

**LAPORAN TUGAS KECIL 4 IF2211 STRATEGI ALGORITMA
SEMESTER II TAHUN 2019/2020**

**Ekstraksi Informasi dari Artikel Berita dengan Algoritma
Pencocokan String**



Dipersiapkan oleh:

Muchammad Ibnu Sidqi - 13518072

Sekolah Teknik Elektro dan Informatika - Institut Teknologi Bandung

Jl. Ganesha 10, Bandung 4013

Daftar Isi

1. Pencocokan string Knuth-Morris-Pratt (KMP), Boyer-Moore, Regex.....	3
2. Kode Program.....	4
3. Dokumentasi.....	6
4. Lampiran.....	11

1. Pencocokan String Knuth-Morris-Pratt (KMP)

Algoritma Knuth-Morris-Pratt (KMP) mencari pola dalam teks dengan urutan pencarian dari kiri-kanan (seperti algoritma brute force). Perbedaannya terletak pada pencocokan pola huruf yang digunakan. pencocokan pola huruf pada KMP berdasar pada panjang prefix dan suffix dari pola yang dicari yang mana pergeseran dilakukan dengan menggunakan $s = (\text{panjang pola dicari}) - (\text{panjang pola huruf prefix} = \text{suffix})$ hingga mencapai akhir dari teks. Dengan menggunakan algoritma KMP didapatkan kompleksitas asimptotik $O(m)$ untuk mencari pinggiran dan $O(n)$ untuk pencarian string sehingga diperoleh kompleksitas waktu asimptotik $T = O(n + m)$. Pada algoritma KMP terdapat kekurangan saat mengelola teks dalam skala besar.

2. Pencocokan String Boyer-Moore

Algoritma Boyer-Moore mencari pola dalam teks yang memiliki 2 teknik dalam urutan pencarian yaitu:

A. Teknik *looking-glass*

cari pola dalam teks dengan bergerak mundur melalui pola, mulai dari akhir

B. Teknik *character-jump*

- ketika ketidakcocokan terjadi pada $T[i] \neq x$
- karakter dalam pola $P[j]$ bukan sama dengan $T[i]$

sehingga dalam proses nya akan terdapat 3 kasus yaitu:

- A. Jika P berisi x di suatu tempat, maka cobalah untuk bergeser P ke kanan untuk menyelaraskan kejadian terakhir dari x dalam P dengan $T[i]$.
- B. Jika P berisi x di suatu tempat, tetapi bergeser ke kanan sampai kejadian terakhir tidak mungkin, maka geser P ke kanan dengan 1 karakter ke $T[i + 1]$.
- C. Jika case 1 dan 2 tidak berlaku, maka pindahkan P lalu sejajarkan $P[0]$ dengan $T[i + 1]$.

Kompleksitas waktu asimptotik Boyer-Moore dalam kondisi worst case adalah $O(nm + A)$ dengan n adalah panjang teks, m adalah banyak pencocokan teks yang terjadi pada setiap huruf, dan A adalah faktor pengecekan alfabet terakhir. Dengan demikian, algoritma Boyer-Moore tidak cocok untuk pencocokan pola yang terlalu kecil seperti *binary*, tetapi cocok untuk pencocokan pola dalam skala besar.

3. Pencocokan String Regex

Pencocokan string regex mencari string pada teks dengan berdasar pada pola regular expression. Regular expression adalah deretan karakter spesial yang mendefinisikan sebuah pola dalam pencarian teks. Perbedaan algoritma regex dengan algoritma lain adalah pendefinisian suatu pola dapat berdasar pada pola yang umum(tidak spesifik) sehingga pencarian string terkait menjadi lebih mudah dan ringkas. Kelemahan pada regex adalah pendefinisian pola umum cenderung lebih sulit sehingga rentan terjadinya kesalahan. Salah satu implementasi pencocokan string dengan regex dapat ditemukan pada salah satu pustaka bahasa python yaitu pustaka *re*.

