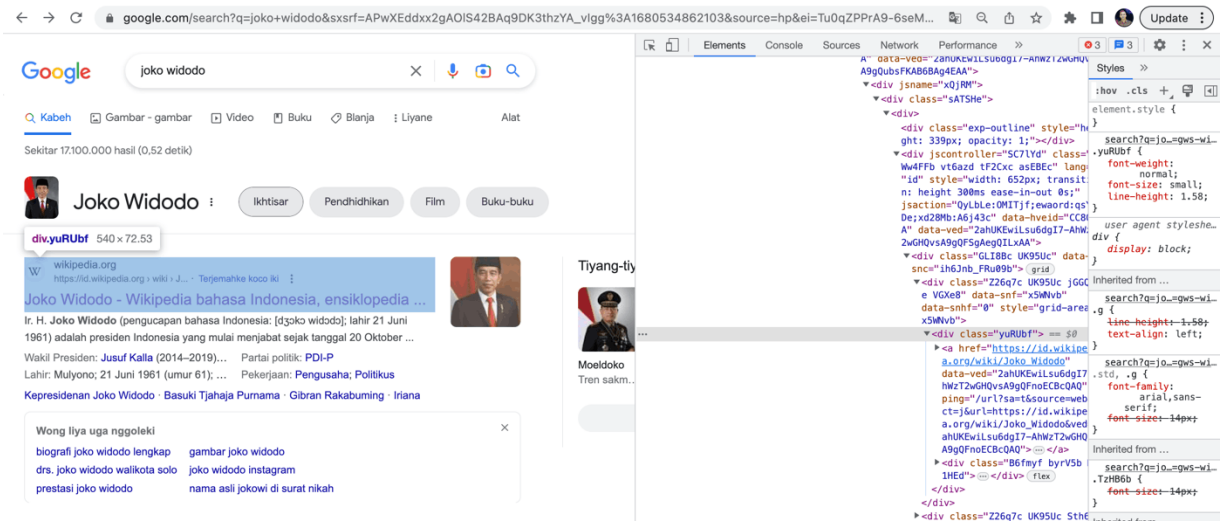
### 2.1. Struktur Elemen Halaman Google Search Engine

Elemen halaman Google Search Engine terdiri atas halaman *search* dan *pagination*. Pada halaman *search*, pengguna akan disajikan dengan kotak pencarian dan tombol pencarian. Pada halaman *pagination,* pengguna akan disajikan hasil pencarian sesuai urutan penomoran halaman

*ISSN: 1829-667X / E-ISSN: 2460-9021*
*DOI:10.31515/telematika.v20i2.9575*

dan filter pencarian. *Pagination* pada Google Search Engine merupakan seluruh hasil pencarian dari mesin pencari [8] yang ditampilkan untuk memudahkan pengguna menelusuri daftar item atau artikel yang panjang. Selain Google Search Engine, terdapat Yahoo! dan Bing yang dapat digunakan sebagai mesin pencari [9]. Namun, Google merupakan salah satu mesin pencari yang paling populer dibandingkan mesin pencari lainnya [10]. Hasil kualitas pencarian informasi dari Google Search Engine menjadi faktor utama bagi pengguna. Posisi urutan dan nomor halaman dari hasil pencarian berpengaruh terhadap interaksi pengguna dengan persepsi dan pengetahuan mereka sehingga menjadi salah satu faktor apakah suatu tautan akan mendapatkan lebih banyak perhatian untuk ditelusuri oleh pengguna [11].

Pada Gambar 2 menunjukkan inspeksi elemen pada halaman hasil pencarian pada Google Search Engine. Kode sintaks *div class* "yuRUbf" ini merupakan kelas *Cascading Style Sheets* (CSS) yang digunakan untuk mengelompokkan tautan yang muncul pada hasil pencarian. Sebagai contoh, jika pada halaman pertama pada hasil pencarian Google Search Engine terdapat tautan sebanyak 10 website, maka hasil *scraping* akan mengumpulkan sebanyak 10 tautan tersebut dan akan bertambah sesuai dengan jumlah halaman *pagination* yang dicari.




