

# **LAPORAN TUGAS BESAR 2 ALJABAR LINIER DAN GEOMETRI IF2123**



**Disusun oleh:**

**Kahfi Soobhan Zulkifli (13519012)**  
**Muhammad Akram Al Bari (13519142)**  
**Jauhar Wibisono (13519160)**

**INSTITUT TEKNOLOGI BANDUNG**  
**2020**

<b>BAB I</b>	<b>3</b>
<b>DESKRIPSI MASALAH</b>	<b>3</b>
A. Deskripsi Singkat	3
B. Spesifikasi Tugas	4
<b>BAB II</b>	<b>5</b>
<b>TEORI SINGKAT</b>	<b>5</b>
A. Retrieval Information	5
B. Vektor	6
C. Cosine Similarity	7
<b>BAB III</b>	<b>8</b>
<b>IMPLEMENTASI PROGRAM</b>	<b>8</b>
A. Modul terms.py	8
B. Modul tabel.py	11
C. Modul rank.py	12
D. Modul main.py	12
<b>BAB IV</b>	<b>13</b>
<b>EKSPERIMEN</b>	<b>13</b>
<b>BAB V</b>	<b>21</b>
<b>KESIMPULAN</b>	<b>21</b>
A. Kesimpulan	21
B. Saran	21
C. Refleksi	21
<b>DAFTAR PUSTAKA</b>	<b>22</b>

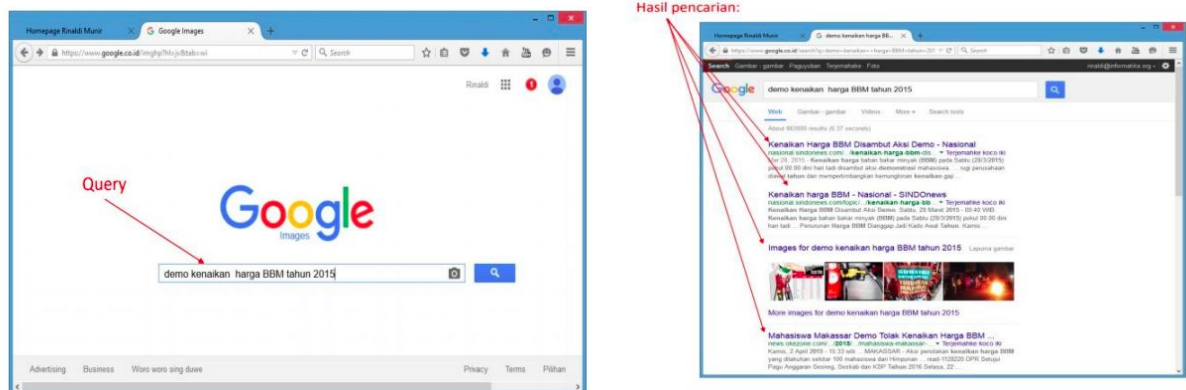
# BAB I

## DESKRIPSI MASALAH

### A. Deskripsi Singkat Abstraksi

Hampir semua dari kita pernah menggunakan search engine, seperti google, bing dan yahoo! search. Setiap hari, bahkan untuk sesuatu yang sederhana kita menggunakan mesin pencarian Tapi, pernahkah kalian membayangkan bagaimana cara search engine tersebut mendapatkan semua dokumen kita berdasarkan apa yang ingin kita cari?

Sebagaimana yang telah diajarkan di dalam kuliah pada materi vector di ruang Euclidean, temu-balik informasi (information retrieval) merupakan proses menemukan kembali (retrieval) informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Biasanya, sistem temu balik informasi ini digunakan untuk mencari informasi pada informasi yang tidak terstruktur, seperti laman web atau dokumen.



Gambar 1. Contoh penerapan Sistem Temu-Balik pada mesin pencarian

Ide utama dari sistem temu balik informasi adalah mengubah search query menjadi ruang vektor. Setiap dokumen maupun query dinyatakan sebagai vektor  $w = (w_1, w_2, \dots, w_n)$  di dalam  $R^n$ , dimana nilai  $w_i$  dapat menyatakan jumlah kemunculan kata tersebut dalam dokumen (term frequency). Penentuan dokumen mana yang relevan dengan search query dipandang sebagai pengukuran kesamaan (similarity measure) antara query dengan dokumen. Semakin sama suatu vektor dokumen dengan vektor query, semakin relevan dokumen tersebut dengan query. Kesamaan tersebut dapat diukur dengan cosine similarity dengan rumus:

$$\text{sim}(\mathbf{Q}, \mathbf{D}) = \cos \theta = \frac{\mathbf{Q} \cdot \mathbf{D}}{\|\mathbf{Q}\| \|\mathbf{D}\|}$$

Pada kesempatan ini, kalian ditantang untuk membuat sebuah search engine sederhana dengan model ruang vector dan memanfaatkan cosine similarity.