4. Kode Program

4.1. Knuth-Morris-Pratt (KMP)

```
class kmp:
    def kmp(self, pattern)
        ''' I.S Pattern yang akan dicari, F.S indeks dari pattern(bila
        tidak ditemukan akan menghasilkan -1'''

    def computeFail(self, pattern)
        ''' I.S Pattern yang akan dicari, F.S suffix=prefix dari sub
        pattern yang akan dicari'''

    def convertText(self,name_file)
        ''' I.S direktori file yang akan dicocokkan, F.S list of string
        dari file yang sudah diekstraksi'''
```

4.2. Boyer-Moore

```
class boyce:
    def bmMatch(self, pattern)
        ''' I.S Pattern yang akan dicari, F.S indeks dari pattern(bila
        tidak ditemukan akan menghasilkan -1'''

    def buildLast(self, pattern)
        ''' I.S Pattern yang akan dicari, F.S faktor A dalam pencocokan
        Boyer-Moore'''

    def convertText(self,name_file)
        ''' I.S direktori file yang akan dicocokkan, F.S list of string
        dari file yang sudah diekstraksi'''
```

4.3. Regex

```
class regex:

    def getBaseOfDate(self)
        ''' I.S basis dari tanggal(tanggal berita) kosong, F.S basis
        dari tanggal sudah ditemukan.'''

    def getDate(self)
        ''' I.S kumpulan string hasil ekstraksi, F.S list of tanggal
        dari string hasil ekstraksi.'''
```

```

def getDigit(self)
''' I.S kumpulan string hasil ekstraksi, F.S list of
jumlah/angka dari string hasil ekstraksi.'''

def getIndexOfDay(self)
''' I.S kumpulan string hasil ekstraksi, F.S list of indeks dari
tanggal hasil ekstraksi.'''

def getIndexOfDay(self)
''' I.S kumpulan string hasil ekstraksi, F.S list of indeks dari
jumlah/angka hasil ekstraksi.'''

def getRealDigit(self)
''' I.S kumpulan indeks jumlah/angka hasil ekstraksi, F.S
Jumlah/angka terkait dengan pattern.'''

def checkData(self)
''' I.S list yang akan dicek apakah kosong atau tidak, F.S Bila
kosong mengembalikan true.'''

def getRealDate(self)
''' I.S kumpulan indeks tanggal hasil ekstraksi, F.S tanggal
terkait dengan pattern.'''

def regexMatch(self)
''' I.S kumpulan string hasil ekstraksi, F.S list of hasil yang
akan ditampilkan dengan format [tanggal,jumlah/angka,kalimat]
yang terkait dengan pattern.'''

```

4.4. extraction

```

class extraction:

    def __init__(self,uploaded_files, text, option)
''' melakukan inisiasi direktori file(unploded_files),
pattern(text), dan option(untuk memilih metode mana yang akan
digunakan). '''

    def getHasil(self)
''' mendapatkan hasil dari ekstraksi string dengan metode
terkait'''

```

4.5. flask program(coba.py)

```

def upload_file()

```

```
''' rute untuk mengupload file dari direktori '''

def select_algorithm()
''' rute untuk memilih algoritma yang digunakan serta pattern apa
yang akan dicocokkan '''

def results():
''' rute untuk melakukan proses serta mengeluarkan hasil dari
pencocokan string '''

def perihal():
''' rute untuk menampilkan informasi pembuat. '''
```

5.Dokumentasi Program

•

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.

1. Metode Regex

Extraction Information

Harap Masukkan File

No file chosen

Pilih Algoritma yang akan digunakan

☒ KMP
 ☐ Boyer Moore
 ☐ Regex

Hasil dari ekstraksi informasi pada teks

Keyword: Terkonfirmasi Positif

Jumlah: 421

Waktu: sabtu, 11 apr 2020 20:07 wib

Kalimat: 421 Orang di Jabar Terkonfirmasi Positif COVID-19

Jumlah: 421

Waktu: sabtu (11/4/2020) pukul 18.43 wib

Kalimat: 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.

[coba lagi](#)

[perihal](#)

New button.

Snipping Tool is moving...

