BAB II

TINJAUAN PUSTAKA

2.1 *Web Scraping*

Menurut Uzun (2020) *Web Extraction* atau yang lebih dikenal dengan w*eb scraping* merupakan sebuah teknik untuk mengekstrak data dari sebuah website dan menyimpannya ke sistem file atau basis data untuk suatu kepentingan. Biasanya, data pada web yang diekstraksi perlu melalui protokol *Hypertext Transfer Protocol (HTTP)* atau melalui *web browser* (Zhao, 2020). Sedangkan menurut Luscombe et al., (2022) *Web scraping* juga dikenal sebagai ekstraksi informasi di internet secara otomatis, yaitu metode penting untuk menghasilkan data untuk mendapatkan pengetahuan. Sebagai contoh, mencari dan menyalin nama dan nomor telepon, atau perusahaan dan URL (Chaib & Salah-ddine, 2021).

Beberapa jenis *web scraping* yang lazim digunakan yaitu *Regular Expression* (Regex), *Hypertext Markup Language Document Object Model* (HTML DOM), *Xpath*, dan *CSS Selector*, bahkan termasuk *copy paste* Darmawan et al., (2022) *CSS Selector* dinyatakan sebagai bagian dari *markup style* yang berlaku untuk mencocokkan tag dan atribut dalam *markup* dan merupakan metode *web scraping* yang dipilih dalam penelitian ini dikarenakan memiliki *upstream* dan *downstream latency* yang paling efisien di antara jenis web scraping yang lain.

2.2 *Bot Chat*

Menurut Poongodi et al., (2019) Chatbot adalah sebuah kode komputer yang berinteraksi dengan manusia secara mandiri tanpa membutuhkan bantuan manusia. Perluasan platform *bot chat* menjadikannya lazim di dunia perpesanan dan jejaring

sosial dan lebih mudah dibuat dan digunakan, namun *bot chat* dengan kecerdasan artifisial masih jarang dan tidak terlalu umum (Tebenkov & Prokhorov, 2021).

2.3 Telegram

Telegram adalah salah satu program obrolan paling populer yang tersedia saat ini Penkov & Taneva (2021). Platform jejaring sosial sumber terbuka berbasis cloud yang disebut Telegram diperkenalkan pada 2013. Penggunanya tidak hanya dapat bertukar informasi melalui pesan, tetapi dapat berupa media atau dokumen (Vasilaras et al., 2022).

Country	Number of Downloads
India	70.48 million
Russia	24.15 million
United States	20.03 million
Indonesia	19.61 million
Brazil	18.04 million
Egypt	11.05 million
Mexico	8.33 million
Ukraine	7.02 million
Vietnam	6.95 million
Turkey	6.48 million
Philippines	6.31 million
Nigeria	5.45 million

Gambar 2.1 10 Negara dengan Pengguna Telegram Terbanyak

Indonesia sebagai negara ke-4 terbanyak (Gambar 2.1) memiliki peluang untuk dapat mengakses ke fitur layanan pesan instan. Dilansir dari laman (Popster, 2023) yang menjadi topik paling populer di Telegram adalah berita (82%), politik dan hiburan (59%), dan edukasi (55%).

Telegram sebagai aplikasi penyedia layanan pesan instan memiliki keunggulan dengan fitur open API untuk mesin bot dan juga akses yang terbuka dan mudah untuk diimplementasikan dengan banyak perangkat Aris Widya & Airlangga (2020). Telegram juga memiliki banyak keunggulan jika dibandingkan aplikasi pesan instan lainnya seperti WhatsApp adalah dukungan Bot, pengeditan pesan terkirim berbasis cloud, grup publik hingga 200.000 anggota, memiliki fitur login dengan banyak perangkat tanpa harus keluar dari perangkat sebelumnya, dan juga fungsi saluran online secara default dengan Telegram, yang berguna untuk mengirimkan informasi ke pelanggan saluran dalam bentuk pesan siaran (Manna & Ghosh, 2018).