**Gambar 2**. Struktur Elemen Halaman Hasil Pencarian pada Google Search Engine

## 2.2. Teknik-teknik web scraping

Terdapat beberapa teknik web scraping secara umum untuk memperoleh data dari suatu halaman website, antara lain Parsing HTML, DOM Parsing, Xpath, dan Reguler Expression (RegEx). Berikut penjelasan masing-masing teknik web scraping:

### 2.2.1. Parsing HTML

Parsing HTML adalah proses menganalisis dan menafsirkan struktur dan konten dokumen HTML yang digunakan untuk membuat halaman web dan konten web lainnya. Parsing HTML melibatkan penggunaan bahasa pemrograman atau alat khusus untuk mengekstrak elemen atau data tertentu dari dokumen HTML sehingga dapat digunakan untuk berbagai keperluan seperti web scraping, penambangan data, atau membuat aplikasi web yang disesuaikan [12].

### 2.2.2. DOM Parsing

DOM merupakan standar untuk mendapatkan, mengubah, menambah, atau menghapus elemen HTML. DOM Parsing melibatkan pembuatan pohon DOM dari dokumen HTML atau XML, yang terdiri dari node yang mewakili berbagai elemen dan konten dalam dokumen. Setelah pohon DOM dibangun, selanjutnya dapat menggunakan API DOM untuk menelusuri dan memanipulasi node dalam pohon, termasuk menambah atau menghapus elemen, mengubah atribut, atau mengekstrak data dari dokumen [13].

### 2.2.3. XPath

XPath merupakan elemen utama di XSLT standar (eXtensible Stylesheet Language Transformation). XML (eXtensible Markup Language) dokumen dapat dinavigasi elemen dan atributnya oleh XPath [14]. Xpath merupakan bahasa untuk memilih node di dokumen XML dan juga dapat digunakan dengan HTML. Ekspresi XPath yang paling berguna adalah lokasi *path*. Lokasi *path* setidaknya menggunakan lokasi satu langkah untuk mengidentifikasi sekumpulan node dalam dokumen. Jalur lokasi paling sederhana adalah jalur yang memilih simpul akar dokumen dengan menggunakan garis miring "/". Simbol tersebut adalah *root* dari file sistem Unix dan juga node akar dari sebuah dokumen.

### 2.2.4. Reguler Expression (RegEx)

Reguler Expression (RegEx) merupakan sebuah formula dengan pola yang spesifik yang mendeskripsikan himpunan kata beberapa alfabet. Regex dapat digunakan untuk menyesuaikan pola dengan beberapa kata. Selain menjadi gagasan mendasar dalam teori bahasa formal, RegEx banyak digunakan dalam ilmu komputer untuk menentukan pola pencarian. Secara formal, RegEx didefinisikan dengan pola *p* dari ukuran *m* dan urutan simbol (teks) *t* panjang *n*, tujuan dari pencocokan RegEx adalah untuk memeriksa apakah *substring* dari *t* dapat diturunkan dari *p* [15].

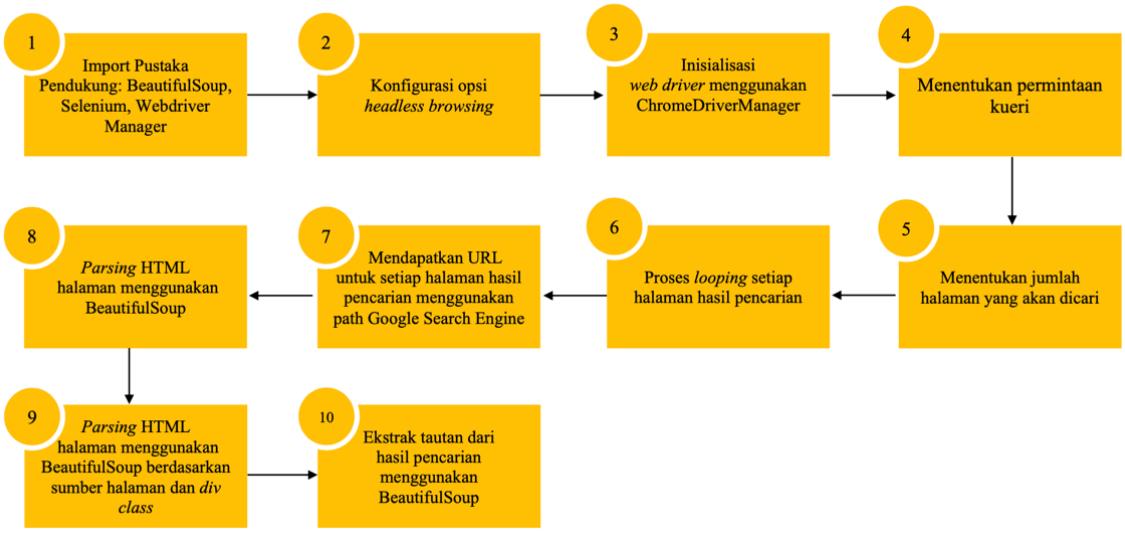
## 2.3. BeautifulSoup untuk Web Scraping