In a future update, Snipping Tool will be moving to a new home. Try improved features and snip like usual with Snip & Sketch.

2. Metode KMP

Extraction Information

Harap Masukkan File

No file chosen

Pilih Algoritma yang akan digunakan

☒ KMP
 ☐ Boyer Moore
 ☐ Regex

Hasil dari ekstraksi informasi pada teks

Keyword: Terkonfirmasi Positif

indeks pada file: 19

[coba lagi](#)

[perihal](#)

3. Metode Boyer-Moore

Extraction Information

Harap Masukkan File

No file chosen

Pilih Algoritma yang akan digunakan

☒ KMP
 ☐ Boyer Moore
 ☐ Regex

Hasil dari ekstraksi informasi pada teks

Keyword: Terkonfirmasi Positif

indeks pada file: 19

[coba lagi](#)

[perihal](#)

1891 Jiwa Tewas Sehari, Rekor Tertinggi Korban Covid-19 di AS Jakarta, CNBC

Indonesia - Amerika Serikat (AS) mencetak rekor tertinggi dalam kasus kematian akibat virus corona (COVID-19). Dilansir The Economic Times, sebanyak 1.891 jiwa tewas dalam 24 jam terakhir.

Lonjakan angka tersebut menjadikan total kematian akibat virus corona di AS mencapai 38.664 kasus pada Sabtu (18/4/2020) waktu setempat.

Jumlah tersebut menjadi kasus kematian akibat COVID-19 tertinggi di negara manapun di dunia. Di waktu yang sama, kematian akibat virus corona melonjak 100.000 di Eropa, terhitung hampir dua pertiga dari 157.539 kematian di seluruh dunia.

Hingga Minggu (18/4/2020), AS tercatat memiliki 738.913 kasus terjangkit corona. Jumlah itu menjadikan AS sebagai negara dengan kasus COVID-19 dan kematian tertinggi di dunia.

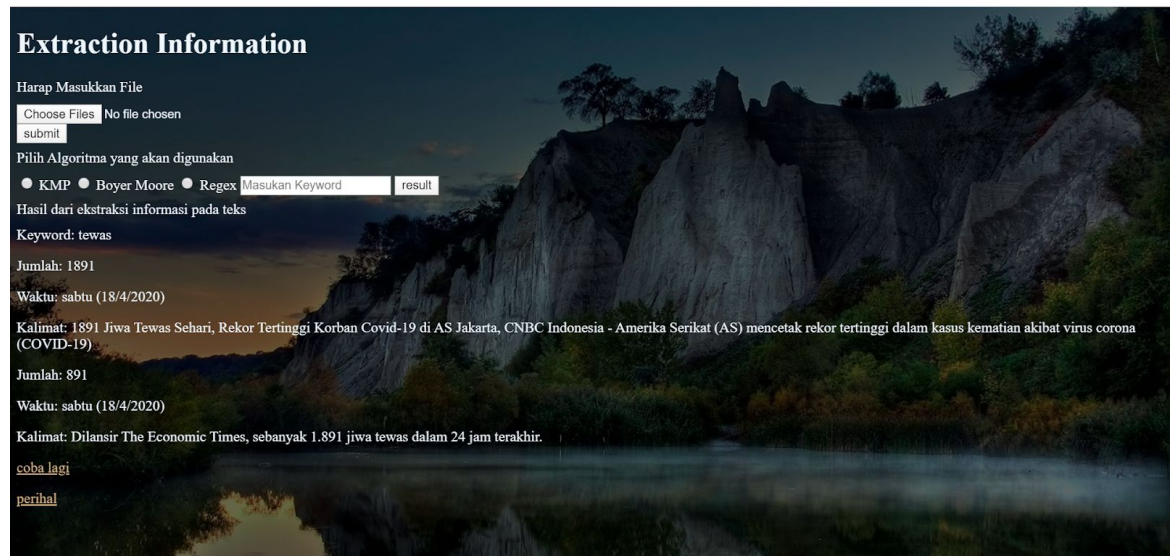
Negara bagian New York menjadi wilayah dengan kasus corona terbanyak di AS, yakni mencapai 241.041 kasus terjangkit, dengan 17.671 kasus kematian.

