Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung

Tugas Kecil IV IF2211 Strategi Algoritma
Ekstraksi Informasi dari Artikel Berita dengan Algoritma Pencocokan String

Batas pengumpulan : Rabu, 22 April 2020 jam 13.
Arsip pengumpulan : Laporan dan kode program dikumpulkan ke dropbox pada tautan berikut: http://irklab.site/tucil4stima. Laporan berisi:

a. Deskripsi singkat pencocokan string Knuth-Morris-Pratt (KMP), Boyer-Moore, Regex.
b. Kode program,
c. Screen-shot input-output program.

Algoritma pencocokan string (pattern) Knuth-Morris-Pratt (KMP) dan Algoritma Boyer-Moore merupakan algoritma yang lebih baik daripada brute force. Pada Tugas Kecil IV kali ini Anda diminta membuat aplikasi sederhana ekstraksi informasi dengan kedua algoritma tersebut, plus menggunakan regular expression (regex). Teks yang akan Anda proses adalah teks berita berbahasa Indonesia seperti contoh berikut ini (jabar11042020.txt).

> 421 Orang di Jabar Terkonfirmasi Positif COVID-19
>
> Yudha Maulana - detikNews
>
> Sabtu, 11 Apr 2020 20:07 WIB
>
> **Bandung** - Angka positif virus Corona atau COVID-19 di Jawa Barat menembus angka 400 kasus. Laman Pusat Informasi dan Koordinasi COVID-19 Jabar (Pikobar) pada Sabtu (11/4/2020) pukul 18.43 WIB, mencatat terdapat 421 orang yang terkonfirmasi positif COVID-19.
>
> Dibandingkan sehari sebelumnya, jumlah tercatat yaitu 388 orang. Terjadi penambahan 8,5 persen atau 33 kasus per harinya. Sementara itu, secara nasional terdapat 3.842 kasus positif COVID-19.
>
> Dari 421 kasus tersebut, 40 orang meninggal dunia dengan keterangan terpapar COVID-19. Sedangkan, angka kesembuhan di Jabar masih tetap berada di angka 19 orang.
>
> Per hari jumlah Orang Dalam Pemantauan (ODP) di Jabar mencapai 28.775 orang. Sebanyak 15.363 di antaranya masih menjalani proses pemantauan dan 13.412 orang lainnya telah selesai menjalani proses pemantauan.
>
> Sementara itu jumlah Pasien Dalam Pengawasan (PDP) mencapai 2.278 orang. Tercatat 1.344 orang masih menjalani proses pengawasan dan 934 orang lainnya telah selesai menjalani proses pengawasan.

Pada kumpulan teks berita korban covid-19 ini, informasi penting dari pengguna adalah jumlah korban dan waktunya. Oleh karena itu, informasi yang akan diekstraksi adalah angka (diberi warna biru) dan waktu (diberi warna merah).

Pengguna aplikasi ini akan memberikan masukan berupa folder yang berisi kumpulan teks berita, keywords, dan hasil ekstraksi jumlah dan waktunya. Karena sebagian besar kalimat mengandung angka, aplikasi akan memfilter angka berdasarkan keywords dari pengguna, seperti ‘terkonfirmasi positif’, ‘meninggal dunia’, ‘Orang Dalam Pemantauan’, ‘ODP’, ‘Pasien Dalam

Pengawasan', 'PDP' atau keyword lainnya. Hasilnya berupa pasangan angka dan waktu, serta kalimat yang mengandung informasi tersebut. Waktu yang diambil harus berada dalam satu kalimat dengan angka tersebut. Jika tidak ada, gunakan tanggal artikel yang tercantum. Jika terdapat lebih dari satu angka, pilih angka yang paling dekat dengan keyword. Berikut contohnya.