BeautifulSoup adalah pustaka Python yang biasa digunakan untuk mem-*parsing* dokumen HTML dan XML. Ini memberikan cara yang sederhana dan efisien untuk menavigasi dan mencari melalui elemen dan konten halaman web atau dokumen, BeautifulSoup menjadi alternatif yang baik untuk aplikasi web scraping dan data mining [16]. Kelebihan dari BeautifulSoup antara lain kemudahan dalam penggunaan dan efisiensi yang diterapkan. Dengan menggunakan BeautifulSoup dapat mengambil tag HTML, atribut, dan konten di dalamnya dengan mudah sehingga dapat melakukan ekstraksi data dengan lebih efektif dan efisien. Selain itu, BeautifulSoup juga menyediakan fitur untuk memperbaiki struktur dokumen HTML yang tidak valid atau rusak sehingga memudahkan dalam melakukan ekstraksi data pada dokumen yang tidak sempurna.

## 2.4. Web Scraping untuk Pengumpulan Teks menjadi Struktur List 2D

Penelitian ini mengusulkan implementasi web scraping untuk mengumpulkan teks dari Google Search Engine menjadi struktur list 2D. Tahapan utama pada web scraping ini adalah proses *parsing* HTML untuk mengekstrak tautan (URL) pada halaman Google Search Engine seperti yang ditunjukkan pada Gambar 3, dan proses *parsing* HTML untuk mengekstrak *body* teks dari halaman website pada masing-masing tautan yang telah dikoleksi seperti yang ditunjukkan pada Gambar 4.

*ISSN: 1829-667X / E-ISSN: 2460-9021*
*DOI:10.31515/telematika.v20i2.9575*




**Gambar 3**. *Parsing* HTML untuk Mengekstrak Tautan (URL) pada Halaman Google Search Engine


**Gambar 4**. *Parsing* HTML untuk Mengekstrak *Body* Teks (URL) dari Masing-masing Tautan Halaman Website

## 3. Hasil dan Pembahasan

Hasil implementasi web scraping pada Google Search Engine untuk mengumpulkan teks ke dalam struktur list 2D didapatkan sejumlah tautan dan waktu pemrosesan berdasarkan masing-masing kueri yang diinputkan. Selain itu, tiap tautan telah diperoleh pengumpulan teks dari hasil *scraping* pada masing-masing halaman website. Pada Gambar 5 menunjukkan kode perintah mengeksekusi file Python untuk web scraping tautan (URL) pada Google Search Engine. Pada file Python tersebut mencakup penggunaan pustaka BeautifulSoup, Selenium, Web Driver Manager (Chrome), input kueri, dan jumlah *pagination.* Selain itu, struktur URL pada Google

*ISSN: 1829-667X / E-ISSN: 2460-9021*
*DOI:10.31515/telematika.v20i2.9575*

Search Engine yang terdiri dari *protocol*, *domain*, *path*, *query string*, dan tambahan jumlah *pagination* dituliskan secara eksplisit pada kode Python sebagai berikut *url = "http://www.google.com/search?q=" + query + "&start=" + str((page - 1) * 10).* Selanjutnya, untuk mengumpulkan tautan pada tiap-tiap website dari hasil *scraping* halaman Google Search Engine menggunakan *div class* "yuRUbf". Hasil *scraping* yang berisi tautan-tautan website dikumpulkan secara kolektif di suatu file berekstensi *.txt.


