# Pembasan

Pembahasan mengenai pembuatan engine yang sudah dikerjakan akan dibagi menjadi 3 bagian yaitu proses Scraping Data, Text Summarization, dan Analysis Sentiment. Penjelasan lebih lanjut terkait 3 bagian pembuatan engine tersebut akan dijelaskan pada sub bab di bawah ini.

## Scraping Data

Scraping Data pada pembuatan engine tersebut menggunakan library dari BeautifulSoup untuk mengambil data pada sebuah website. BeautifulSoup4 hanya bisa mengambil data tersebut dengan mempelajari sebuah halam website yang akan dilakukan Scraping. Cara kerja dari BeautifulSoup4 ini sendiri yaitu dengan mepelajari halaman website yang akan dilakukan Scraping, sehingga jika ingin mengambil Data Scraping dari berbeda halaman website maka diperlukannya analisa *source code* yang terdapat pada *domain* atau website tersebut. Pengambilan Scraping Data berita yang akan dilakukan pada tugas ini yaitu dengan menganalisa *source code* yang terdapat pada website berita Indonesia yaitu CNCBIndonesia, Detik.com, Kompas.com, Kumparan, Liputan6 dan TribunNews. Gambaran alur proses pengguna menggu engine scraping akan dijelaskan pada Gambar 4. 1 di bawah ini.

|  |
| --- |
| C:\Users\risma\Desktop\FILE VSION MCA.jpg |

Gambar 4. 1Alur Proses User Menggunakan Engine Scraping

Gambar 4. 1 merupakan Alur Proses User Menggunakan Engine Scraping yang telah dibuat. Pengguna hanya tinggal memasukan parameter yang tersedia yaitu Tanggal Berita dan Pencarian Berita yang akan dicari. Engine nantinya akan mencari berdasarkan parameter yang dimasukan oleh user di Class yang telah dibuat diatas yaitu terdapat 6 Class Berita yaitu Class CNCBIndonesia, Detik.com, Kompas.com, Kumparan, Liputan6 dan TribunNews. Data yang telah diambil dari 6 website berita tersebut nantinya akan disimpan kedalam *database.* Proses Engine dari masing-masing class akan dijelaskan pada Gambar 4. 2di bawah ini.

|  |
| --- |
| C:\Users\risma\Desktop\AlurKerjaClass.jpg |

Gambar 4. 2 Alur Kerja Engine Class Berita

Gambar 4. 2 merupakan alur kerja dari engine class scraping berita secara umum dilakukan. Pembuatan engine dengan mengimport library BeautifulSoup4, mysql.connector, Exception, dan Datetime. Parameter yang didapatkan dari engine tersebut yaitu url berita, Tanggal Berita, dan kata kunci pencarian berita tersebut. Selanjutnya dengan menggunakan library dari BeautifulSoup4 mengambil semua content yang terdapat pada url tersebut. Analisa Tag HTML pada website ini yaitu mencari data-data yang akan diambil pada website tersebut dengan parameter pengambilan tag yaitu tag html dan class yang digunakan. Pada website berita itu hasil yang didapatkan yaitu berupa per halaman dari berita yang ditampilkan, jadi pengambilan semua url detail dari berita tersebut dan di simpan pada variable DataURLDetail. Selanjutnya dilakukan perulangan scraping pada proses Scraping Data dan engine akan menvalidasi data yang akan diambil, apakah sudah pernah di scraping atau belum. Jika sudah maka dilakukan ke pengecekan apakah DataURLDetail masih tersedia, jika masih maka akan dilakukan scraping lagi hingga data pada DataURLDetail itu habis. Data yang berhasil di dapatkan akan disimpan kedalam database. Penjelasan selanjutnya akan menjelaskan source code dari engine Scraping Data.

