**TEXT SUMMARIZATION MENGGUNAKAN ALGORITMA TEXTRANK LIBRARY SPACY DAM GRABBING DATA MENGGUNAKAN LIBRARY BEATIFULSOAP4**

**MULTI CHANNEL ACCESS - MDI (A)**

Diajukan guna memenuhi sebagian persyaratan dalam rangka menyelesaikan Pendidikan Sarjana Strata Satu (S1) Program Studi Teknologi Informasi



I Made Rismawan Nugraha   
NIM :1805551114

**PROGRAM STUDI TEKNOLOGI INFORMASI  
FAKULTAS TEKNIK  
UNIVERSITAS UDAYANA  
2021**

# Deskripsi

Kebutuhan akan informasi di jaman yang moderen ini merupakan suatu kebutuhan yang sangat penting dan mudah didapat karena adanya perkembangan teknologi komunikasi berdampak pada penggunaan internet untuk mempublikasi media digital. Media digital dapat berupa artikel, blog, maupun situs-situs berita yang berisi informasi seperti dokumen (*teks*). Sebuah media digital kebanyakan dibuat dengan penulisan kata yang berulang sehingga menemukan sebuah representasi yang tepat dari informasi tekstual tersebut merupakan sebuah hal yang penting untuk membuat pemahaman dari keseluruhan teks. Sebuah dokumen tersebut dapat diringkas oleh suatu sistem tanpa menghilangkan informasi yang penting maka pembaca dapat menghemat waktu, karena pembaca dapat memahami dan mengetahui informasi penting dari dokumen tersebut tanpa harus membaca isi dokumen secara keseluruhan.

Meringkas pada umumnya dapat dilakukan dengan mudah jika pembaca relative familiar terhadap teks yang sudah dibaca atau jika teks tersebut tidak terlalu panjang. Hasil rangkuman setiap individu yang bersifat subjektif sedangkan untuk menghasilkan sebuah rangkuman yang baik dan bagus perlu pertimbangan yang bersifat objektif. Meringkas dokumen secara manual oleh manusia, membutuhkan banyak biaya dan waktu apabila dokumen tersebut banyak dan panjang sehingga diperlukan sistem peringkas otomatis (*automatic summarization)* untuk mengatasi banyaknya biaya dan waktu tersebut (Mustaqhfiri, Abidin, & Kusumawati, 2012).

*Automatic Summarization* yang dirancang harus efisien terhadap waktu dan efektif terhadap ketepatan penyajian informasi. Ada beberapa metoda dan algortima yang digunakan dalam membuat engine *Automatic Summarization*. Algoritma yang digunakan untuk peringkasan artikel adalah algoritma TextRank. TextRank merupakan *graph-based ranking* *algorithm* (graf dengan model pemeringkatan) untuk pemrosesan teks. TextRank yang digunakan sebagai metode yang menghasilkan ekstraksi berupa kalimat (*TextRank for Sentence Extraction*). TextRank sangatlah fleksibel karena dapat digunakan pada berbagai bahasa tanpa mengubah algoritmanya. Hal ini dikarenakan TextRank tidak memerlukan data training untuk proses pengelolahan dokumen (Pragantha, & Tarumanagara, 2017).

# Metode Pembuatan

Bagian Progress akan menjelaskan progress atau tahapan yang akan dilakukan penulis. Progress berisikan tahapan dari pembuatan *Automatic Summarization* yang dapat dibagi menjadi 2 yaitu proses pembuatan *engine* *Automatic Summarization* dan *engine Grabbing Data.* Penjelasan lebih lanjut mengenai progress akan dijabarkan sebagai berikut.