## B. Spesifikasi Tugas

Buatlah program mesin pencarian dengan sebuah website lokal sederhana. Spesifikasi program adalah sebagai berikut:

1. Program mampu menerima *search query*. *Search query* dapat berupa kata dasar maupun berimbuhan.
2. Dokumen yang akan menjadi kandidat dibebaskan formatnya dan disiapkan secara manual. Minimal terdapat 15 dokumen berbeda sebagai kandidat dokumen. **Bonus:** Gunakan web scraping untuk mengekstraksi dokumen dari website.
3. Hasil pencarian yang terurut berdasarkan similaritas tertinggi dari hasil teratas hingga hasil terbawah berupa judul dokumen dan kalimat pertama dari dokumen tersebut. Sertakan juga nilai similaritas tiap dokumen.
4. Program disarankan untuk melakukan pembersihan dokumen terlebih dahulu sebelum diproses dalam perhitungan cosine similarity. Pembersihan dokumen bisa meliputi hal-hal berikut ini:
  - a. Stemming dan Penghapusan stopwords dari isi dokumen.
  - b. Penghapusan karakter-karakter yang tidak perlu.
5. Program dibuat dalam sebuah website lokal sederhana. Dibebaskan untuk menggunakan *framework* pemrograman website apapun. Salah satu *framework* website yang bisa dimanfaatkan adalah Flask (Python), ReactJS, dan PHP.
6. Kalian dapat menambahkan fitur fungsional lain yang menunjang program yang anda buat (unsur kreativitas diperbolehkan/dianjurkan).
7. Program harus modular dan mengandung komentar yang jelas.
8. Dilarang menggunakan library cosine similarity yang sudah jadi.

## BAB II

# TEORI SINGKAT

### A. Retrieval Information

Sistem temu balik informasi (*information retrieval*) digunakan untuk menemukan kembali informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Salah satu aplikasi umum dari sistem temu kembali informasi adalah *search-engine* atau mesin pencarian yang terdapat pada jaringan internet. Pengguna dapat mencari halaman-halaman Web yang dibutuhkannya melalui mesin tersebut.

Ukuran efektivitas dalam pencarian ditentukan oleh faktor *precision* dan *recall*. *Precision* adalah rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen yang ditemukan oleh *search-engine*. Sedangkan *recall* adalah rasio jumlah dokumen relevan yang ditemukan kembali dengan total jumlah dokumen dalam kumpulan dokumen yang dianggap relevan.

Dalam sistem temu balik informasi, hanya mendapatkan dokumen yang relevan saja tidaklah cukup. Tujuan yang harus dipenuhi adalah bagaimana mendapatkan dokumen yang relevan dan tidak mendapatkan dokumen yang tidak relevan. Tujuan lainnya adalah bagaimana menyusun dokumen yang telah didapatkan agar bisa ditampilkan terurut berdasarkan tingkat relevansi yang lebih tinggi ke tingkat relevansi yang lebih rendah. Penyusunan dokumen tersebut disebut sebagai perankingan dokumen. Model Ruang Vektor dan Model Probabilistik adalah 2 model pendekatan untuk melakukan hal tersebut.

Model ruang vektor dan model probabilistik adalah model yang menggunakan pembobotan kata dan perankingan dokumen. Dalam model ruang vektor, dokumen dan *query* direpresentasikan sebagai vektor dalam ruang vektor yang disusun dalam indeks term, kemudian dimodelkan dengan persamaan geometri. Sedangkan model probalistik membuat asumsi-asumsi distribusi term dalam dokumen relevan dan tidak relevan dalam orde estimasi kemungkinan relevansi suatu dokumen terhadap suatu *query*. Dalam pengerjaan tugas ini, akan digunakan pendekatan dalam model ruang vektor untuk melakukan temu balik informasi.

