

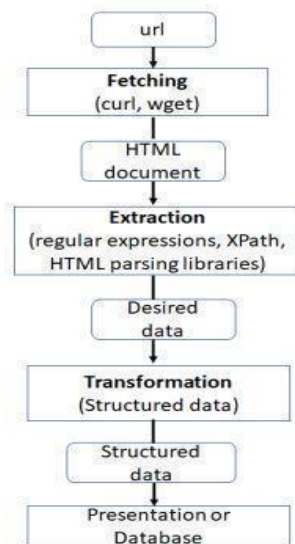
2. LANDASAN TEORI

2.1 Tinjauan Pustaka

1. *Web Data Scraping*

Web data scraping adalah sebuah teknik untuk mengubah data *web* yang tidak terstruktur menjadi data yang terstruktur untuk disimpan dan dianalisis ke dalam sebuah *spreadsheet* atau *database* (Khder, 2021). Proses *Web Scraping* dibagi menjadi 3 langkah, yaitu:

- a. *Fetching stage*: mengakses web yang memiliki informasi relevan dengan *HTTP* protokol (Khder, 2021).
- b. *Extraction stage*: setelah menerima halaman HTML, data yang penting harus diekstraksi. *Regular expressions*, *HTML parsing libraries*, dan *XPath queries* digunakan pada langkah ini (Khder, 2021).
- c. *Transformation stage*: setelah data relevan terkumpul, maka data tersebut dapat dikonversikan menjadi *format* terstruktur untuk ditampilkan atau disimpan (Khder, 2021).



Gambar 2.1 *Web Scraping Process* (Khder, 2021)

2. Toko Rental Kostum XYZ

Toko Rental Kostum XYZ adalah *Costume Center* yang berpusat di kota Surabaya, yang menjual dan menyewakan aneka kostum pilihan berkualitas, lengkap, bersih, dan terjangkau. Gudang Kostum sudah dipercaya banyak lembaga Perusahaan maupun Pemerintahan, Sekolah, maupun pribadi yang sedang mengadakan ACARA-ACARA spesial seperti *Gala Dinner* perusahaan, Ulang tahun perusahaan, Kostum untuk Ulang Tahun dan Perayaan lainnya, Acara 1001 Malam di Akhir Tahun, Lomba Fashion anak, *Modelling*, dan juga kostum Fotografi. Dengan menjaga kualitas layanan, Gudang Kostum telah menyewakan kostum terbaiknya di seluruh Indonesia.

3. *Regular Expression*

Regular Expression (juga dikenal sebagai *regex* atau bahkan hanya *re*) adalah rangkaian karakter (huruf, angka, dan karakter khusus) yang membentuk pola yang dapat digunakan untuk mencari teks guna melihat apakah teks tersebut berisi rangkaian karakter yang cocok dengan pola. Misalnya, Anda mungkin memiliki pola yang didefinisikan sebagai tiga karakter diikuti oleh tiga angka. Pola ini dapat digunakan untuk mencari pola serupa pada *string* lainnya.

4. Katalog

Menurut Fathmi (2004), katalog berarti daftar berbagai jenis koleksi perpustakaan yang disusun menurut sistem tertentu. Pada katalog *online* Toko Rental XYZ dibuat agar pengunjung dapat melakukan pencarian kostum sesuai dengan kebutuhan.

2.2 Tinjauan Studi

2.2.1 Penerapan *Web Scraping* sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis berdasarkan *Keyword* (Flores, V. A. et al., 2020)

- Masalah yang diangkat pada penelitian ini adalah artikel ilmiah dapat dicari melalui internet secara bebas, namun proses pencarian ini sering kali memakan waktu dikarenakan banyaknya artikel ilmiah yang ada di internet.
- Metode yang digunakan dalam penelitian ini menggunakan teknik *Web Scraping* sebagai teknik pengambilan data pada suatu situs *web* dan memanfaatkan metode *Depth First Search* (DFS).
- Hasil uji coba dari penelitian ini adalah sebanyak 134 artikel yang berhasil diunduh menunjukkan bahwa sistem ini dapat bekerja dengan sangat baik untuk melakukan pencarian artikel ilmiah dengan nilai *precision* sebesar 0,87 dan *recall* sebesar 0,99.
- Perbedaan penelitian yang dilakukan dalam skripsi ini adalah penggunaan *web scraping* akan dilakukan pada *Instagram* dan targetnya adalah akun yang memberikan jasa sewa pakaian.

