Implementasi *Web Scraping* Pada Situs Berita Menggunakan Metode *Supervised learning*

Edwin Hari Agus Prastyo
Program Studi S1 Teknik Informatika, Fakultas Teknologi Informasi, Universitas Hasyim Asy'ari
Email : edwinhap08@gmail.com
IGL Putra Eka Prismana
Program Studi S1 Teknik Informatika, Fakultas Teknologi Informasi, Universitas Hasyim Asy'ari
Email : ekaprismana@gmail.com
Radityo Wiratsongko
Program Studi S1Sistem Informasi, Fakultas Teknologi Informasi, Universitas Hasyim Asy'ari
Email : wiratsongko@gmail.com

Abstrak

Negara indonesia adalah salah satu pengguna internet tertinggi didunia termasuk dalam penetrasi informasi di internet media berita online. Namun pada umumnya situs berita tidak hanya menampilkan informasi berita saja, Kebanyakan situs juga menampilkan informasi-lain seperti iklan dan juga bentuk navigasi yang menggangu pembaca situs berita serta menggangu kenyamanan pembaca, dari permasalahan tersebut penelitian ini bertujuan dapat menerapkan teknik *web scraping* dengan metode *supervised learning* dan menganalisa bentuk DOM *tree* dan XPath situs berita. Metode pendekatan *supervised learning* adalah metode yang digunakan dalam penelitian ini, yang merupakan salah satu metode *machine learning*. Dengan digabungkannya teknik web scraping ini dengan pembelajaran *supervised learning* bertujuan agar dapat mengimplementasikan dan mengoptimalkan teknik *web scraping* untuk mengumpulkan informasi berita dari berbagai situs. Untuk melakukan *web scraping* dasar nya yaitu mengetahui pola DOM, sturuktur *XPath* sebagai data model atau *selector* di setiap situs. Hasil penelitian berupa aplikasi web *scrap* yang dapat mengambil konten situs berita tanpa copy paste dan data tersebut disimpan dalam *database* dan ditampilkan ke bentuk aplikasi user buat pembaca tanpa adanya iklan dan navigasi yang mengganggu pembaca.

Kata Kunci: *web scraping*, python, *supervised learning*, *XPath*, DOM *tree*.

Abstract

Indonesia is one of the highest internet users in the world, including in the penetration of information on the internet, online news media. But in general news sites not only display news information, but Most sites also display other information such as advertisements and also forms of navigation that interfere with news site readers and interfere with reader comfort, from these problems this study aims to implement web scraping techniques with supervised learning methods and analyzing the form of DOM tree and XPath news sites. The supervised learning approach method is the method used in this study, which is one of the methods of machine learning. By combining these web scraping techniques with supervised learning, the aim is to be able to implement and optimize web scraping techniques to gather news information from various sites. To do basic web scraping that is knowing DOM patterns, XPath structure as a data model or selector at each site. The results of research in the form of a web scrap application that can retrieve news site content without copy paste and the data is stored in a database and displayed to the user application form for the reader without any ads and navigation that disturb the reader.

Keywords: web scraping, supervised learning, XPath, DOM *tree*.

Implementasi *Teknik Web Scraping* Pada Situs Berita Menggunakan Metode *Supervised learning*

PENDAHULUAN

Negara indonesia adalah salah satu pengguna internet tertinggi didunia. Berdasarkan badan survai infografis penestrasi & perilaku Pengguna internet indonesia mencapai 68% dari 264,18 jt jiwa penduduk indonesia (APPJI, 2018), banyaknya pengguna melakukan penetrasi informasi melalui internet.

Informasi merupakan hal yang sangat terpenting dalam kehidupan, sumber informasi dapat melalui situs berita. Situs berita adalah sebuah *web*site menyediakan berbagai informasi dari berbagai sumber berita untuk ditampilkan kepada pengguna**.**

Namun pada umumnya situs tidak hanya menampilkan informasi dari sumber luar, kadang mereka juga menampilkan informasi-informasi dalam situs *web*site mereka sendiri seperti iklan dan navigasi yang menggangu pembaca situs berita serta menggangu kenyamanan pembaca, dan juga sebagian orang mungkin masih mengumpulkan data yang ada di*web*site dengan menyalin satu persatu yang ada di*web*site, namun jika *website* yang anda kelola adalah situs berukuran besar dengan ribuan data, tentu perkerjaan tersebut memakan waktu yang lama, Jadi untuk mengatasi masalah tersebut peneliti melakukan salah satu teknik extraksi konten *web* menggunakan *machine learning.*