### B. Vektor

Vektor adalah objek geometri yang memiliki besaran dan arah. Vektor dilambangkan dengan tanda panah ( $\rightarrow$ ). Besar vektor proporsional dengan panjang panah dan arahnya bertepatan dengan arah panah. Vektor dapat melambangkan perpindahan dari suatu titik A ke B.

Untuk mencari panjang sebuah vektor dalam ruang euklidian  $n$  dimensi, dapat digunakan cara berikut:

$$||a|| = \sqrt{(a_1 * a_1) + (a_2 * a_2) + \dots + (a_n * a_n)}$$

Dalam pembahasan mengenai vektor, kita juga mengenal istilah ruang vektor. Ruang vektor adalah struktur matematika yang dibentuk oleh sekumpulan vektor, yaitu objek yang dapat dijumlahkan dan dikalikan dengan suatu bilangan, yang dinamakan skalar. Skalar

umumnya adalah bilangan riil, tetapi kita juga dapat merumuskan ruang vektor dengan perkalian skalar dengan bilangan kompleks, bilangan rasional, atau bahkan meda. Operasi penjumlahan dan perkalian vektor mesti memenuhi persyaratan tertentu yang dinamakan aksioma. Contoh ruang vektor adalah vektor Euklides yang sering digunakan untuk melambangkan besaran fisika seperti gaya. Dua gaya dengan jenis sama dapat dijumlahkan untuk menghasilkan gaya ketiga, dan perkalian vektor gaya dengan bilangan riil adalah vektor gaya lain. Vektor yang melambangkan perpindahan pada bidang atau pada ruang tiga dimensi juga membentuk ruang vektor.

Ruang vektor merupakan subjek dari aljabar linear, dan dipahami dengan baik dari sudut pandang ini, karena ruang vektor dicirikan oleh dimensinya, yang menspesifikasikan banyaknya arah independen dalam ruang. Teori ruang vektor juga ditingkatkan dengan memperkenalkan struktur tambahan, seperti norma atau hasil kali dalam. Ruang seperti ini muncul dengan alamiah dalam analisis matematika, dalam bentuk ruang fungsi berdimensi takhingga, dengan vektornya adalah fungsi.

Saat ini ruang vektor diterapkan di seluruh bidang matematika, sains dan rekayasa. Ruang vektor adalah konsep aljabar linear yang sesuai untuk menghadapi sistem persamaan linear, menawarkan kerangka kerja untuk deret Fourier (yang digunakan dalam pemampatan citra), atau menyediakan lingkungan yang dapat digunakan untuk teknik solusi persamaan diferensial parsial. Ruang vektor juga memberikan cara abstrak dan bebas koordinat untuk berurusan dengan objek geometris dan fisis seperti tensor. Ruang vektor dapat dirampatkan ke beberapa arah, dan menghasilkan konsep lebih lanjut dalam geometri dan aljabar abstrak.

### C. Cosine Similarity

Cosine similarity adalah ukuran kemiripan (*similarity*) diantara dua vektor tidak nol dalam ruang perkalian dalam. Cosine similarity didefinisikan sebagai nilai cosinus dari sudut yang dibentuk diantara dua vektor tersebut, yang juga bernilai sama dengan perkalian dalam dari kedua vektor dalam bentuk vektor satuannya. Nilai dari cosinus(0) adalah 1, dan akan bernilai kurang dari 1 untuk setiap sudut dalam interval  $(0, \pi]$  radian. Maka dari itu, cosine similarity ditinjau berdasarkan orientasi vektor, bukan besar dari vektor.

Cosine similarity biasa digunakan dalam ruang besaran positif, di mana hasilnya pasti akan berada dalam interval  $[0, 1]$ . Interval ini berlaku untuk berapapun jumlah dimensi ruang vektor, dan cosine similarity paling umum digunakan dalam ruang vektor positif yang memiliki dimensi banyak. Sebagai contoh, cosine similarity digunakan dalam sistem temu baik informasi dan *text mining*.

Salah satu kelebihan dari cosine similarity adalah ia tergolong *low-complexity*, terlebih lagi untuk vektor sparse, karena hanya dimensi tidak nol saja yang perlu diperhatikan.