Di tengah lonjakan kasus corona di Negeri Paman Sam tersebut, Presiden Donald Trump malah berencana melonggarkan kebijakan pembatasan pergerakan secara bertahap.

Langkah tersebut tetap akan dilakukan Trump, meskipun sempat ditolak oleh sejumlah gubernur negara bagian yang khawatir pelonggaran pembatasan pergerakan akan memicu penyebaran corona gelombang kedua di sana.

Secara global, sudah ada 2.331.727 kasus terjangkit, dengan 160.759 kasus kematian, dan 597.194 kasus berhasil sembuh, menurut data Worldometers.

1. Metode Regex



Extraction Information

Harap Masukkan File

No file chosen

Pilih Algoritma yang akan digunakan

☒ KMP ☐ Boyer Moore ☐ Regex

Hasil dari ekstraksi informasi pada teks

Keyword: tewas

Jumlah: 1891

Waktu: sabtu (18/4/2020)

Kalimat: 1891 Jiwa Tewas Sehari, Rekor Tertinggi Korban Covid-19 di AS Jakarta, CNBC Indonesia - Amerika Serikat (AS) mencetak rekor tertinggi dalam kasus kematian akibat virus corona (COVID-19)

Jumlah: 891

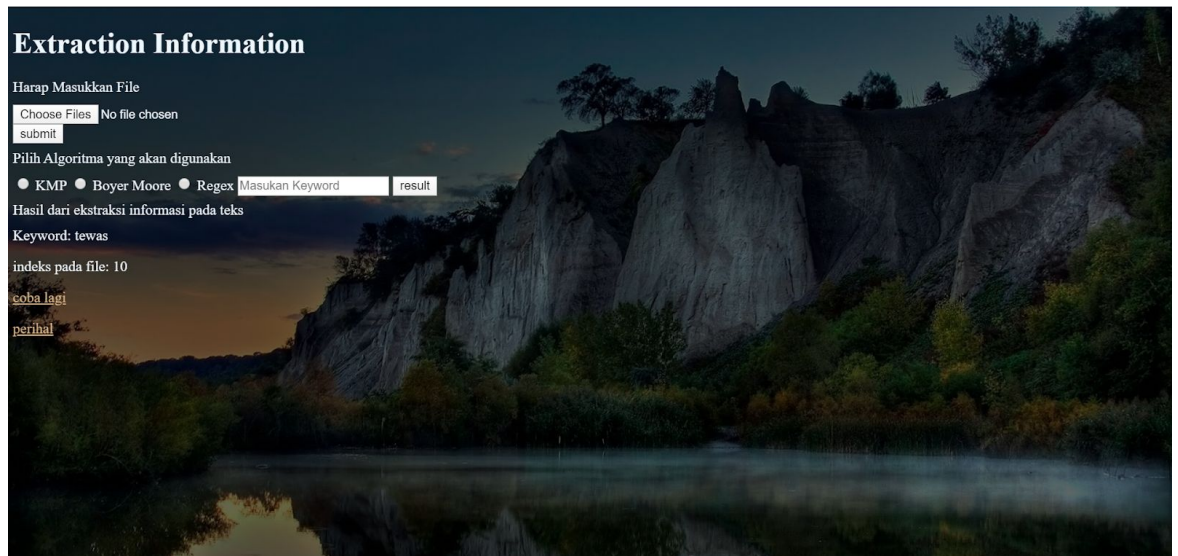
Waktu: sabtu (18/4/2020)

Kalimat: Dilansir The Economic Times, sebanyak 1.891 jiwa tewas dalam 24 jam terakhir.