## Grabbing Data

Grabbing adalah suatu teknik mengambil text atau secara umum obyek yang ada di situs lain yang nantinya dipergunakan untuk mendapatkan data dari tampilan sebuah website tersebut kemudian diletakkan ke website kita. Kesimpulannya jika terdapat perubahan data pada website parent yang kita lakukan grabbing tersebut maka otomatis data yang kita ambil juga pasti akan berubah tanpa harus merubahnya secara manual. Pada tahapan pembuatan engine Grabbing data saya menggunakan *library open source* yang dapat digunakan yaitu library BeautifulSoup. Penjelasan terkait dengan library tersebut dapat dilihat pada penjelasan berikutnya.

### BeautifulSoup

BeautifulSoup merupakan salah satu *library open source* yang dapat digunakna pada Python yang memungkinkan untuk melakukan scraping dengan mudah dan cepat. BeautifulSoup hadir dengan metode-metode sederhana sehingga dapat dengan mudah melakukan proses navigasi, pencarian dan juga modifikasi struktur data situs yang di-scraping. BeautifulSoup dapat mengekstrak data, baik dari file XML maupun HTML dengan mengubah markup menjadi parse tree. Parsing HTML adalah suatu teknik untuk memisahkan suatu teks dari *tag-tag code* HTML sehingga menghasilkan sebuah data yang spesifik. Pengambilan data yang biasanya dilakukan secara manual dapat dilakukan secara otomatis dengan menggunakan BeautifulSoup. (Wesley, 2019).

|  |
| --- |
|  |

Gambar 2. 1 Source Code Testing Grabing Data

Gambar 2.1 merupakan tahapan awal di mana saya akun melakukan pengambilan data pada website. Jadi setelah meng-GET data pada URL yang telah di deklarasikan di awal, parse data tersebut menggunakan library dari beatifulsoup dan mengambil element html pada URL tersebut. Jadi di awal saya mengambil tanggal tayang dan juga nama bioskop yang dapat dilihat pada gambar di atas. Jadi pada dasarnya kita mengambil suatu informasi melalui tag-tag HTML yang ada pada sebuah *article* atau media digital yang ada.

## Text Summarization

Text SummarizationPeringkasan teks otomatis (*automatic text summarization*) adalah pendekatan yang bisa digunakan untuk meringkas aritkel yang panjang menjadi lebih pendek dan ringkas, sehingga hasil rangkuman teks yang relative pendek bisa mewakilkan teks yang panjang tersebut. Pada dasarnya menggunakan prinsip *Natural Language Processing* dan algoritma untuk membuat sebauh engine tersebut. Pada penelitian ini saya menggunakan algoritma dari TextRank yang nantinya akan dijelaskan lebih detail pada penjelasan berikutnya.

### TextRank

TextRank adalah algoritme tanpa pengawasan untuk peringkasan otomatis dari teks yang juga dapat digunakan untuk mendapatkan kata kunci terpenting dalam sebuah dokumen yang diperkenalkan oleh Rada Mihalcea dan Paul Tarau. Algoritma menerapkan variasi PageRankdi atas grafik yang dibangun khusus untuk tugas meringkas. Ini menghasilkan peringkat elemen-elemen dalam grafik: elemen yang paling penting adalah elemen yang lebih baik mendeskripsikan teks. Pendekatan ini memungkinkan TextRank untuk membuat ringkasan tanpa kebutuhan suatu kalimat yang korpus. Pelatihan atau pelabelan dan memungkinkan penggunaan algoritma dengan bahasa yang berbeda-beda.

|  |
| --- |
|  |

Gambar 2. 2 Skema Peringkasan Menggunakan Algoritma TextRank

Gambar 2.2 merupakan cara kerja dari teknik peringkasan algoritma TextRank yang dimana prosesnya adalah mencari kalimat yang berada di daftar urutan teratas akan dipilih menjadi ringkasan ekstraktif untuk dokumen tersebut. Banyaknya kalimat yang akan menjadi hasil ringkasan tergantung dari persentase nilai kompresi yang ditentukan oleh user. Pengguna dapat menentukan nilai kompresi dari sebuah artikel atau sebuah paragraph yang diinputkan. Terdapat sebuah *library open source* yang dapat digunakan dalam menerapkan algoritma TextRank tersebut yaitu library spaCy yang berguna sebagai teknik peringkasan sebuah paragraph dengan menggunakan bahasa python.Penjelasan lebih lanjut mengenai *library* spaCy tersebut akan dijelaskan pada penjelasan berikutnya.

