Tugas Besar III IF2211 Strategi Algoritma

Pencarian Berita pada News Aggregator dengan Algoritma Pencocokan String



Dipersiapkan oleh :

WTS > Solusi Tubes 3 Stima Minat PM

Adrian Mulyana Nugraha 13515075

Vincent Hendryanto Halim 13515089

Mikhael Artur Darmakesuma 13515099

**Program Studi Teknik Informatika**

**Sekolah Teknik Elektro dan Informatika**

**Institut Teknologi Bandung**

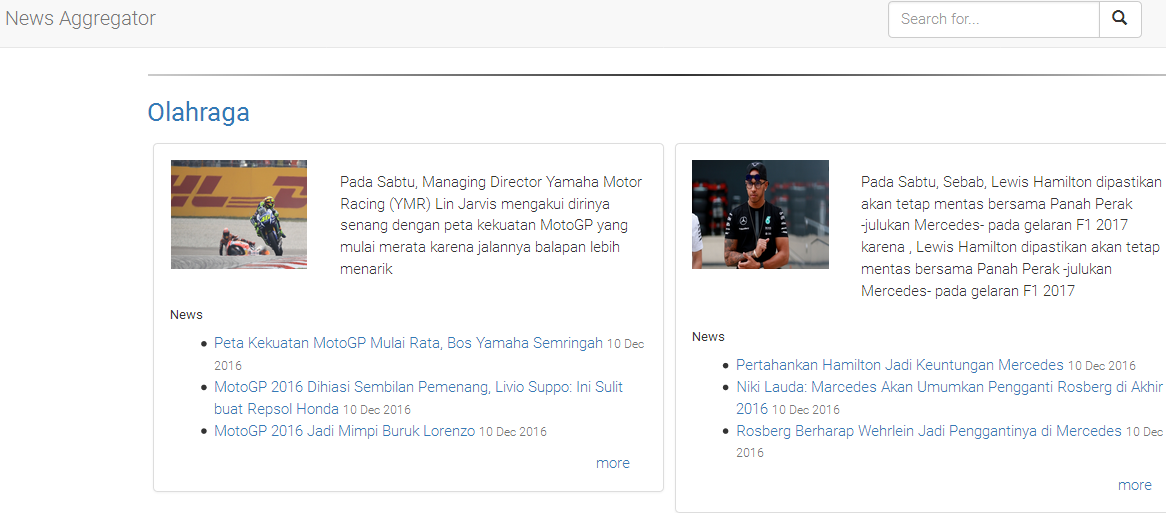
**2017**

# Deskripsi Masalah

Sistem agregasi berita (*news aggregator*) dikembangkan untuk membantu pembaca berita dengan mengumpulkan informasi berita dari berbagai sumber dan menyajikannya dalam satu tempat. Dengan sistem ini, pembaca tidak perlu mencari berita sendiri, dan aplikasi dapat mengambil berita sesuai kebutuhan dari pembaca (Lasica, 2003). Pencarian berita merupakan salah satu fitur yang terdapat pada sistem ini.



Gambar 1. News aggregator dari Google



Gambar 2. Prototipe News aggregator dari Lab GAIB ITB



Gambar 3. News aggregator dari situs www.newslookup.com

Algoritma pencocokan *string* (*pattern*) Knuth-Morris-Pratt (KMP) dan Algoritma Boyer-Moore merupakan algoritma yang lebih baik daripada *brute force*. Pada Tugas Besar III kali ini Anda diminta membuat aplikasi sederhana pencarian berita pada *news aggregator* dengan kedua algoritma tersebut, plus menggunakan *regular expression* (*regex*). Teks yang akan Anda proses adalah kumpulan berita berbahasa Indonesia. Pengguna aplikasi ini akan memberikan masukan berupa keyword pencarian, dan menghasilkan daftar berita yang diurut berdasarkan tanggal berita.

Pencocokan *string* yang anda buat adalah *exact matching* (untuk KMP dan BM) jadi artikel berita yang diproses mengandung string yang tepat sama dengan *keyword* yang dimasukkan oleh pengguna. Sedangkan bila menggunakan *regex* maka tidak selalu *exact matching*. Pencarian juga tidak bersifat *case sensitive*, jadi huruf besar dan huruf kecil dianggap sama (hal ini dapat dilakukan dengan mengganggap seluruh karakter di dalam *pattern* dan teks sebagai huruf kecil semua atau huruf kapital semua).

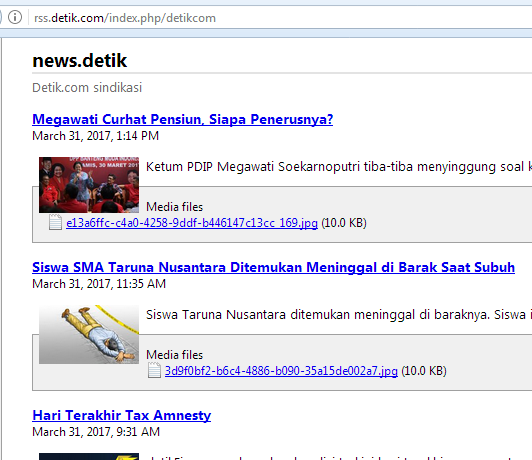
Kumpulan berita diambil secara otomatis menggunakan *crawler* berbasis *RSS* (*rich site summary* atau *really simple syndication*) dari situs berita daring berbahasa Indonesia. Saat membaca *RSS* dengan *XML parser*, informasi yang dibutuhkan berupa judul, tanggal berita, dan *URL* berita. Berikut daftar *RSS* yang dapat digunakan:

<http://rss.detik.com/index.php/detikcom>

<http://tempo.co/rss/terkini>

<http://rss.vivanews.com/get/all>

<http://www.antaranews.com/rss/terkini>



Gambar 4. Isi RSS detik.com jika dilihat dari browser

Untuk setiap *URL*, unduh artikelnya dan lakukan *parsing*. Artikel berupa file HTML dan tidak hanya mengandung konten berita, tetapi masih mengandung *header*, *footer*, iklan, dan tambahan informasi pada situs berita tersebut. Untuk itu, dilakukan *parsing* HTML untuk mendapatkan hanya teks konten berita dan foto yang terkait berita tersebut. Salah satu library *html* *parser* yang dapat digunakan misalnya adalah <https://jsoup.org/> untuk Java, *RSS parser* untuk .NET (<http://stackoverflow.com/questions/684507/rss-parser-in-net>) dan *XML parser* dalam Bahasa C# ( <http://stackoverflow.com/questions/642293/how-do-i-read-and-parse-an-xml-file-in-c> ),

# Dasar Teori

# Analisis Pemecahan Masalah

# Implementasi dan Pengujian

# Kesimpulan dan Saran