Telegram bot tidak membutuhkan nomor tambahan saat pembuatannya. Akun Telegram bot berfungsi hanya sebagai interface dari program atau sistem yang dibuat dalam server, Kusuma (2019). Telegram menjadi objek penelitian ini karena pengguna yang banyak dan kemudahan dalam akses dan juga gratis.

2.4 *Python*

Python memiliki kumpulan pustaka yang lengkap untuk mengekstraksi konten digital yang tersebar di internet. Di antara pustaka-pustaka yang tersedia, tiga pustaka berikut ini banyak digunakan untuk tujuan tersebut: BeautifulSoup, LXml, dan RegEx, Thivaharan et al., (2020). Selain untuk mengekstrak data *HTML* yang mentah untuk berbagai keperluan, *request* pada python dapat digunakan untuk mengirim form dan mengakses *API* (Prakash & Rashid, 2017).

Penggunaan bahasa python untuk keperluan komputasi ilmiah telah mendapatkan momentum dalam beberapa tahun terakhir. Menjadi fakta bahwa

bahasa python memiliki kompleksitas dan mudah dibaca dengan dilengkapi library ilmiah yang menjadi pendukung dalam penerapan karakteristiknya (Tejedor et al., 2017).

Python dapat mengekstrak informasi dari sebuah situs web dan melakukan *parsing* untuk mendapatkan informasi yang dibutuhkan. Data tersebut dikirim ke server web yang dihosting di internet, dan program yang berjalan di server akan mengambil data dari skrip Python (Vikrant et al., 2021). Menurut Chandrika et al., (2020) *BeautifulSoup* merupakan *library* yang efisien untuk mengekstrak atau mem-*parsing* informasi yang berguna dari situs web yang dihosting secara online dalam penelitian ini penulis menggunakan *BeautifulSoup-4* untuk mengambil tag HTML dan mengekstrak data.

2.5 Sci-Hub

Website Sci-Hub memungkinkan pengguna untuk mengunduh versi PDF dari artikel ilmiah, termasuk banyak artikel yang dibatasi aksesnya di situs jurnal Himmelstein et al., (2018). meski begitu Website Sci-Hub dianggap kontroversial, tetapi banyak peneliti dari berbagai belahan dunia (termasuk dari negara dan lembaga yang memiliki akses legal ke makalah penelitian) menggunakan layanan yang disediakannya (Andročec, 2017).

Pendiri Sci-Hub, Alexandra Elbakyan dari Kazakhstan, mengklaim bahwa tujuan utama proyek Sci-Hub adalah untuk menghindari pembatasan hak cipta guna mempercepat perkembangan ilmu pengetahuan, terutama di negara-negara berkembang di mana para peneliti tidak memiliki akses institusional ke pembatasan akses penerbit.

Sci-Hub memberikan akses ke lebih dari 48 juta artikel yang dibatasi aksesnya oleh pembayaran, tanpa memperhatikan hak cipta, Greshake (2023). Fakta bahwa Sci-Hub dianggap sebagai upaya heroik untuk menyelesaikan masalah penelitian yang dibatasi oleh pembayaran mewakili kemajuan yang relatif sedikit dari gerakan Open Access (Priego, 2016).

2.6 Penelitian Terkait

Dalam penelitian ini tentunya berkaitan dengan penelitian-penelitian sebelumnya yang didapatkan ketika melakukan studi literatur melalui jurnal, pencarian di internet dan perpustakaan, serta diskusi dengan dosen pembimbing dan orang-orang yang memiliki pengetahuan yang berkaitan dengan penelitian dan orang yang memiliki pengalaman dalam pembuatan bot chat dan web scraping. Berikut beberapa penelitian sebelumnya yang memiliki keterkaitan dengan penelitian ini dapat dilihat pada tabel 2.1.