**Gambar 5**. Eksekusi File Python untuk Web Scraping Tautan (URL) pada Google Search Engine

Pada Tabel 1 menunjukkan hasil web scraping tautan (URL) pada halaman Google Search Engine dengan menampilkan informasi kueri, jumlah halaman, jumlah tautan, dan waktu pemrosesan. Kueri yang diinputkan disesuaikan dengan isu dan berita terkini di Indonesia misalnya tokoh penting Presiden, bulan Ramadhan dan Idul Fitri, tragedi kerusuhan dan bencana alam, kenaikan harga bahan pokok, minyak, dan emas, serta berita lainnya. Jumlah tautan yang diperoleh yang paling sedikit adalah 56 tautan dan yang paling banyak adalah 151 tautan, sedangkan waktu pemrosesan untuk memperoleh tautan pada masing-masing kueri yang paling cepat adalah 1 menit 6.3 detik dan yang paling lama adalah 2 menit 49.1 detik. Ini menunjukkan bahwa hasil *scraping* pada masing-masing input kueri telah diperoleh sejumlah tautan yang optimal dengan waktu pemrosesan yang efisien.

**Tabel 1**. Hasil Web Scraping Tautan (URL) pada Halaman Google Search Engine

| No. | Kueri | Jumlah Halaman | Jumlah Tautan | Waktu Pemrosesan |
|---|---|---|---|---|
| 1. | Joko Widodo | 35 | 151 | 2 menit 15.6 detik |
| 2. | Ramadhan di Indonesia | 35 | 135 | 1 menit 22.8 detik |
| 3. | Kenaikan Harga BBM | 35 | 133 | 1 menit 21.9 detik |
| 4. | Tragedi Kanjuruhan | 35 | 94 | 1 menit 37.9 detik |
| 5. | Transaksi Rp349 Triliun Kemenkeu | 35 | 87 | 1 menit 14.2 detik |
| 6. | Piala Dunia U-20 batal di Indonesia | 35 | 56 | 1 menit 6.3 detik |
| 7. | Pemilu Indonesia 2024 | 35 | 91 | 2 menit 10.4 detik |
| 8. | Gempa Bumi di Indonesia | 35 | 95 | 1 menit 52.6 detik |
| 9. | Sepeda Motor listrik | 35 | 127 | 1 menit 46.4 detik |
| 10. | Istri Pejabat Pamer Harta | 35 | 66 | 1 menit 34.7 detik |
| 11. | Harga Bahan Pokok Jelang Ramadhan | 35 | 122 | 1 detik 55.5 detik |
| 12. | Insiden Petasan Ramadhan-Idul Fitri | 35 | 79 | 1 detik 34.0 detik |
| 13. | Mudik Lebaran 2023 | 35 | 79 | 1 detik 43.5 detik |
| 14. | Tol Macet | 35 | 92 | 2 menit 26.0 detik |
| 15. | Harga Emas Naik Turun | 35 | 101 | 1 menit 30.4 detik |

*ISSN: 1829-667X / E-ISSN: 2460-9021*
*DOI:10.31515/telematika.v20i2.9575*

| | | | | |
|---|---|---|---|---|
| 16. | Seleksi Ujian Masuk Perguruan Tinggi | 35 | 101 | 1 menit 30.4 detik |
| 17. | Kunjungan Turis Asing | 35 | 163 | 1 menit 55.4 detik |
| 18. | Pembayaran THR Karyawan | 35 | 126 | 2 menit 49.1 detik |
| 19. | Kasus Bullying Anak | 35 | 145 | 1 menit 49.9 detik |
| 20. | Kenaikan Harga Minyak Dunia | 35 | 114 | 2 menit 14.1 detik |

Pada Tabel 2 menunjukkan daftar tautan (URL) pada halaman Google Search Engine dengan salah satu contoh kueri yakni "Pemilu Indonesia 2024". Hasil *scraping* tautan dari kueri tersebut di antaranya diperoleh dari Wikipedia, Detik, Kompas, Badan Pengawas Pemilihan Umum (Bawaslu), CNN Indonesia, Komisi Pemilihan Umum (KPU), Pikiran Rakyat, dan lainnya.