2.2.2 *Social Media Web Scraping using Social Media Developers API and Regex* (Dewi, L. C. et al., 2019)

- Masalah yang diangkat pada penelitian ini bahwa *post* pada media sosial terkadang tidak relevan untuk *user* karena adanya redundansi pada informasi.
- Metode yang digunakan dalam penelitian adalah *Web Scraping* dengan menggunakan *API Facebook* dan *Twitter* yang sudah tersedia, serta menggunakan metode *Regular Expression*.
- Hasil dari penelitian ini adalah bahwa: kelebihan informasi bisa ditekan ke dalam data terstruktur pada *database*, terdapat jumlah minimum informasi redundan yang dapat dikonversikan menjadi analisis *trend*, dan relevansi informasi dapat diubah dengan mengganti *regex format*.
- Perbedaan penelitian yang dilakukan dengan skripsi ini adalah finalisasi dari penelitian skripsi akan berupa *web* dan tidak berhenti di *data base* saja dengan target hanya pada *Instagram*.

2.2.3 *Web Scraping using Natural Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis* (Pichiyan, V. et al., 2023)

- Masalah yang diangkat adalah tantangan yang ditimbulkan oleh pertumbuhan konten digital secara eksponensial di internet, sehingga menyulitkan pengguna untuk menemukan informasi relevan secara efisien.
- Metode yang digunakan dalam penelitian ini adalah *web scraping* dengan *Natural Language Processing* (NLP). Teknik yang digunakan dalam pendekatan gabungan ini untuk mengekstraksi wawasan berharga dari data teks tidak terstruktur.
- Hasil dari penelitian ini adalah peran penting *web scraping* dalam meningkatkan sistem pengambilan informasi dengan memungkinkan ekstraksi data yang efisien dari web yang luas. Metode ini menekankan tingkat akurasi 98,5% dalam menyampaikan informasi secara efektif.
- Perbedaan penelitian dalam pengerjaan skripsi ini adalah penggunaan *web scraping* pada suatu *website* dan dalam skala yang lebih kecil untuk mengumpulkan data yang ada.

Tabel 2.1 Tabel Daftar Pustaka

Nama Penulis	Judul (Tahun)	Masalah	Metode	Hasil
Flores, Permatasari, dan Jasa	Penerapan <i>Web Scraping</i> Sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan	Masalah yang diangkat pada penelitian ini bahwa <i>post</i> pada media sosial terkadang tidak relevan untuk <i>user</i> karena adanya redundansi	Metode yang digunakan dalam penelitian ini menggunakan teknik <i>Web Scraping</i> sebagai teknik pengambilan data pada suatu situs <i>web</i> dan	Hasil dari penelitian ini adalah bahwa: kelebihan informasi bisa ditekan ke dalam data terstruktur pada <i>database</i> , terdapat jumlah minimum

	Keyword (2020)	pada informasi.	memanfaatkan metode <i>Depth First Search</i> (DFS).	informasi <i>redundant</i> yang dapat dikonversikan menjadi analisis <i>trend</i> , dan relevansi informasi dapat diubah dengan mengganti <i>regex format</i> .
Dewi, Meiliana, and Chandra	<i>Social Media Web Scraping using Social Media Developers API and Regex</i> (2019)	Masalah yang diangkat pada penelitian ini bahwa <i>post</i> pada media sosial terkadang tidak relevan untuk <i>user</i> karena adanya redundansi pada informasi.	Metode yang digunakan dalam penelitian adalah <i>Web Scraping</i> dengan menggunakan <i>API Facebook</i> dan <i>Twitter</i> yang sudah tersedia, serta menggunakan metode <i>Regular Expression</i> .	Perbedaan penelitian yang dilakukan dengan skripsi ini adalah finalisasi dari penelitian skripsi akan berupa <i>web</i> dan tidak berhenti di <i>data base</i> saja dengan target hanya pada <i>Instagram</i> .
Vijayaragavan Pichiyan, S.	<i>Web Scraping using Natural</i>	Masalah yang diangkat	Metode yang digunakan	Perbedaan penelitian

Muthulingam, G. Sathar, Sunanda Nalajala, Akhil Ch, Manmath Nath Das	<i>Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis</i>	adalah tantangan yang ditimbulkan oleh pertumbuhan konten digital secara eksponensial di internet, sehingga menyulitkan pengguna untuk menemukan informasi relevan secara efisien	dalam penelitian ini adalah web scraping dengan Natural Language Processing (NLP). Teknik yang digunakan dalam pendekatan gabungan ini untuk mengekstraksi wawasan berharga dari data teks tidak terstruktur.	dalam pengerjaan skripsi ini adalah penggunaan web scraping pada suatu <i>website</i> dan dalam skala yang lebih kecil untuk mengumpulkan data yang ada.
--	---	---	---	--