Tabel 2.1 Penelitian Terkait

No.	Judul	Penulis	Metode Penelitian	Hasil
1.	Chat Bot sebagai implementasi Pemanfaatan Teknologi Artificial Intelligence dengan Channel Telegram	(Suparno, 2020)	Membuat QnA untuk keperluan sekolah dengan menggunakan Microsoft Azure yang dihubungkan dengan API Telegram	Waktu respon yang cukup cepat membuat waktu tunggu dari penerima pesan menjadi lebih kecil.
2.	Pembuatan Bot Telegram Untuk Mengambil Informasi dan Jadwal Film Menggunakan PHP	Anggiat Cokrojoyo, Justinus Andjarwirawan, Agustinus Noertjahyana	Menggunakan API Telegram dengan Bahasa pemrograman yang digunakan untuk merancang Bot akan menggunakan bahasa Hypertext PreProcessor (PHP)	Bot dapat berfungsi dengan baik di klien aplikasi Telegram baik di ponsel pintar maupun di klien komputer.
3.	*A review of programming languages for web scraping from software repository sites*	(Prakash & Rashid, 2017)	Mereview empat bahasa pemrograman C, Java, PHP dan Python terkait perpustakaan dan metode penggunaanya pada *web scraping* dan *data extraction.*	Python merupakan bahasa terbaik yang dapat digunakan untuk *web scraping.* Hal ini karena adanya modul-modul yang mumpuni seperti *Scrapy, Selenium, Spiders* dan lain-lain.
4.	Implementasi *Web Scraping* dan *Text Mining* untuk Akuisisi dan Kategorisasi Informasi Laman Web Tentang Hidroponik	(A Priyanto & M R Ma'arif, 2018)	Mengumpulkan informasi pada halaman web menggunakan *web scraping* dilanjutkan dengan melakukan pengelompokkan kedalam beberapa kategori menggunakan *text mining*.	Mengotomasi proses akuisisi informasi khususnya informasi-informasi yang bersumber dari artikel atau tulisan bebas di internet dan mengelompokkannya kedalam beberapa kategori.

No.	Judul	Penulis	Metode Penelitian	Hasil
5.	Perbandingan Metode *Web Scraping* Menggunakan *CSS Selector* dan *Xpath Selector*	(Rizaldi & Putranto, 2017)	Membandingkan dua metode *web scraping* dengan parameter pengukuran waktu, penggunaan memori, penggunaan data dan jumlah data	Penggunaan metode *XPATH* untuk *web scraping* situs berita menghasilkan artikel yang lebih lengkap dibandingkan dengan menggunakan metode *CSS Selector*. Metode *XPATH* juga lebih unggul dalam pengukuran waktu
6.	*An Overview On Web Scraping Techniques And Tools*	(Saurkar & Gode, 2018)	Menganalisis teknik *web scraping* dan alat-alatnya	Ada banyak teknik *web scraping* yang bisa digunakan diantaranya *Classical copy and paste*, *Hypertext Transfer Protocol (HTTP) Programming*, *Hyper Text Markup Language (HTML) Parsing*, *Document Object Model (DOM) Parsing*, *Web Scraping Software* dan *Computer vision web-page analysers* dan alat-alat yang digunakan pun beragam diantaranya *Mozenda*, *Visual Web Ripper*, *Web Content Extractor*, *Import.io* dan *Scrapy*
7.	*DECO: Polishing Python Parallel Programming*	(Sherman & Hartog, 2016)	Mengusulkan penyederhanaan teknik pemrograman paralel tradisional yang meminimalkan interaksi programmer dan tidak memerlukan pengetahuan pemrograman paralel	Penyederhanaan teknik pemrograman bersamaan yang ditargetkan pada pemrograman dengan sedikit pemahaman tentang pemrograman bersamaan