**Tabel 2**. Daftar Tautan (URL) pada Halaman Google Search Engine dengan Kueri "Pemilu Indonesia 2024"

| **No.** | **Tautan** |
|---|---|
| 1. | https://id.wikipedia.org/wiki/Pemilihan_umum_Presiden_Indonesia_2024 |
| 2. | https://id.wikipedia.org/wiki/Pemilihan_umum_legislatif_Indonesia_2024 |
| 3. | https://news.detik.com/pemilu/d-6521367/apa-saja-yang-dipilih-dalam-pemilu-2024-cek-infonya-di-sini |
| 4. | https://regional.kompas.com/read/2023/02/04/162049878/daftar-6-nama-capres-cawapres-2024-yang-muncul-pada-musra-jateng-ada-ganjar?page=all |
| 5. | https://pekalongankab.bawaslu.go.id/berita/detail/sah-pemilu-14-februari-2024-disepakati-dpr-pemerintah-dan-penyelenggara-pemilu |
| 6. | https://pekalongankab.bawaslu.go.id/berita/detail/sah-pemilu-14-februari-2024-disepakati-dpr-pemerintah-dan-penyelenggara-pemilu |
| 7. | https://indonesiabaik.id/infografis/tahapan-dan-jadwal-pemilu-2024 |
| 8. | https://www.cnnindonesia.com/tag/pemilu-2024 |
| 9. | https://infopemilu.kpu.go.id/Pemilu/Peserta_pemilu |
| 10. | https://news.detik.com/pemilu |
| … | … |
| 90. | https://parwainstitute.id/article/5-parpol-sepakat-milenial-harus-ambil-peran-di-pemilu-2024 |
| 91. | https://indobalinews.pikiran-rakyat.com/politik/pr-883871975/wacana-penundaan-pemilu-2024-poros-peduli-indonesia-bertentangan-dengan-konstitusi-dan-demokrasi |

Pada Gambar 6 menunjukkan kode perintah mengeksekusi file Python untuk web scraping teks pada masing-masing tautan (URL) halaman website. File yang berisi tautan (URL) dari hasil scraping sebelumnya digunakan sebagai input dan juga menyediakan file output sebagai tempat untuk mengumpulkan teks hasil *scraping*. Proses *scraping* teks menggunakan kode Python berikut *elems = driver.find_element_by_tag_name('body').text* untuk membaca halaman HTML pada bagian *body* teks.


**Gambar 6**. Eksekusi File Python untuk Web Scraping Teks pada Masing-masing Tautan (URL) Halaman Website

*ISSN: 1829-667X / E-ISSN: 2460-9021*
*DOI:10.31515/telematika.v20i2.9575*

Pada Tabel 3 menunjukkan hasil web scraping teks dari tautan (URL) halaman web dengan salah satu contoh kueri "Pemilu Indonesia 2024". Seluruh teks disimpan dalam bentuk struktur list 2D, di mana teks disimpan berdasarkan indeks nomor urut tautan dan indeks nomor urut teks yang sudah difilter minimal sebanyak 5 kata. Proses *cleansing* yang telah dilakukan misalnya menghapus '\n' pada teks mentah. Terlihat pada output bahwa teks dikoleksi dalam bentuk struktur list 2D yang ditandai dengan kurung siku [[…],[…]].