### Fungsi Main Engine Scraping Data

Fungsi main dari *engine Scraping Data* ini merupakan inti dari *engine* .Main Engine berguna dalam memanggil fungsi-fungsi atau Class yang telah dibuat. Fungsi main dari *engine Scraping Data* ini dapat dilihat pada Gambar 4. 3 sebagai berikut.

|  |
| --- |
|  |

Gambar 4. 3 Engine Master Scraping Data

Gambar 4. 3 merupakan *source code* dari Engine Master Scraping Data. Engine Master karena setiap fungsi dan Class akan dipanggil semuanya di source code tersebut. Kernel pertama adalah bagian import semua Class yang telah dibuat berdasarkan website berita di Indonesia. Pada bagian kernel ketiga atau yang bergambarkan kotak merah tersebutlah pengguna dapet mengganti parameter yang diinginkan. Parameter yang dimasukan adalah Keyword dan Tanggal Berita itu terbit. Parameter tersebut nantinnya digunakan pada fungsi scrapingData, dimana fungsi tersebut berguna dalam melakukan scraping pada tiap-tiap Class pada kernel bagian 2 yaitu def scrapingData. Penjelasan secara umum dari Class akan dijelaskan pada gambar di bawah ini. Data url yang didapatkan pada masing-masing Class tersebut dibuat dengan pengambilan data pada *database*.

|  |
| --- |
|  |

Gambar 4. 4 Parameter Pemanggilan Setiap Class

Gambar 4. 4 merupakan parameter yang digunakan dalam pemanggilan setiap class tersebut. Pengaturan parameter pencarian berita dan tanggal berita dibuat pada engine secara lansung agar pengguna dapat secara fleksibel menggunakan atau mengganti isi dari parameter dengan mudah.

### Pembuatan Class Scraping

Class Scraping merupakan bagian source code yang digunakan dalam Scraping Data berdasarkan url website berita. Terdapat 6 Class yang digunakan dalam melakukan scraping. 6 Class tersebut sebelumnya sudah dianalisa berdasarkan isi tag HTML yang digunakan pada website tersebut. 6 Class tersebut sesuai dengan pengambilan list berita yang akan diambil yaitu CNCBIndonesia, Detik.com, Kompas.com, Kumparan, Liputan6 dan TribunNews.

|  |
| --- |
|  |

Gambar 4. 5 Class Scraping pada Engine

Gambar 4. 5 merupakan Class yang dibuat dalam melakukan Scraping pada website berita. Terdapat 6 Class yang terdiri dari CNBC, Detik, Kompas, Kumparan, Lipuitan6, dan Tribunnews. Dari ke 6 Class tersebut terdapat scraping data yang sama tetapi pemanggilan Tag , Class , dan Id pada masing-masing Class tersebut berbeda. Struktur dari semua Class yang dibuat semuanya mirip dengan file yang lainnya,tetapi dalam tidak bisa di pergunakan untuk semua website dikarenakan setiap website memiliki karakter tag atau class HTML yang berbeda-beda. Jika ingin mengambil data dari website yang lain, terlebih dahulu kalian harus menganalisa website tersebut. Penjelasan lebih dalam mengenai isi dari Class akan dijelaskan secara umum menggunakan Class CNBC.

#### Import Library Class

Import Library Class merupakan tahapan yang terpenting sebelum melakuakan Scraping Data. Import Library memudahkan developer dalam menggunakan fungsi-fungsi yang sudah ada sesuai dengan keperluan dari library yang digunakan. Penjelasan lebih lanjut terkait dengan library yang digunakan dalam Class Scraping ini akan dijelaskan sebagai berikut.

|  |
| --- |
|  |

Gambar 4. 6 Improt Library File Class Berita

Gambar 4. 6 merupakan library yang digunakan pada Class Scraping s

1. Exception : Berguna dalam menghandle error yang terjadi pada engine agar ketika terdapat error pada engine, engine tidak langsung berhenti berjalan.
2. BeautifulSoup : Berguna dalam mengambil semua data HTML berdasarkan url yang dimasukan. User dapat mengambil data pada website dengan mendeklarasi Tag HTML dan Class yang digunakan.
3. Datetime : Berguna dalam pengolahan waktu yang dimasukan oleh pengguna.
4. Get : Mengecek response yang terdapat pada url tersebut benar adanya.
5. Mysql.connector : Berguna dalam melakukan pengolahan data dan penyimpanan data pada database.

