Implementasi *Web Scraping* Pada Media Sosial Instagram

Rio Baskara
Fakultas Teknologi Industri
Universitas Islam Indonesia
Yogyakarta, Indonesia
18523254@students.uii.ac.id

Fayruz Rahma, S.T., M.Eng.
Fakultas Teknologi Industri
Universitas Islam Indonesia
Yogyakarta, Indonesia
175230101@uii.ac.id

***Abstract*—Dinas Komunikasi dan Informatika Daerah Istimewa Yogyakarta (Diskominfo DIY) memiliki sebuah sistem analitik berbasis *big data*, yang difokuskan pada pengembangan data analitik dan pendukung pengambilan keputusan, serta merujuk pada dimensi-dimensi *Jogja Smart Province* (JSP) yang diberi nama Jogja Center. Penelitian ini bertujuan untuk memenuhi kebutuhan data Jogja Center yang menjadi dasar kebutuhan analisis. Pengumpulan data yang akan digunakan menggunakan teknik *web scraping* dengan bahasa pemrograman *Python*. Metode yang akan digunakan terdiri dari *analysis*, *coding*, dan *testing*. Hasil dari penelitian ini berupa data dalam bentuk format dokumen .CSV yang telah dibersihkan. Dapat disimpulkan bahwa dengan menggunakan teknik *web scraping* dapat memudahkan pengguna dalam mengumpulkan data secara cepat.**

Keywords—Web Scraping, Python.

I. PENDAHULUAN

Teknologi adalah penerapan pengetahuan ilmiah dengan tujuan memudahkan kehidupan manusia, menimbulkan perubahan dan manipulasi kehidupan. Seiring dengan berjalannya zaman, perlunya mengembangkan teknologi guna memajukan kehidupan bangsa. Kebutuhan informasi sebagai salah satu alasan, teknologi perlu dikembangkan agar pengolahan data dan informasi lebih mudah dilakukan, mempermudah untuk mendapatkan suatu informasi yang dibutuhkan. *Web Scraping* adalah salah satu teknik pengambilan data, umumnya berupa halaman-halaman web yang terdapat di internet.

Web scraping bukanlah bagian dari *data mining* karena *web scraping* terfokus pada cara memperoleh data melalui pengambilan data, sedangkan *data mining* berupaya memahami tren atau pola dari data yang telah diperoleh.

Dengan *Web Scraping*, data yang dikumpulkan akan terpusat kepada tujuan yang telah ditentukan sehingga akan memudahkan dalam proses pencarian. Menciptakan sebuah program yang dapat mempelajari dokumen dengan bahasa pemrograman HTML pada situs yang dituju merupakan salah satu cara untuk mengembangkan teknik *web scraping*.

II. LANDASAN TEORI

A. *Web Scraping*

Web scraping merupakan sebuah teknik yang digunakan untuk mengumpulkan data secara manual yaitu dengan melakukan *copy paste* data yang diinginkan maupun secara otomatis yaitu dengan membuat sebuah program atau kode yang dapat melakukan proses pengambilan data dari sebuah halaman web [1]. Dalam melakukan *web scraping* terdapat beberapa metode yang dapat dilakukan yaitu:

1. Menyalin data secara manual
2. *Regular Expression*
3. *Parsing HTML*
4. Analisa *Document Object Model* (DOM)

Tidak dipungkiri teknik *web scraping* memiliki kekurangan yaitu sampai saat ini belum ada teknik web scraping yang 100% efektif. Selain itu, hasil yang didapatkan tidak selalu rapi, maka perlu juga untuk memahami struktur halaman *website* yang dituju. Karena tidak semua data dapat diekstrak dengan mudah, sering kali program harus dijalankan berulang kali yang mengakibatkan akses terhadap halaman web tersebut terblokir. Namun, ada pula manfaat dari melakukan *web scraping* yaitu data-data yang didapatkan akan lebih terfokus yang dapat memudahkan dalam pencarian sesuatu.