> Keyword: *terkonfirmasi positif*
> Hasil ekstraksi informasi:
> Jumlah: *421;* Waktu: Sabtu, 11 Apr 2020 20:07 WIB
> 421 Orang di Jabar **Terkonfirmasi Positif** COVID-19. (jabar11042020.txt)
>
> Jumlah: *421;* Waktu: *Sabtu (11/4/2020) pukul 18.43 WIB*
> Laman Pusat Informasi dan Koordinasi COVID-19 Jabar (Pikobar) pada Sabtu (11/4/2020) pukul 18.43 WIB, mencatat terdapat 421 orang yang **terkonfirmasi positif** COVID-19. (jabar11042020.txt)
>
> Keyword: *meninggal dunia*
> Hasil ekstraksi informasi:
> Jumlah: *40;* Waktu: Sabtu, 11 Apr 2020 20:07 WIB
> Dari 421 kasus tersebut, 40 orang **meninggal dunia** dengan keterangan terpapar COVID-19. (jabar11042020.txt)

Terdapat dua jenis pencocokan string yang Anda lakukan. Pertama, exact match dengan keyword yang diberikan pengguna untuk memfilter kalimat yang akan diproses informasinya. Semua teknik (KMP, BM, dan regex) bisa digunakan untuk fitur ini. Kedua, ekstraksi jumlah dan waktu dari kalimat hasil exact match dengan menggunakan regex.

Pencarian 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).

Spesifikasi program :

1. Aplikasi yang anda buat merupakan aplikasi web yang menerima keyword pencarian, misalnya "*terkonfirmasi positif*". Tampilan antarmuka pengguna seperti berikut.

My InfoExtraction App

Folder : <browse>

Keyword : <keyword>

Algoritma :

- Boyer-Moore
- KMP
- Regex

1. Jumlah: … ; Waktu: ….
   <kalimat> (<namafile>)

2. Jumlah: … ; Waktu: ….
   <kalimat> (<namafile>)

...

Perihal

Perihal : link ke halaman tentang program dan pembuatnya
Anda dapat menambahkan menu lainnya, gambar, logo, dan sebagainya

2. Aplikasi menggunakan hasil implementasi algoritma KMP, Boyer-Moore, dan Regex dengan menggunakan bahasa python. Pencocokan string dilakukan pada konten berita (teks).

Data Uji
Data uji yang digunakan dapat anda tentukan sendiri, minimal terdapat folder yang berisi 10 teks berita. Posting dapat berbahasa Indonesia atau Inggris.

Lain – lain :
1. Anda dapat menambahkan fitur fungsional lain yang menunjang program yang anda buat (unsur kreatifitas diperbolehkan/dianjurkan).
2. Program berbasis web dan dapat dikembangkan dengan salah satu kakas: php, flask, javascript.
3. Program implementasi Boyer-Moore dan KMP menggunakan bahasa python.
4. Program harus modular dan mengandung komentar yang jelas.
5. Mahasiswa harus membuat program sendiri kecuali library file dan regex, tetapi belajar dari contoh-contoh program serupa yang sudah ada tidak dilarang (tidak boleh mengkopi source code

dari program orang lain). Program harus dibuat sendiri, tidak boleh sama dengan teman. Keterlambatan pengumpulan akan mengurangi nilai.
6. Program disimpan di dalam folder StrAlgo4-xxxxx. Lima digit terakhir adalah NIM Anda. Di dalam folder tersebut terdapat tiga folder bin, src dan doc yang masing-masing berisi :
a. Folder src berisi source code dari program
b. Folder test berisi data uji.
c. Folder doc berisi dokumentasi program dan readme
7. Semua pertanyaan menyangkut tugas ini harus dikomunikasikan melalui milis agar dapat dicermati oleh semua peserta kuliah IF2211 (milis IF2211@students.if.itb.ac.id).

Tambahkan cek list berikut (centang dengan v) di dalam laporan anda untuk memudahkan Asisten dalam menilai:

Poin	Ya	Tidak
1. Program berhasil dikompilasi		
2. Program berhasil running		
3. Program dapat menerima input dan menuliskan output.		
4. Luaran sudah benar untuk data uji		

- Selamat mengerjakan -