Dalam Klasifikasi pada *supervised learning* dilakukan dengan melakukan *training* (pelatihan) untuk membentuk model. *Classifier* (algoritma klasifikasi) akan membentuk model yang beradaptasi sesuai dengan fitur-fitur yang ada pada data. Tujuan dalam menerapkan metode *supervised learning* agar dapat mengimplementasikan dan mengoptimalkan teknik *web scraping*, serta mempermudah penguna untuk mengolah informasi dari hasil teknik *web scraping* digunakan untuk mengumpulkan informasi berita dari berbagai situs. Untuk melakukan *web scraping* dasar nya yaitu mengetahui sturuktur DOM *tree* pada situs berita.

Peneliti menganalisa pola DOM *tree* dan Xpath nantinya digunakan untuk bahan dalam data latih. DOM *tree* merupakan standar stuktur dokumen html untuk membentuk sebuah halaman *web*site, sedangakan *Xpath* adalah bahasa *query* dari representasi dari DOM mengahasilkan sebuah *node-node* tertentu. Penelitian rizaldi menyatakan bahwa pengunaan Xpath lebih baik dibandingkan CSS *selector* maka dari itu penelitian ini memakai Xpath.

Dari penelitian hasil penelitian rizaldi. Maka Dalam penelitian ini mengambil judul Implementasi *Web Scraping* Pada Situs Berita Menggunakan Metode *Supervised Learning*.dibawah ini beberapa pengertian yang berhubungan dengan penelitian ini.

A. *Web scraping*

Web scraping **(**pengikisan *web*) adalah praktik mengumpulkan data melalui manusia menggunakan *web browser*, untuk mengatasi *web* yang tidak menyediakan API, *Web scraping* meminta data dalam tag HTML dan kemudian mem-parsing data tersebut untuk mengekstrak yang diperlukan informasi. Dalam penerapanya, pengikisan *web* mencakup beragam teknik pemrograman dan teknologi, seperti analisis data dan keamanan informasi. Pengikisan *web* Mengambil data HTML dari nama domain, Parsing data itu untuk informasi target, Menyimpan informasi target secara opsional, dan pindah ke halaman lain untuk mengulangi proses *scraping* (Mitchell, 2015)

B. *Content Extraction using Machine learning* (Ekstraksi Konten menggunakan Pembelajaran mesin)

Ekstraksi Konten menggunakan pembelajaran Mesin Pembelajaran mesin yang merupakan sebuah cabang kecerdasan buatan adalah tentang konstruksi dan studi sistem yang dapat belajar dari data *training*, inti dari pembelajaran mesin berkaitan dengan representasi dan generalisasi. Generalisasi adalah properti yang sistem akan bekerja dengan baik pada sebuah data yang tidak terlihat (S. Ajoudanian and M. Jazi, 2009).

C. *Supervised Learning*

Pada penelitian Dougherty pengertian *supervised learning* merupakan salah satu metode untuk mengklasifikasikan masing masing objek dalam data ke beberapa kelas. Pada *supervised learning* setiap objek pada suatu data memiliki fitur, yaitu ciri-ciri yang ada pada masing-masing objek. Setiap objek dalam suatu data memiliki jumlah fitur yang sama. Fitur digunakan sebagai input untuk menentukan kelas pada objek. Dalam *supervised learning*, kelas dari masing-masing objek sudah diketahui. Oleh karena itu, permasalahan yang dihadapi dalam *supervised learning* adalah bagaimana memetakan objek ke dalam kelas yang tepat menggunakan fitur-fitur yang dimiliki oleh setiap objek.