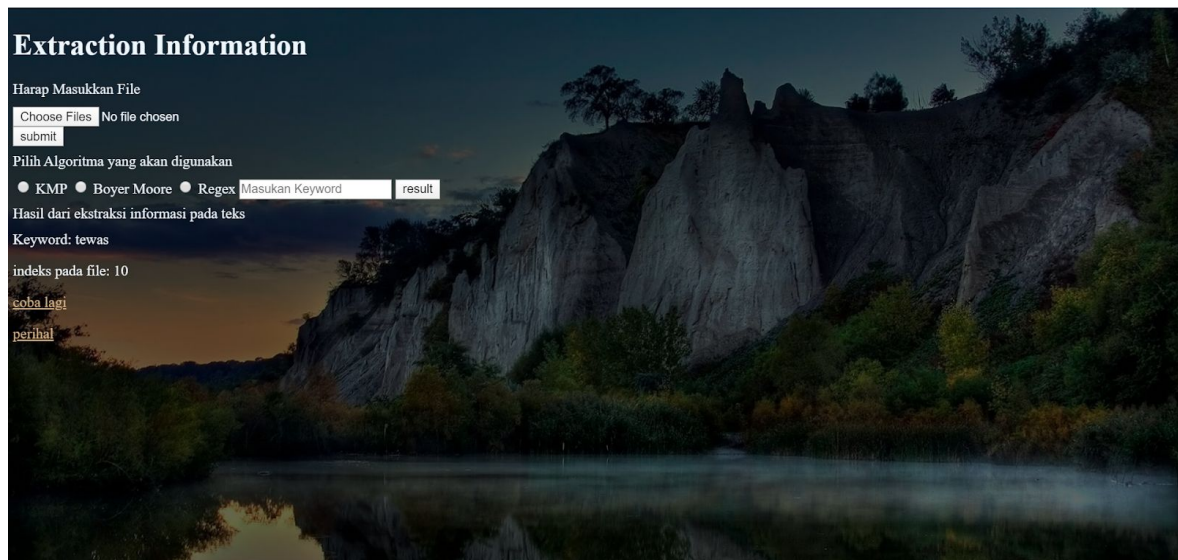
[coba lagi](#)

[perihal](#)

2. Metode KMP



3. Metode Boyer-Moore



Tambah 325, Kasus Positif Corona di Indonesia Tembus 6.248 Per 18 April 2020
 Herianto Batubara - detikNews
 Sabtu, 18 Apr 2020 15:54 WIB
 Jakarta - Kasus positif virus Corona di Indonesia mencapai angka 6.248 per hari ini Sabtu, 18 April 2020. Angka itu bertambah 325 orang.
 "Konfirmasi positif menjadi 6.248 orang," kata juru bicara pemerintah terkait penanganan wabah virus Corona, Achmad Yurianto, dalam konferensi pers yang ditayangkan YouTube BNPB, Sabtu (18/4/2020). Data tersebut dikumpulkan hingga pukul 12.00 WIB hari ini. Sebelumnya, kasus positif virus Corona sebanyak 5.923 orang per 17 April 2020. "Ada 325 kasus baru," ujar Yuri. Data kasus virus Corona diperbarui setiap hari. Warga juga dapat mengakses situs covid19.go.id untuk melihat perkembangan kasus virus Corona.

1. Metode Regex

Extraction Information

Harap Masukkan File

No file chosen

Pilih Algoritma yang akan digunakan

☒ KMP ☐ Boyer Moore ☐ Regex

Hasil dari ekstraksi informasi pada teks

Keyword: tewas

Jumlah: -

Waktu: -

Tidak ditemukan kalimat yang mengandung tewas.

[coba lagi](#)

[perihal](#)

2. Metode KMP

Extraction Information

Harap Masukkan File

No file chosen

Pilih Algoritma yang akan digunakan

☒ KMP ☐ Boyer Moore ☐ Regex

Hasil dari ekstraksi informasi pada teks

Keyword: tewas

Tidak ditemukan tewas pada file.

[coba lagi](#)

[perihal](#)

3. Metode Boyer-Moore

Extraction Information

Harap Masukkan File

No file chosen

Pilih Algoritma yang akan digunakan

☒ KMP ☐ Boyer Moore ☐ Regex

Hasil dari ekstraksi informasi pada teks

Keyword: tewas

Tidak ditemukan tewas pada file.

[coba lagi](#)

[perihal](#)

Lampiran

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