Cosine dari dua vektor tidak nol dapat dihitung dengan menggunakan rumus *Euclidian dot product* :

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \|\mathbf{B}\| \cos \theta$$

Bila diberikan dua vektor,  $A$  dan  $B$  maka nilai cosine similarity,  $\cos(\theta)$ , direpresentasikan dengan menggunakan perkalian dot product dan norma sebagai berikut:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

Hasil perhitungan dengan rumus di atas memiliki range hasil  $[-1,1]$ . Di mana  $-1$  artinya adalah tepat berlawanan,  $1$  artinya sama persis,  $0$  mengindikasikan ortogonalitas, dan nilai diantaranya mengindikasikan tingkat similarity atau dissimilarity.

Dalam kasus *information retrieval*, cosine similarity dari dua dokumen akan memiliki range nilai  $[0,1]$ , karena frekuensi kemunculan term tidak bisa negatif. Sudut diantara kedua vektor frekuensi tidak dapat lebih besar daripada  $90^\circ$ .

## BAB III

### IMPLEMENTASI PROGRAM

#### A. Modul *terms.py*

Pada modul *terms.py*, berisi fungsi-fungsi yang melakukan pengolahan data yang berkaitan dengan ke-linguistik-an. Mulai dari fungsi untuk melakukan *web scraping*, hingga pengolahan input query dan file dokumen agar siap untuk digunakan dalam algoritma sistem temu balik informasi. Fungsi-fungsi yang ada pada modul ini:

Fungsi	Spesifikasi
<b>WebScrapingKontenByUrl</b> (url)	Fungsi ini adalah fungsi yang digunakan untuk melakukan Web Scraping dan mengambil konten berita dari suatu laman dalam website. Fungsi ini menggunakan library <i>newspaper3k</i> dan untuk menggunakan fungsi ini, cukup masukan <i>url</i> dari laman web yang ingin discrap kontennya. Fungsi ini akan mengembalikan hasil scraping dari konten website dalam bentuk <b>string</b> .
<b>FormatNamaFile</b> (namaFile)	Fungsi ini berguna untuk melakukan pengecekan apakah input nama file yang dimasukkan oleh pengguna sudah tepat atau belum. Format yang diterima dalam program ini hanyalah format file <b>.txt</b> . Argumen yang diterima dalam fungsi ini adalah argumen nama file dalam bentuk <b>string</b> .
<b>SaveKontenTxt</b> (namaFile, konten)	Digunakan untuk melakukan save file ke dalam perangkat. User memasukkan argumen berupa nama dari file yang digunakan untuk menyimpan konten, dan juga memasukkan argumen kedua berupa konten yang ingin di save ke dalam file. Kedua argumen ini bertipe <b>string</b> .
<b>BacaKontenTxt</b> (namaFile)	Digunakan untuk melakukan pembacaan isi dari suatu file <b>.txt</b> . Argumen yang dimasukan adalah nama dari file yang ingin dibaca isinya. Argumen serta hasil kembali dari fungsi ini bertipe <b>string</b> .
<b>RegexCleaning</b> (stringKotor)	Berfungsi untuk melakukan <i>pembersihan</i> terhadap suatu <b>string</b> untuk menghilangkan karakter-karakter yang