**Tabel 3**. Hasil Web Scraping Teks dari Tautan (URL) Halaman Web dengan Kueri "Pemilu Indonesia 2024"

| No. | Tautan | Teks |
|---|---|---|
| 1. | https://id.wikipedia.org/wiki/Pemilihan_umum_Presiden_Indonesia_2024 | [['…., 'Dari Wikipedia bahasa Indonesia, ensiklopedia bebas', 'Artikel ini adalah bagian dari seri', 'Pemilihan Umum Presiden Indonesia 2024 adalah sebuah proses demokrasi untuk memilih Presiden dan Wakil Presiden Republik Indonesia untuk masa bakti 2024–2029 yang akan dilaksanakan pada Rabu, 14 Februari 2024. Pemilihan ini akan menjadi pemilihan presiden langsung kelima di Indonesia. Presiden petahana Joko Widodo dan mantan Presiden Susilo Bambang Yudhoyono tidak dapat maju kembali dalam pemilihan ini karena dicegah oleh undang-undang yang melarang periode ketiga untuk seorang presiden. Pemilihan umum ini akan dilaksanakan bersamaan dengan Pemilihan Umum anggota DPR, DPD, dan DPRD seluruh Indonesia sementara Pemilihan Umum Kepala Daerah baru akan dilaksanakan pada bulan November.', 'Pemilihan umum Presiden dan Wakil Presiden diatur dalam Pasal 6A dan Pasal 22E Undang-Undang Dasar Negara Republik Indonesia Tahun 1945 dan oleh Undang-Undang tentang Pemilihan Umum. Pasangan calon Presiden dan Wakil Presidden diusulkan oleh partai politik atau gabungan partai politik yang memperoleh sedikitnya 20% kursi Dewan Perwakilan Rakyat atau sedikitnya 25% suara nasional pada pemilihan umum sebelumnya. Dengan begitu hanya PDI-P saja yang dapat mengusulkan pasangan calon tanpa berkoalisi. Pemilihan umum Presiden dan Wakil Presiden dilakukan dengan dua putaran apabila pada putaran pertama tidak ada pasangan calon yang memperoleh lebih dari 50% suara dengan sedikitnya 20% suara yang tersebar di lebih dari setengah provinsi di Indonesia. Hingga saat ini, pemilihan umum Presiden dan Wakil Presiden dua putaran hanya pernah terjadi pada Pemilihan Umum Presiden dan Wakil Presiden 2004.', 'Potret resmi Jokowi dan pelantikan kedua pada tahun 2019.', 'Pasal 7 Undang-Undang Dasar Negara Republik Indonesia Tahun 1945 menyatakan:', '" Presiden dan Wakil Presiden memegang jabatan selama lima tahun, dan sesudahnya dapat dipilih kembali dalam jabatan yang sama, hanya untuk satu kali masa jabatan. "', ….'], |

*ISSN: 1829-667X / E-ISSN: 2460-9021*
*DOI:10.31515/telematika.v20i2.9575*

| No. | Tautan | Teks |
|---|---|---|
| | | ['Ikuti Wikipedia bahasa Indonesia di Facebook, Twitter, Instagram, dan Telegram', 'Dari Wikipedia bahasa Indonesia, ensiklopedia bebas', ' 2019 14 Februari 2024 2029 ', '(Dewan Perwakilan Rakyat: 580; Dewan Perwakilan Daerah: 152', ' Ketua Megawati Soekarnoputri Airlangga Hartarto Prabowo Subianto', 'Aliansi Koalisi Indonesia Bersatu Koalisi Kebangkitan Indonesia Raya', 'Ketua sejak 24 Maret 1999 13 Desember 2017 20 September 2014', 'Kursi saat ini 128 85 78', ' Ketua Surya Paloh Muhaimin Iskandar Agus Harimurti Yudhoyono', 'Aliansi Koalisi Perubahan untuk Persatuan Koalisi Kebangkitan Indonesia Raya Koalisi Perubahan untuk Persatuan', 'Ketua sejak 25 Januari 2013 25 Mei 2005 15 Maret 2020', 'Kursi saat ini 59 58 54', ' Ketua Ahmad Syaikhu Zulkifli Hasan Muhamad Mardiono', 'Aliansi Koalisi Perubahan untuk Persatuan Koalisi Indonesia Bersatu Koalisi Indonesia Bersatu', 'Ketua sejak 10 Mei 2020 1 Maret 2015 5 September 2022', 'Kursi saat ini 50 44 19', 'Pemilihan |
| 2. | https://id.wikipedia.org/wiki/Pemilihan_umum_legislatif_Indonesia_2024 | Umum Anggota Dewan Perwakilan Rakyat, Dewan Perwakilan Daerah, dan Dewan Perwakilan Rakyat Daerah 2024 (biasa disingkat Pemilu Legislatif 2024) adalah Pemilihan Umum Indonesia yang akan diselenggarakan pada tanggal 14 Februari 2024 untuk memilih anggota Dewan Perwakilan Rakyat (DPR), anggota Dewan Perwakilan Daerah (DPD), serta anggota Dewan Perwakilan Rakyat Daerah (DPRD Provinsi maupun DPRD Kabupaten/Kota) se-Indonesia periode 2024–2029.', 'Pemilihan umum presiden dan legislatif Indonesia 2024', 'Pemilu Legislatif (Pileg) tahun tersebut dilaksanakan bersamaan dengan Pemilihan umum Presiden Indonesia 2024 (Pilpres) dan Pemilihan Kepala Daerah (Pilkada).', 'Pelaksanaan Pileg, Pilpres dan Pilkada di waktu bersamaan ini masih menimbulkan kontroversi, bahkan digugat ke Mahkamah Konstitusi.', 'Alokasi kursi parlemen[sunting \| sunting sumber]', 'Pemilihan umum legislatif di Indonesia: Februari 2024', 'Nasional Dewan Perwakilan Rakyat (DPR) 580', 'Nasional Dewan Perwakilan Daerah (DPD) 152', 'Provinsi Dewan Perwakilan Rakyat Daerah I (DPRD I) 2,372', 'Kabupaten/Kota Dewan Perwakilian Rakyat Daerah II (DPRD II) 17,510', 'Peserta pemilihan umum Anggota DPR[sunting \| sunting sumber]', 'Dari 40 partai politik yang mendaftar, hanya terdapat 17 partai yang memenuhi syarat administrasi dan verifikasi faktual secara nasional. Verifikasi ini mencakup keberadaan pengurus inti parpol di tingkat pusat, keterwakilan perempuan minimal 30% dan domisili kantor tetap di tingkat DPP. Kemudian, di tingkat Provinsi, ada tambahan syarat, yakni memenuhi keanggotaan di 75% Kabupaten/Kota di 34 provinsi. Syarat terakhir, yakni status sebaran pengurus minimal 50% kecamatan pada 75% Kabupaten/Kota di 34 provinsi. Urutan partai politik peserta Pemilu Legislatif 2024 adalah sebagai berikut.', …'], |
| … | … | … ] |