D. *Python*

Python adalah salah satu bahasa pemrograman berbasis dekstop atau *web*, Bahasa pemrograman saat ini populer digunakan oleh banyak pengembang. Python salah satu bahasa pemrograman populer yang digunakan oleh banyak pengembang. Survei menurut situs bahasa pemrograman versi www.tiobe.com, Python berada diperingkat ke-4 pada tahun 2019. Python juga dapat dipakai untuk *enterprise*. Pyhton masuk Dalam tingkatan bahasa pemrograman, termasuk *high level language*. Python adalah salah satu bahasa pemrograman yang dapat digunakan untuk membangun aplikasi, baik itu berbasis *desktop*, *web* atau berbasis *mobile.*

METODE

Bab ini menjelaskan tentang gambaran tahapan untuk menjawab perumusan masalah sehingga dapat mencapai tujuan dari penelitian.

3.1. Objek Penelitian

Penelitian yang akan dilakukan ini, objek penelitiannya adalah sebuah situs berita berjumlah 100 laman situs berita. Daftar situs berita yang di ekstraksi konten *web* pada tabel 3.1.

3.1 Tabel daftar situs berita

No	Nama situs	Url
1	Media Indonesia	https://mediaindonesia.com
2	Kompas	https://www.kompas.com
3	Bisnis Indonesia	https://teknologi.bisnis.com
4	Pikiran Rakyat	https://www.pikiran-rakyat.com
5	Cek & Ricek	https://ceknricek.com
6	Siwalima	https://siwalimanews.com
7	Waspada	http://waspada.co.id
8	Analisa	https://analisadaily.com
9	Tribun Timur	https://makassar.tribunnews.com
10	Kedaulatan Rakyat	https://krjogja.com
11	Harian Jogja	https://jogjapolitan.harianjogja.com
12	Suara Merdeka	https://www.suaramerdeka.com
13	Solo Pos	https://www.solopos.com
14	Koran Sindo	https://jateng.sindonews.com
15	Sindo Weekly	https://sumeks.co
16	Sumatera Ekspres	http://www.sindoweekly.com
17	Radar Palembang	http://www.radar-palembang.com
18	Tribul Sumsel	https://sumsel.tribunnews.com
19	Palempang Ekspres	https://palembang.tribunnews.com
20	Republika	https://republika.co.id

Tabel lanjutan 3.1 daftar situs berita

No	Nama situs	Url
21	Antara	https://www.antaranews.com
22	Okezone	https://www.okezone.com/
23	merdeka	https://www.merdeka.com/
24	detik	https://www.detik.com/
25	liputan	https://www.liputan6.com/

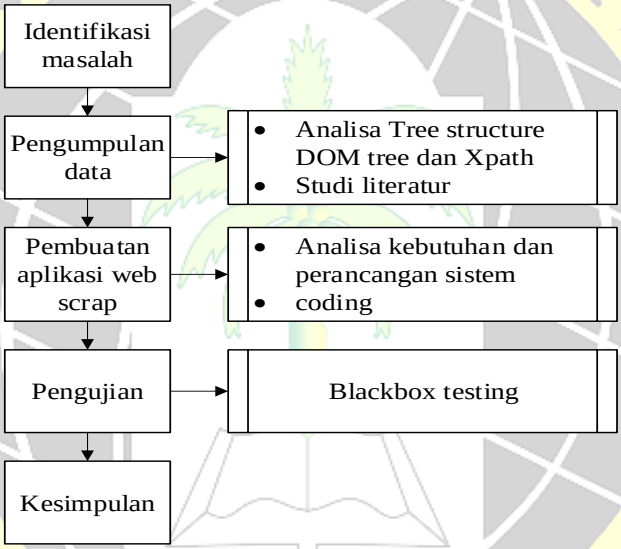
3.2. Variabel Penelitian

Dari semua list situs didalam objek penelitian, diambil hanya data tertentu yang akan diekstraksi dan disimpan ke dalam beberapa variabel, Variabel yang pertama adalah link yang berisi tautan untuk menuju halaman *web* yang memuat berita. kedua adalah judul berita, yang berisi judul berita, berikutnya adalah isi konten berita, dan Keempat variabel XPath *Selector*.

3.3. Jenis Penelitian

Penelitian ini jenis penelitian dan pengembangan. Penelitian ini bertujuan untuk membuat aplikasi yang menggunakan teknik *web scraping* unutk mengambil data secara otomatis dalam laman berita berdasarkan *tag* yang disimpan di basis data menggunakan pendekatan *supervised learning*. Peneitian ini menghasilkan berupa model berupa struktur *tree* dari dokumen html setiap *web*site yang di *scrap*.

