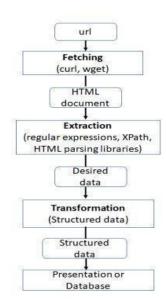
2. LANDASAN TEORI

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

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

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