### spaCy

spaCy adalah perpustakaan untuk Natural Language Processing dengan Python. SpaCy ini mendukung tokenisasi dan pelatihan untuk 60+ bahasa. ni menampilkan kecepatan canggih dan model jaringan saraf untuk penandaan, penguraian, pengenalan entitas bernama, klasifikasi teks, dan banyak lagi, pembelajaran multi-tugas dengan transformator yang telah dilatih sebelumnya seperti BERT, serta sistem pelatihan siap produksi dan model yang mudah. pengemasan, penyebaran, dan manajemen alur kerja. spaCy adalah open-source software, dirilis di bawah lisensi MIT.

# Progress

Bagian progress ini menjelaskan tentang progress pembuatan engine yang sudah dikerjakan. Adapun penjelasan progress akan dijelaskan lebih lanjut sebagai berikut.

1. **Grabbing Data**

Grabbing data pada pembuatan engine tersebut menggunakan library dari BeautifulSoup untuk mengambil data pada sebuah website. Setelah mendalami mempelajari library tersebut hanya dapat mengambil data berdasarkan tag html yang terdapat pada domain tersebut, sehingga jika ingin mengambil data scraping dari berbeda domain maka diperlukannya analisa source code yang terdapat pada domain tersebut. Progress yang telah dibuat kali ini mengambil data scraping pada salah satu website berita CBNC Indonesia berdasarkan berita terbaru atau kata kunci yang ingin dimasukan oleh user. Penjelasan lebih lanjutnya akan dijelaskan pada gambar berikutnya.

|  |
| --- |
|  |

Gambar 3. 1 Mengambil Jumlah Page pada Domain Website CBNC

Gambar 3.1 merupakan proses pengambilan data halaman pada domain CBNC Indonesia dan juga anda dapat mengatur configurasi kata kunci dari berita yang akan dicari pada website tersebut. Nah untuk mencoba mengambil data dari lain website nantinya rencanannya akan dibuatkan validasi pencarian website tersebut. Mengambil semua data halaman tersebut berguna untuk mengambil semua data yang berkatian dari kata kunci tersebut, karena biasanya dia user hanya mengambil data pada page pertama saja.

|  |
| --- |
|  |

Gambar 3. 2 Mengambil URL Berita Detail

Gambar 3.2 merupakan penjelasan pembuatan fungsi dari penyimpan data yang telah diambil kedalam database dan pada kernel bagian pertama menjelaskan tentang pengambilan url dari headline berita yang ada pada CBNC Indonesia. Karena untuk mendapatkan content dari berita tersebut mempunyai link yang berbeda-beda jadi, link dari detail berita tersebut akan disimpan kedalam array dan akan di looping untuk mengambil content dari detail berita tersebut.

|  |
| --- |
|  |

Gambar 3. 3 Mengambil Semua Data Pada URL Detail yang diSimpan

Gambar 3.3 menjelaskan terkait dengan pengambilan semua data pada CBNC Indonesia. Skema pengambilan data yang telah dibuat saya buatkan validasi pada database agar tidak terdapat data double yang disimpan pada database. Sebagai parameter pembeda saya menggunakan link dari detail berita tersebut untuk menentukan apakah sebuah data tersebut sudah ada pada database. Ketika tidak ada maka engine akan mencoba mengambil menggunakan fungsi 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.

|  |
| --- |
|  |

Gambar 3. 4 Hasil Penyimpanan Pada Database

Gambar 3.4 merupakan hasil dari scraping database dan melakukan penyimpanan kedalam database mysql yang masih berada dilocal computer saya.