No.	Judul	Penulis	Metode Penelitian	Hasil
8.	*Web Scraping and Naïve Bayes Classification for Job Search Engine*	(Slamet et al., 2018)	Penyederhanaan pencarian kerja melalui konstruksi dan kolaborasi teknik pengikisan web dan klasifikasi menggunakan *Naïve Bayes* di mesin pencari	Menghasilkan aplikasi yang efektif dan efisien bagi pengguna untuk mencari pekerjaan potensial yang sesuai dengan minat.
9.	Analisis *Web Scraping* Untuk Data Bencana Alam Dengan Menggunakan Teknik *Breadth-First Search* Terhadap 3 Media Online	(Sonya, 2016)	Melakukan *scraping* pada data yang tidak terstruktur di beberapa media online dan mengklasifikasikannya menjadi data yang terstruktur menggunakan teknik *Breadth-First*	Menghasilkan data yang terstruktur berupa tabel dengan beberapa field yaitu no, hari/tanggal, waktu posting, judul, deskripsi, gambar, dan link halaman artikel
10.	*PyCOMPSs: Parallel computational workflows in Python*	(Tejedor et al., 2017)	Melakukan penelitian terhadap permasalahan umum yang dialami pengguna dalam menggunakan pemrograman paralel pada *Python*	Menyajikan PyCOMPSs, sebuah kerangka kerja yang memfasilitasi pengembangan alur kerja komputasi paralel dalam Python
11.	*Web Scraping*	Zhao (2020)	Menganalisa *web scraping* secara terperinci	Mendeskripsikan *web scraping* secara terperinci
12.	Penggunaan Telegram Bot Pada Telegram Messenger Dengan Metode Web scraping Untuk Sistem Peminjaman Infrastruktur Di Uin Maulana Malik Ibrahim Malang	Afrizal Dwi Kusuma (2019)	Menerapkan *web scraping* pada bot chat telegram untuk sistem informasi peminjaman Infrastruktur	Hasilnya adalah waktu respon rata-rata dari menu dengan panjang 111 adalah 1,22 detik sama dengan menu dengan panjang konten 255

Berbagai penelitian yang sudah dilakukan untuk menjawab hal dasar yang dibutuhkan penelitian ini, beberapa parameter yang akan digunakan pada penelitian ini yaitu, kecepatan respon, kesesuaian data yang diinput dan dihasilkan, jenis data yang dikirim dan diterima. Beberapa penelitian berfokus pada cara merekam kecepatan respon mengulang proses yang dilakukan pengguna.

Terdapat penelitian yang mengutip program yang dapat melakukan hal serupa namun tidak dapat menghasilkan paper secara banyak sekaligus yang mengharuskan pengguna menginputkan DOI artikel atau *paper* secara satu persatu sebelum kemudian menerima file yang dimaksud.

2.7 Tabel *State of the Art*

Beberapa penelitian sudah dilakukan berhubungan dengan penelitian yang saat sedang dilakukan, peneliti sudah memperluas *state of the art* dari unsur – unsur bot chat telegram yang sekarang sudah banyak digunakan pada teknologi mesin, serta dukungan pembaca layar yang mampu mengenali keadaan sekitar. Pada dasarnya penelitian ini berfokus pada teknologi komputer dengan menerapkan metode *web scraping* pada *bot chat* telegram

Selain itu, para peneliti sudah menjawab beberapa pertanyaan yang dibutuhkan untuk memulai penelitian ini, dengan mengeksploitasi kelebihannya untuk menangkap parameter yang dibutuhkan, mengembangkan metode untuk menghasilkan suatu sistem yang menggunakan metode web scraping pada *bot chat* telegram. Hal tersebut dapat dilihat pada matrik penelitian seperti pada tabel 2.2.

Tabel 2.2 Matriks Penelitian

No	Judul	Ruang Lingkup Penelitian								Penulis
		Metode			Objek			Bahasa Pemrograman		
		Artificial Intelligence	Webhook	Web Scraping	Telegram Channel	Telegram Bot	Bot Chat lainnya	Phyton	PHP	
1.	Chat Bot sebagai implementasi Pemanfaatan Teknologi Artificial Intelligence dengan Channel Telegram	V			V			V		Ade Saputra
2.	Pembuatan Bot Telegram Untuk Mengambil Informasi dan Jadwal Film Menggunakan PHP			V					V	Anggiat Cokrojoyo, Justinus Andjarwirawan, Agustinus Noertjahyana
3.	Penggunaan Telegram Bot Pada Telegram Messenger Dengan Metode Webhook Untuk Sistem Peminjaman Infrastruktur Di Uin Maulana Malik Ibrahim Malang		V			V			V	Afrizal Dwi Kusuma