#### Constructor Class Scraping Berita

Constructor merupakan suatu method yang akan memberikan nilai awal pada saat suatu class yang dibuat saat melakukan pemanggilan. Penjelasan lebih lanjut terkait dengan Constructor Class Scraping Berita akan dijelaskan sebagai berikut.

|  |
| --- |
|  |

Gambar 4. 7 Constructors Class Scraping Berita

Gambar 4. 7 merupakan Constructor pada class scraping berita. Pemanggilan sebuah class scraping berita tersebut harus menggunakan 3 parameter yang digunakan yaitu urlberita, pencarian berita, dan tanggal tersebut berita tersebut. Method constructor itu juga mendeklarasikan sebuah objek dengan koneksi database yang digunakan pada engine. Terdapat kondisi apabila tanggal berita yang dicari kosong, maka engine akan menggunakan tanggal sekarang untuk mencari sebuah berita.

#### Fungsi Simpan dan Validasi Database

Fungsi *database* dari *engine scraping data* adalah melakukan penyimpanan data dan melakukan validasi data pada *database*. Method pengolahan *database* ini dapat dilihat pada Gambar 4. 8 sebagai berikut.

|  |
| --- |
|  |

Gambar 4. 8 Fungsi Pengolahan Database (Validasi dan Simpan Data)

Gambar 4. 8 meperlihatkan terdapat 2 fungsi yaitu simpanDatabase dan validasiDatabase. Fungsi dari simpanDatabase itu berguna dalam melakukan penyimpanan data setelah melakukan grabbing data yang telah dilakukan. Dengan memanggil fungsi tersebut data yang akan diinput berupa type data tuple sebagai parameter pemanggil fungsi simpanDatabase() dan juga *connection* database yang diambil dari deklarasi connection database diawal. Fungsi dari validasiDatabase() ini berguna dalam melakukan pengecekan terhadap *database* apakah data yang akan discraping sudah terinput atau belum pada *database*. Fungsi tersebut berguna dalam menghindari terjadinya data double pada database. Parameter yang digunakan pada fungsi tersebut adalah *connection* database dan urlDetail dari database tersebut, apakah sudah terdapat urlDetail tersebut pada database dan jika belum maka akan dilakukan input data menggunakan fungsi simpanDatabase().

#### Fungsi Mengambil Halaman pada Website Berita

Fungsi getAllPageWebsite dari engine *scraping data* adalah melakukan pengambilan jumlah halaman yang terdapat pada pencarian berita tersebut. Website berita dalam melakukan pencarian sebuah berita, pasti menyajikan daftar berita itu dalam bentuk list daftar berita yang disajikan dalam page atau halam website. Untuk mengambil semau berita yang terdapat pada semua page tersebut, dengan menganalisa dan mengambil halam terkahir dari berita dan melakukan looping pada url yang telah di deklarasikan untuk mendapatkan urlDetail semua berita yang dicari. Fungsi getAllPageWebsite ini dapat dilihat pada Gambar 4. 9 sebagai berikut.

|  |
| --- |
|  |

Gambar 4. 9 Fungsi getAllPageWebsite pada Class Scraping Berita

Gambar 4.8 merupakan fungsi yang dibuat untuk mengambil jumlah halaman yang terdapat pada website berita tersebut. Return yang diberikan dari pemanggilan fungsi ini sendiri adalah integer jumlah page. Terdapat kondisi apabila terdapat kata kunci pada pencarian maka dia akan masuk ke website index berita dengan tanggal yang dimasukan pengguna ataupun default pengguna telah di atur pada fungsi constractor. Kenapa proses ini diperlukan karena dalam melakukan scraping seorang biasanya hanya mengambil halaman pertama saja, jadi disini tujuannya mengambil jumalah halaman yang berguna dalam mengambil semua data berita yang ada.