3.4. Prosedur penelitian

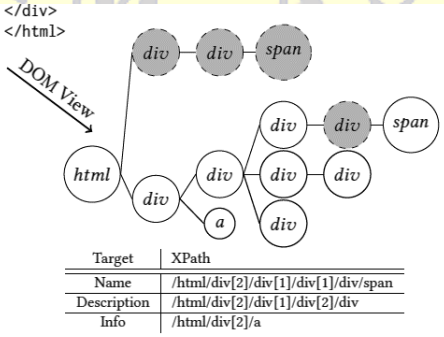


Gambar 3.1 Prosedur penelitian

3.5. Metode Pengumpulan Data

3.5.1. Analisa *Tree structure* DOM *tree* dan Xpath

Dalam analisa ini klasifikasi *tag* secara manual dan setelah mengamati struktur html dan *tag*-nya, Sebuah contoh dari sebuah dokumen HTML dimodelkan sebagai pohon ditunjukkan dalam Gambar 3.2.



Gambar 3.2 dokumen HMTL dan model *tree*

3.6. Analisa Kebutuhan Dan Perancangan Sistem

3.6.1. Analisa kebutuhan

Setelah mencari studi literatur peneliti melakukan analisa kebutuhan fungsional dan non fungsional.

A. Kebutuhan fungsional

1. Sistem ini dapat mengekstrak laman *web* untuk diambil *tag* judul dan *tag* konten berita, disimpan kedalam *database*.
2. Sistem dapat mengklasifikasikan *tag XPath* konten berita.
3. Sistem dapat melakukan ekstraksi fitur.
4. Sistem dapat mengelola daftar situs yang akan di ekstrak.
5. Sistem dapat menyimpan hasil extraksi konten dalam *database* sql.
6. Sistem dapat mengelola user aplikasi *scraping.*
7. Sistem dapat mengimplementasikan sesuai metode *supervised learning.*
8. Sistem memunculkan jumlah konten yang di extrak.
9. Sistem dalam pada data latih dapat menyimpan url, *tag XPath selector*.
10. Sistem dalam pada data uji dapat melakukan pembacaan model dari data latih.

B. Kebutuhan Non-Fungsional

1. Perangkat Keras (*Hardware*)

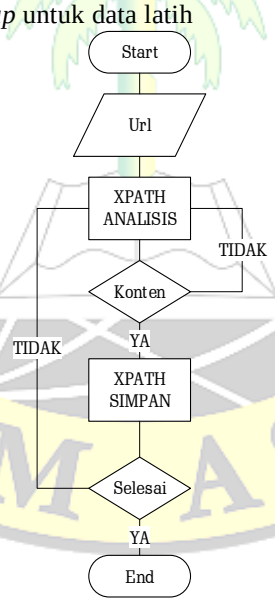
   Spesifikasi minimum yang dibutuhkan agar rekayasa perangkat lunak dapat dijalankan dengan maksimal adalah sebagai berikut:

   - Laptop/Komputer processor minimal AMDA A9 dan RAM minimal 4 GB.
   - Kapasitas Hardisk minimal 1 tera - Modem/Wifi.

2. Perangkat Lunak (*Software*)
   - *Text editor* menggunakan *visual studio code* dan *sublime text*.
   - Menggunakan bahasa pemrograman python.
   - Sql *database*

3.6.2. Perancangan Sistem

1. *Flowcart* sistem *web scrap* untuk data latih

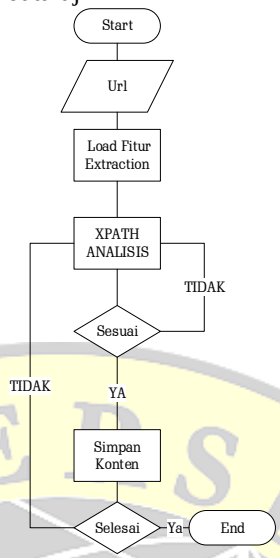


Gambar 3.4 *flowacart* untuk data latih

Penjelasan :