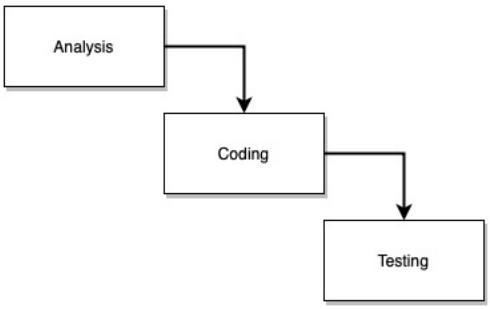
B. *Extract, Transform, Load* (ETL)

Extract, transform, load atau dapat disingkat menjadi ETL merupakan prosedur umum dalam menyalin data dari suatu sumber ke sistem yang dituju [2]. Proses ini dibagi menjadi tiga tahap yaitu:

1. *Extract*: Ekstraksi merupakan proses paling penting karena dapat menjadi sebuah acuan keberhasilan tahap selanjutnya. Sebagian besar dari proyek *database* menggabungkan data dari sumber yang berbeda. Setiap sumber juga dapat menggunakan format data yang berbeda. Format data pada umumnya berbentuk JSON, namun ada juga yang memiliki format XML.
2. *Transform*: Pada tahap transformasi data disiapkan untuk dimuat pada target akhir dengan melakukan serangkaian fungsi pada data yang telah diekstrak. Proses *transform* ini berfungsi untuk membersihkan data atau memisahkan data-data yang tepat untuk digunakan. Namun terdapat tantangan ketika sistem yang berbeda melakukan interaksi yaitu antarmuka dan komunikasi sistem yang relevan. Kumpulan karakter yang mungkin tersedia di satu sistem mungkin tidak tersedia di sistem lainnya.
3. *Load*: Tahap *load* merupakan tahap dimana data dimasukan ke target akhir. Proses *load* dapat dibagi menjadi menjadi dua, yaitu *full load* dan *incremental load*. *Full load* merupakan metode untuk memuat seluruh data secara bersamaan untuk menjadi catatan baru pada *database*. Metode ini berguna untuk menghasilkan data yang tumbuh secara eksponensial namun sulit untuk diatur. Sedangkan incremental *load* merupakan metode untuk memuat data secara *interval* terjadwal. Karena *incremental load* membandingkan data yang masuk dengan data yang sudah ada, metode ini menghasilkan data tambahan

jika ditemukan data yang baru dan memudahkan untuk diatur.

III. METODOLOGI



Gambar 1. *Waterfall Model*

Pada **Gambar 1** diperlihatkan metode yang digunakan adalah *classic life cycle* atau *waterfall* yang terdiri dari: 1) *Analysis*, 2) *Coding*, dan 3) *Testing* [3].

A. *Analisis*

Pada tahap awal yaitu analisis yang merupakan tahap dimana dilakukan perencanaan data apa yang nantinya akan dikumpulkan. Penulis melakukan koordinasi dengan anggota tim lainnya perihal data apa yang dibutuhkan.

B. *Pembuatan Kode*

Dengan bahasa pemrograman Python, rancangan yang telah dibuat akan diterjemahkan dalam bahasa pemrograman tersebut. Kode yang telah dibuat berfungsi untuk menjalankan aplikasi agar berjalan sesuai dengan yang telah direncanakan. Aplikasi tersebut hanya dapat dijalankan terhadap akun yang bersifat publik dengan metode *parsing HTML*.

```
results = api.user_feed('3535868226')
next_max_id = results.get('next_max_id')
hasil = json.dumps(results, indent = 4)
# print(hasil)

while next_max_id:
    results = api.user_feed('3535868226', max_id = next_max_id)
    next_max_id = results.get('next_max_id')

with open('data_new.json', 'w') as outfile:
    json.dump(results, outfile, indent = 4)
```