Dengan demikian, hasil *scraping* yang telah dikumpulkan dalam bentuk struktur list 2D dapat digunakan untuk input data dan dilakukan analisis berikutnya, misalnya digunakan untuk mengubah dari teks ke vektor, pemodelan pembelajaran mesin, analisis sentimen, dan model analisis lainnya.

## 4. Kesimpulan dan Saran

Web scraping menjadi alternatif teknik pengumpulan data tidak terstruktur, di mana data mentah dapat disimpan dalam berbagai format yang terstruktur atau semi struktur misalnya dalam bentuk CSV, JSON, MySQL, dan lainnya. Dalam penelitian ini, data teks mentah telah berhasil disimpan dalam bentuk struktur list 2D. Selain itu, teknik untuk mengekstrak elemen pada halaman website Google Search Engine menggunakan *parsing* HTML dengan pustaka BeautifulSoup, Selenium, dan Web Driver Manager. Penelitian ini telah mencapai dua tujuan utama dalam proses pengumpulan data melalui web scraping yakni proses *parsing* HTML untuk mengekstrak tautan (URL) pada halaman Google Search Engine, dan proses *parsing* HTML untuk mengekstrak *body* teks dari halaman website pada masing-masing tautan yang telah dikoleksi. Kueri yang diinputkan disesuaikan dengan isu dan berita terkini di Indonesia misalnya tokoh penting Presiden, bulan Ramadhan dan Idul Fitri, tragedi kerusuhan dan bencana alam, kenaikan harga bahan pokok, minyak, dan emas, serta berita lainnya. Jumlah tautan yang diperoleh yang paling sedikit adalah 56 tautan dan yang paling banyak adalah 151 tautan, sedangkan waktu pemrosesan untuk memperoleh tautan pada masing-masing kueri yang paling cepat adalah 1 menit 6.3 detik dan yang paling lama adalah 2 menit 49.1 detik. Hasil *scraping* tautan dari kueri tersebut di antaranya diperoleh dari Wikipedia, Detik, Kompas, Badan Pengawas Pemilihan Umum (Bawaslu), CNN Indonesia, Komisi Pemilihan Umum (KPU), Pikiran Rakyat, dan lainnya. Untuk penelitian berikutnya, diperlukan teknik untuk membuat IP secara acak atau menggunakan proxy dalam proses *scraping* teks dikarenakan pada beberapa kali percobaan menggunakan IP yang bersifat tetap pada jaringan internet WIFI berpengaruh terhadap dibloknya IP oleh Google Search Engine. Selain itu, perlu diperlukan beberapa teknik praproses teks agar teks lebih bersih misalnya membersihkan tanda, angka, dan simbolik lainnya.