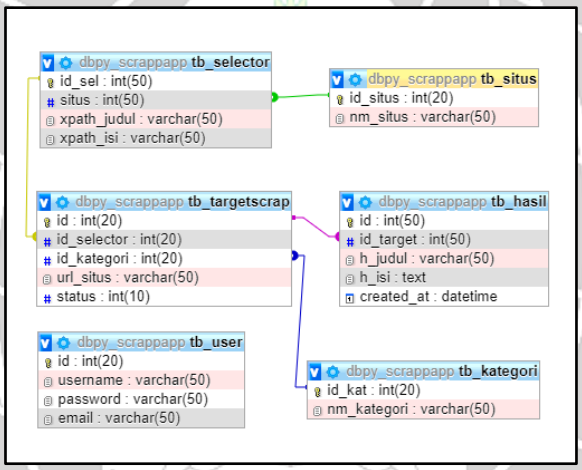
Alur sistem data latih ini dimulai dari start, sistem mendapat inputan URL situs berita, dianalisa menggunakan metode xpath, dari hasil anlisa xpath, diberi keputusan apakah ini xpath konten, jika iya maka simpan xpath dalam *database*, dan sistem ini selesai.

2. *Flowcart* sistem *web scrap* untuk data uji



Gambar 3.5 *flowacart* untuk data uji

3.6.3. Perancangan database



Gambar 3. 6 perancangan database

Gambar 3.7 menjelaskan ada tabel user, situs, kategori, *selector*, target scrap, dan hasil. Pertama membuat tabel user, lalu tabel kategori, tabel situs, tabel *selector* relasi dengan tabel kategori dan situs. Kedua tabel target *scrap* relasi dengan tabel *selector* dan tabel hasil.

HASIL DAN PEMBAHASAN

4.1.Hasil

4.1.1. Hasil Penelitian

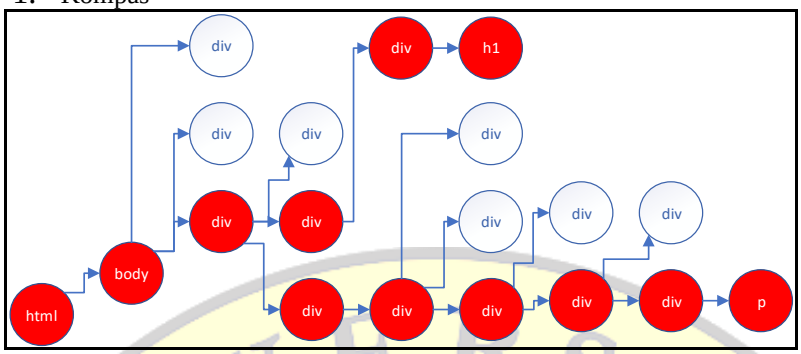
Hasil penelitian pada implementasi teknik web scraping pada situs berita menggunakan metode *supervised learning* adalah sebagai berikut:

a. Mengetahui pola DOM *tree* dan *XPath* yang terdapat dalam sebuah situs berita.
b. Admin dapat mengelola situs berita dan mengelola kategori berita.
c. Admin dapat mengelola model *selector XPath* situs berita.
d. Admin dapat mengelola target laman situs berita yang akan discraping atau diekstrak bagian judul dan isi.
e. User dapat membaca hasil *scraping* berisi judul dan konten berita dalam aplikasi web mobile *newsscrapp*.
f. User dapat membaca berita sesuai kategori yang disediakan dan mencari berita sesuai judul.

4.1.2. Hasil analisa struktur DOM tree dan XPath situs

Berikut analisa situs DOM sesuai daftar situs dalam bab 3 dan hasil analisa *XPath* judul dan isi pada tabel 3.1 Tabel daftar situs berita.

1. Kompas



Gambar 4. 1 Dom tree kompas

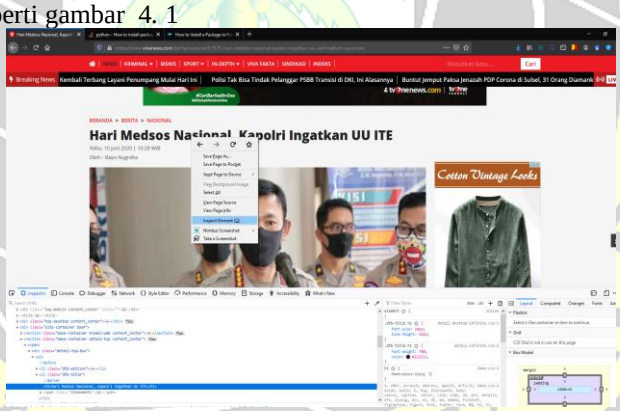
Tabel 4. 1 XPath Kompas

No	Target	XPath
1	Judul	/html/body/div[3]/div[2]/div/h1
2	Isi	/html/body/div[3]/div[3]/div[1]/div[3]/div[2]/div[2]/p