Gambar 2. *Potongan Kode Scraping*

C. *Pengujian*

Akan dilakukan pengujian atau *testing* pada aplikasi *web scraping* yang telah dibuat sebelumnya terhadap akun Instagram dengan *username* “rio_baskara” karena akun tersebut bersifat publik. Hal ini bertujuan untuk mencari kesalahan dalam kode atau logika pemrograman.

```
(base) riobaskara@Rios-MacBook-Air private_api % python scrap.py
Done
```

Gambar 3. *Hasil Proses Pengujian*

Setelah melewati tahap pengujian, didapatilah sebuah aplikasi yang memberikan *output* yaitu hasil pencarian yang kemudian hasil pencarian tersebut akan disimpan.

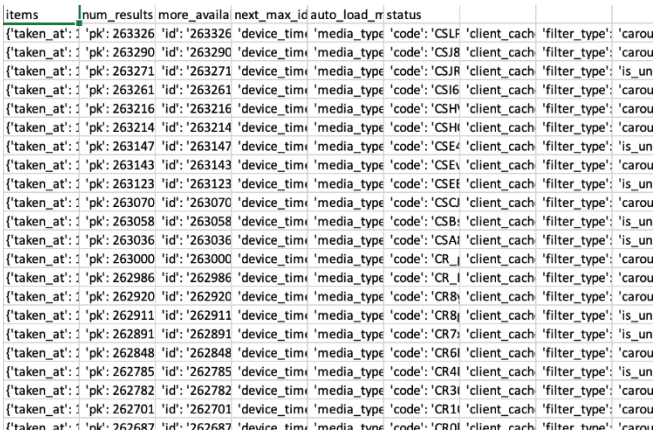
IV. HASIL DAN PEMBAHASAN

A. *Hasil Web Scraping*

Aplikasi dijalankan terhadap akun Instagram “kominfodiy”. Data yang telah di-*scrape* umumnya masih dalam bentuk data Json seperti pada **Gambar 4**, yang kemudian data tersebut dikonversi ke dalam bentuk CSV seperti pada **Gambar 5**. Proses ini dilakukan menggunakan kode dengan tujuan untuk memudahkan ketika nantinya data tersebut akan dilanjutkan ke tahap *data cleaning*.


```
"taken_at": 1470794060,
"pk": 1313438232372944762,
"id": "1313438232372944762_3535868226",
"device_timestamp": 1470793995893,
"media_type": 1,
"code": "BI6RR0LAZ96",
"client_cache_key": "MTMxMzQzODIzMjM3Mjk0NDc2Mg==.2",
"filter_type": 0,
"is_unified_video": false,
"should_request_ads": false,
"user": {
    "pk": 3535868226,
    "username": "kominfodiy",
    "full_name": "#jogjaistimewa | Kominfo DIY",
    "is_private": false,
    "profile_pic_url": "https://instagram.fjog3-1.fna.fbcdn.net/v/t51.2885-19/s150
    "profile_pic_id": "2585479117540205404_3535868226",
    "is_verified": true,
    "follow_friction_type": 0,
    "has_anonymous_profile_picture": false,
    "is_unpublished": false,
    "is_favorite": false,
    "account_badges": [],
    "latest_reel_media": 1630574084
},
"can_viewer_reshare": true,
"caption_is_edited": false,
"like_and_view_counts_disabled": false,
"is_commercial": true,
"commerciality_status": "commercial",
"is_paid_partnership": false,
"playlist_eligibility": false,
"comment_likes_enabled": true,
```


Gambar 4. *Hasil dalam Bentuk Json*



Gambar 5. *Hasil Yang Telah Dikonversi Menjadi CSV*

B. *Hasil Parsing*

Setelah data didapatkan, data tersebut dibersihkan terlebih dahulu. Proses pembersihan data dilakukan secara manual oleh pengguna. Langkah yang dilakukan yaitu data yang tidak relevan atau tidak dibutuhkan dibersihkan. Kemudian mengedit kesalahan pada struktur seperti adanya kesalahan penulisan, kesalahan penggunaan huruf besar dan kecil, dan lain-lain. Hal-hal tersebut dilakukan untuk memudahkan dalam pengolahan data. Hasil *parsing* kemudian disimpan ke dalam bentuk *file* CSV seperti yang ditampilkan pada **Gambar 6** dan diberikan kepada anggota tim lain untuk dianalisis.