#### Fungsi Mengambil URL Detail Berita (getULRinPage)

Fungsi getUrlInPage dari engine scraping data adalah melakukan pengambilan data url dari masing-masing berita yang ada. Pengambilan urlDetail dengan melakukan looping pada jumlah *page* yang didapatkan pada fungsi getAllPageWebsite. Fungsi getUrlInPage ini dapat dilihat pada Gambar 4. 10 sebagai berikut.

|  |
| --- |
|  |

Gambar 4. 10 Fungsi Mengambil Semua URL Berita

Gambar 4. 10 merupakan penjelasan dari Fungsi Mengambil Semua URL Berita. Parameter pemanggilan fungsi tersebut adalah page yang didapatkan dari fungsi getAllPageWesbsite. Maka nantinya akan diambil setiap *url* berita list menggunakan looping dengan url pada page ke berapa. Karena untuk mendapatkan content dari berita tersebut mempunyai harus mempunyai link yang berbeda-beda setiap beritanya, jadi link dari detail berita tersebut akan disimpan kedalam Array dan akan di *looping* untuk mengambil *content* dari detail berita tersebut. Array tersebut nantinya akan digunakan untuk mengambil scraping lebih dalam seperti menggambil data judul,penulis, dan isi berita. Return dari fungsi ini adalah Arrya yang berisikan list-list urldetail berita yang akan digunakan untuk scraping data detail.

#### Fungsi Mengambil Detail Data Berita (scrapingDataDetail)

Fungsi scrapingDataDetail dari engine grabbing data adalah melakukan pengambilan data detail yang terdapat pada urlDetail berita yang ada. Pengambilan data yang diambil berupa judul berita, label berita, penulis, kapan terbitnya, url berita tersebut, dan content dari berita tersebut. Fungsi scrapingDataDetail ini dapat dilihat pada Gambar 4. 11 sebagai berikut.

|  |
| --- |
|  |

Gambar 4. 11 Fungsi Mengambil Semua Data Pada URL Detail

Gambar 4. 11 merupakan sebuah fungsi utama dari sebuah Class Scraping Berita. Pada fungsi ini lah semua fungsi yang telah dibuat sebelumnya yaitu getUrlInPage, getAllPageWebsite, validasiDataDatabase, dan simpanDatabase. Alur pengambilan data dimulai dari pengambilan getUrlInPage yaitu mengambil semua url berita yang terdapat pada halaman website, dengan parameter yang digunakan adalah jumlah page yang didapatkan dari *return* fungsi getAllPageWebsite. Alur selanjutnya yaitu melakukan validasi pada database agar tidak melakukan scraping kemabli yang membuat performa engine melambat dan data double.Validasi tersebut menggunakan fungsi validasiDataDatabase dengan parameter urldetail yang didapatkan. Selanjutnya engine akan mencoba mengambil menggunakan *source* *try catch* untuk menanggulangi terjadinya data missing saat mescraping data. Data yang diambil berupa data judul berita, label berita, penulis, kapan di publishnya sebuah berita, url dan content dari berita tersebut. Selanjutnya data akan disimpan kedalam *database* sesuai dengan fungsi yang dibuat diatas tersebut. Hasil dari grabbing data tersebut dapat dilihat pada Gambar 4. 12 sebagai berikut.

#### Hasil Running Engine Scraping Data (scrapingDataDetail)

Hasil Running Engine Scraping Data dari ke 6 website berita yang telah di analisa tersebut dapat dilihat pada Gambar 4. 12 sebagai berikut.

|  |
| --- |
|  |

Gambar 4. 12 Hasil Scraping Data dari Engine

Gambar 4. 12 merupakan hasil dari setelah engine scrapingData dijalankan. Table yang tidak terisi hanya bagian field summary dan analisa sentiment yang nanti akan dilanjutkan pada engine summarization. Pengambilan dari 6 website berita yang ada di Indonesia itu engine kurang lebih membutuhkan waktu 20 menit dengan total data yang didapatkan 1148 data. Pengambilan data tersebut tergantung dengan katagori pencarian dan tanggal yang ditentukan. Pada pencarian tersebut saya menggunakan kata kunci dari Vaksin dan tanggal 27 December 2021.

## Text Summarization

Text Summarization adalah pembuatan bentuk yang lebih singkat dari suatu teks dengan memanfaatkan suatu algoritma tertentu. Alur proses dari melakukan text summarization melalui beberapa tahapan yaitu Persiapan Data, melakukan Text Preprocessing, Cleansing Data, Case Folding, Tokenizing Data, dan Ssteamming data tersebut agar mendapatkan hasil yang maximal pada perhitungan Similarity kata. Proses Text Summarization ini akan dijelaskan lebih lanjut sebagai berikut.