4.2. Pembahasan

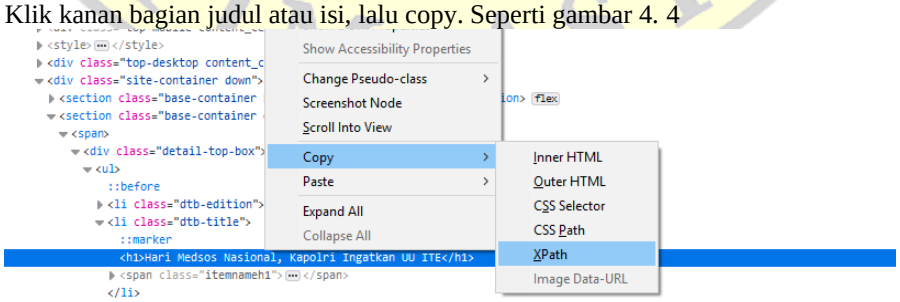
4.2.1. Langkah melakukan pengambilan *XPath* situs

1. Pertama buka browser, masuk kan link situs yang akan ditambahkan dalam aplikasi web scrap ini
2. Lalu klik kanan pada laman browser, pilih bagian judul atau isi.
   Akan muncul seperti gambar 4. 1



Gambar 4. 1 klik kanan laman browser

3. Klik kanan bagian judul atau isi, lalu copy. Seperti gambar 4. 4

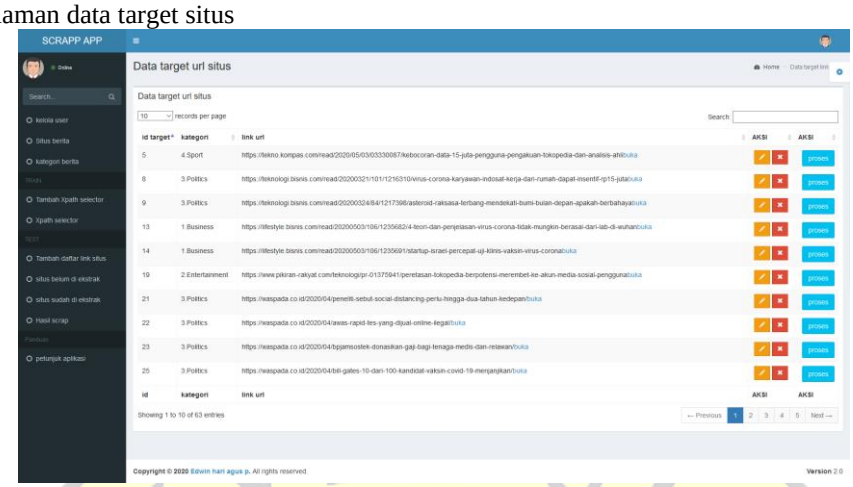


Gambar 4. 2 mengcopy *XPath* situs

4. Copy dan simpan dalam menu *XPath selector* aplikasi web scraping ini.

4.2.2. Implementasi sistem dan hasil proses program

1. Halaman data target situs

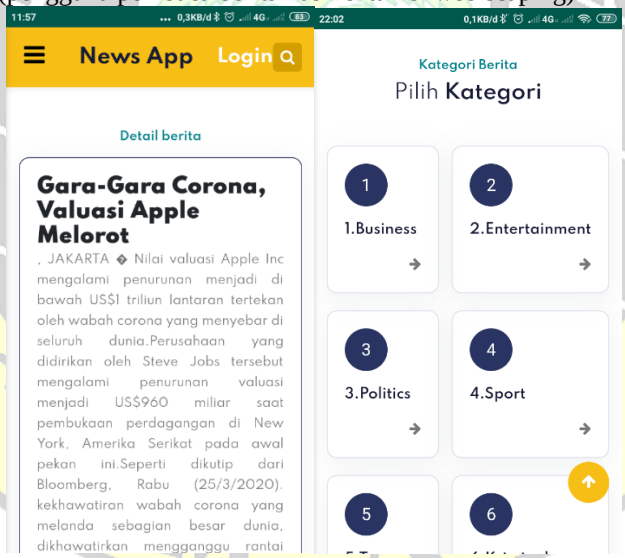


Gambar 4. 3 target situs

Penjelasan dari gambar 4. 3:

Halaman data target ini berisi daftar link situs yang akan discrap atau di ambil konten beritanya melalui *XPath selector* (data latih) yang sudah disediakan sebelumnya. Admin melakukan klik button proses akan melakukan straksi fitur html dan menganalisa *XPath* dicocokkan dengan *XPath selector* atau model *XPath* sebelumnya yang sudah disediakan jika sama lalu akan disimpan konten judul dan isi beritanya jika model *XPath* selector tidak sama dengan situs berita yang akan dilakukan scraping data makan tidak bisa mengambil konten berita tersebut dan sistem akan menampilkan erorr.

2. Halaman user (pengguna pembaca berita hasil ektarksi web scaping)



Gambar 4. 4 membaca berita dan memeilih kategori

Penjelasan dari gambar 4. 4:

Halaman user yang ini menampilkan semua isi berita yang dipilih dengan menekan tombol baca lanjut dan halam pencarian berita berdasarkan kategori berita.

PENUTUP

Simpulan

Berdasarkan hasil dan pembahasan sebelumnya, hasil penelitian ini berupa analisa pola DOM tree dan *XPath* dari situs berita yang diteliti, dan diimplementasikan berupa aplikasi web scrap news berbasis web mengunakan bahasa pemrograman python dengan *framework* flask. Penulis dapat menarik kesimpulan sebagai berikut:

1. Mengetahui pola DOM dan *XPath* situs berita. Mengetahui pola *XPath* yang berbeda dalam sebuah situs berita.

2. Hasil data ekstaksi menggunakan *XPath* yang dikumpulkan lebih lengkap. Dengan menggunakan metode pembelajaran *supervised learning* ini semakin banyak data model latih *XPath* semakin bagus aplikasi ini. Jadi ketika melakukan web scraping tidak menganalisa situs lagi karena sudah ada di data model *XPath*.

Saran

Dari hasil dan pembahasan, serta kesimpulan diatas mempunyai beberapa saran untuk penelitian selanjutnya. Sebagai berikut:

1. Pola *XPath* sebuah situs dalam penelitian ini masih dibentuk secara manual oleh peneliti, untuk penelitian berikutnya lebih baik jika aplikasi dapat membentuk pola *XPath* automatic berdasarkan situs yang diakses.
2. Agar hasil ekstraksi data situs berita dapat dimanfaatkan dengan baik, diperlukan untuk proses penelitian *text mining*, *similar* dan data mining.

DAFTAR PUSTAKA

APJII. 2020, januari 28. *infografis penestrasi & perilaku pengguna internet indonesia*. Retrieved from Asosiasi penyelengara jasa internet indonesia: https://www.apjii.or.id/content/read/39/410/Hasil-Survei-Penetrasi-dan-Perilaku-Pengguna-Internet-Indonesia-2018.

Mitchell, R. 2015. *Web Scraping with Python.* United States of America: O'Reilly Media.

S. Ajoudanian and M. Jazi, 2009. Deep Web Content Mining. *World Academy of Science, Engineering and Technology*

Taufiq Rizaldi, H. A, 2017. Pemanfaatan News Crawling Untuk Pembangunan. *Seminar Nasional Hasil Penelitian* , Hal. 291-295.

Rizaldi, T. d. (2017). Perbandingan Metode Web Scraping Menggunakan CSS Selector dan XPath Selector. TEKNIKA, 6, 43-45.

Haddaway, N. R. (2015). The Use of Web-Scraping Software in Searching for Grey Literature. *Grey J11*, 1, 186–90.

Alshomrani, S. M.-G. (27 January 2015). A Comprehensive Survey on Web Content Extraction Algorithms and Techniques. Jeddah, Saudi Arabia: Faculty of Computing and Information Technology, King Abdulaziz University.