## Daftar Pustaka

[1] S. Praveen and U. Chandra, "NoSQL Products : IT Giants Perspectives," *Int. J. Comput. Intell. Res.*, vol. 13, no. 8, pp. 2125–2133, 2017.

[2] W. G. Swajati, "Kajian Kebijakan dan Sistem Pengelolaan Data Penelitian Indonesia," Jakarta, 2021.

[3] S. C. GlobalStats, "Search Engine Host Market Share Worldwide," *GlobalStats, Stat Counter*, 2023. https://gs.statcounter.com/search-engine-host-market-share (accessed Apr. 03, 2023).

[4] Google, "How Google Search Works," *Google Search*, 2023. https://www.google.com/search/howsearchworks/how-search-works/ (accessed Apr. 03, 2023).

[5] S. Fatima, S. Luqmaan, and N. A. Rasheed, "Web Scraping with Python and Selenium," *IOSR J. Comput. Eng.*, vol. 23, no. 3, pp. 1–5, 2021, doi: 10.9790/0661-2303020105.

[6] A. Rahmatulloh and R. Gunawan, "Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar," *Indones. J. Inf. Syst.*, vol. 2, no. 2, pp. 95–104, 2020, doi: 10.24002/ijis.v2i2.3029.

[7] V. A. Flores, P. A. Permatasari, and L. Jasa, "Penerapan Web Scraping Sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan Keyword," *Maj. Ilm. Teknol. Elektro*, vol. 19, no. 2, pp. 157–162, 2020, doi:

10.24843/mite.2020.v19i02.p06.

[8] L. Gotsev and E. Shoikova, "An Analysis of Scientific Production in Big Data Knowledge Domain on Google Books, YouTube and IEEE Explore® Digital Library," in *Proceedings of the 2020 4th International Conference on Cloud and Big Data Computing*, 2020, pp. 10–14, doi: 10.1145/3416921.3416936.

[9] W. Nel, L. De Wet, and R. Schall, "Randomised Controlled Trial of the Usability of Major Search Engines (Google, Yahoo! And Bing) When using Ambiguous Search Queries," in *Proceedings of the 4th International Conference on Computer-Human Interaction Research and Applications (CHIRA 2020)*, 2020, no. November, pp. 152–161, doi: 10.5220/0010133601520161.

[10] C. Ziakis, M. Vlachopoulou, T. Kyrkoudis, and M. Karagkiozidou, "Important Factors for Improving Google Search Rank," *Futur. Internet*, vol. 11, no. 32, pp. 1–12, 2019, doi: 10.3390/fi11020032.

[11] D. Trielli and N. Diakopoulos, "Search as News Curator: The Role of Google in Shaping Attention to News Information," in *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, 2019, pp. 1–15, doi: 10.1145/3290605.3300683.

[12] P. C. Patil, P. M. Chawan, and P. M. Chauhan, "Parsing of HTML Document," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 4, pp. 320–324, 2012.

[13] M. Radilova, P. Kamencay, R. Hudec, M. Benco, and R. Radil, "Tool for Parsing Important Data from Web Pages," *Appl. Sci.*, vol. 12, no. 12031, pp. 1–18, 2022, doi: 10.3390/app122312031.

[14] R. Gunawan, A. Rahmatulloh, I. Darmawan, and F. Firdaus, "Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath," *Atl. Highlights Eng.*, vol. 2, no. IcoIESE 2018, pp. 283–287, 2019, doi: 10.2991/icoiese-18.2019.50.

[15] A. Backurs and P. Indyk, "Which Regular Expression Patterns Are Hard to Match?," *Proc. - Annu. IEEE Symp. Found. Comput. Sci. FOCS*, pp. 1–33, 2016, doi: 10.1109/FOCS.2016.56.

[16] V. Bhateja, S. C. Satapathy, and H. Satori, *Embedded Systems and Artificial Intelligence*, vol. 1171. Singapore: Advances in Intelligent Systems and Computing, 2020.