datetime	timestamp	image_link	caption	link	media_id	comment
05/08/21	1628129892	https://instagrar	Hai Lur!\nKab	https://www.	2633266855686853562_3535868226	
04/08/21	1628086393	https://instagrar	Hasil laporan	https://www.	2632901954711	Total jumlah keseluruha
04/08/21	1628063733	https://instagrar	ÄúSaya sudah	https://www.	2632711871105217551_3535868226	
04/08/21	1628051653	https://instagrar	Berkaitan den	https://www.	2632610541166	Yang sumbangan 2T, te
03/08/21	1627998671	https://instagrar	Rekapitulasi P	https://www.	2632166088460	Mekanisme nya lewat
03/08/21	1627996397	https://instagrar	Hasil laporan	https://www.	2632147015559	Total jumlah keseluruha
02/08/21	1627916220	https://instagrar	Hai Lur! Menj	https://www.	2631474440667690661_3535868226	
02/08/21	1627911833	https://instagrar	Hasil laporan	https://www.	2631437642981	Terimakasih atas Infony
02/08/21	1627887588	https://instagrar	Hai Lur! \nMa	https://www.	2631234017189875923_3535868226	
01/08/21	1627824884	https://instagrar	Hasil laporan	https://www.	2630708262777	Yok Jogja masih tinggi y
01/08/21	1627809711	https://instagrar	Pemda DIY lak	https://www.	2630580978334312085_3535868226	
01/08/21	1627784612	https://instagrar	Hai Lur!\n\nK	https://www.	2630369108461722592_3535868226	
31/07/21	1627740685	https://instagrar	Hasil laporan	https://www.	2630001952091809450_3535868226	
31/07/21	1627723789	https://instagrar	Berkaitan den	https://www.	2629860217718142644_3535868226	
30/07/21	1627645103	https://instagrar	Hasil laporan	https://www.	2629200148932455629_3535868226	
30/07/21	1627635341	https://instagrar	Gubernur DIY	https://www.	2629118259894690767_3535868226	
30/07/21	1627610659	https://instagrar	Hai Lur!\n\nN	https://www.	2628910891072300586_3535868226	
29/07/21	1627559657	https://instagrar	Hasil laporan	https://www.	2628483375625587372_3535868226	
28/07/21	1627484781	https://instagrar	Halo Sahabat	https://www.	2627855267964750978_3535868226	

***Gambar 6**. Hasil Setelah Parsing*

V. KESIMPULAN

Berdasarkan implementasi *web scraping* yang telah dilakukan, dapat disimpulkan bahwa:

1. *Web scraping* merupakan suatu teknik yang sangat bermanfaat untuk mengumpulkan data secara cepat.
2. *Web scraping* merupakan suatu hal yang sah untuk dilakukan selama tidak disalahgunakan seperti pencurian data, dan lain-lain.
3. Selanjutnya penelitian ini dapat dikembangkan dengan mengaplikasikannya pada media sosial lain dengan menggunakan bahasa pemrograman lainnya.

VI. DAFTAR PUSTAKA

[1] Anand V., Kedar G., Schweta A. An Overview On Web ScrapingTechniques And Tools. IJFRCSCE. 2018; 4(4): 363-367.

[2] Srividya K., Sebastian K. Integrating Big Data: A Semantic Extract-Transform-Load Framework. IEEE. 2015; 48(3): 42-50.

[3] Wijaya Y., Astuti M. Sistem Informasi Penjualan Tiket Wisata Berbasis *Web* Menggunakan Metode *Waterfall*. e-ISSN. 2019; 1(2): 273-276.