	<p>tidak diperlukan dalam proses pengolahan selanjutnya. Fungsi ini memanfaatkan modul <i>re</i>, yakni modul <i>regular expressions</i>. Cara kerjanya adalah dengan mencocokkan input string yang ada dengan pola-pola tertentu, dan apabila pola yang ingin dihilangkan ditemukan, maka fungsi ini akan menghilangkan bagian tersebut. Termasuk yang <i>dibersihkan</i> di dalam fungsi ini adalah karakter-karakter <i>unicode</i>, membuat semua kata menjadi lowercase, dan menghilangkan tanda baca.</p>
<code>StringToArray(stringAwal)</code>	<p>Fungsi yang digunakan untuk mengubah suatu <b>string</b> yang panjang menjadi potongan-potongan kata dalam <b>string</b> dan masing-masing kata disimpan sebagai satu buah elemen array. Fungsi ini memanfaatkan library <i>nlTK</i> dan menggunakan fungsi <i>word_tokenize</i>. Hasil dari fungsi adalah sebuah <b>array of string</b>.</p>
<code>StemmingKonten(arrayOfKata)</code>	<p>Fungsi ini melakukan <i>stemming</i> terhadap suatu <b>array of string</b>. Stemming dilakukan dengan memanfaatkan library <i>nlTK</i>, dan digunakan jenis <i>PorterStemmer</i>.</p>
<code>StopWordsRemove(arrayOfKata)</code>	<p>Melakukan penghapusan <i>stop words</i> yang ada pada suatu <b>array of karakter</b>. Hasil kembaliannya adalah array yang sama namun tanpa elemen <i>stop words</i>. Library <i>nlTK</i> juga digunakan pada fungsi ini.</p>
<code>firstsentence()</code>	<p>Fungsi untuk mengambil kalimat pertama dari suatu <b>string</b> yang panjang. Fungsi <i>sent_tokenize</i> yang merupakan bawaan dari library <i>nlTK</i> digunakan pada fungsi ini.</p>
<code>ArrayIsiFileSiapOlah(namaFile)</code>	<p>Merupakan fungsi gabungan dari fungsi-fungsi kecil sebelumnya. Input berupa namaFile yang akan dibaca kontennya dan kemudian <i>dibersihkan</i> hingga siap untuk diolah. Hasil akhirnya adalah <b>array of string</b>.</p>
<code>ArrayIsiFileSiapOlahNonStem(namaFile)</code>	<p>Sama seperti fungsi di atas, namun fungsi</p>

	ini tidak melakukan <i>stemming</i> terhadap isi file yang akan diolah
KataDalamKamus()	Membuat himpunan kata dari semua dokumen yang ada. Himpunan kata ini akan berperan sebagai <i>ruang vektor</i> yang akan digunakan dalam perhitungan <i>cosine similarity</i> .
KataDalamKamusNonStem()	Sama seperti atas, namun yang diolah adalah kata-kata yang tidak dilakukan proses <i>stemming</i> .
KataDalamDokumen()	Melakukan pembacaan terhadap semua isi file yang ada pada folder <i>test</i> . Hasil pembacaan berupa <b>array of string</b> yang setiap elemennya adalah setiap kata yang ada pada tiap dokumen. <b>array of string</b> tersebut disimpan lagi dalam suatu array, sehingga hasil akhirnya adalah <b>array of array of string</b>
KataDalamDokumenNonStem()	Sama dengan atas, namun yang diproses adalah kata-kata yang tidak distem
KamusDokumen()	Fungsi ini mengembalikan hasil akhir pengolahan dalam bentuk <b>dictionary</b> yang memiliki format <i>key-value</i> = <code>{{kata,dokumen}: frekuensiKata}</code> .
KataDalamDokumen2()	Sama seperti fungsi <i>KataDalamDokumen</i> , namun format kembaliannya berupa <b>dictionary</b> yang isinya <b>array</b> .
KataDalamDokumenNonStem2()	Sama dengan atas, namun tanpa melalui proses <i>stemming</i> .
KamusDokumen2(namaFiles, kamusKata, kataDiDokumen)	Sama seperti fungsi <i>KamusDokumen</i> , namun kembaliannya dalam format <b>dictionary</b> yang berisi <b>array</b> .
GetTermsStem(s)	Fungsi yang menerima argumen berupa <b>string</b> query dari pengguna yang kemudian akan <i>dibersihkan</i> sehingga siap untuk diolah.
GetTermsNonStem(s)	Sama seperti fungsi di atas, namun tidak melakukan proses <i>stemming</i> .

### B. Modul *tabel.py*

modul ini hanya memiliki satu fungsi saja yang digunakan untuk mendapatkan perhitungan kemunculan kata-kata yang ada pada query di setiap dokumen yang dibaca.

Fungsi	Spesifikasi
<code>GetTabel(query)</code>	Fungsi ini merupakan fungsi yang menerima argumen berupa <i>query</i> dari pengguna dalam bentuk <b>string</b> . <i>Query</i> kemudian diolah dan dicocokkan dengan hasil Kamus Kata (ruang vektor) yang dibentuk oleh tiap-tiap dokumen. Kemudian, hasil akhir dari fungsi ini adalah dictionary yang memiliki format <code>{(kataPadaQuery, dokumen): frekuensiKemunculanKata}</code> . Pengolahan data pada fungsi ini memanfaatkan fungsi-fungsi yang telah direalisasikan dalam modul <i>terms.py</i>

### C. Modul *rank.py*

Modul ini adalah modul yang digunakan untuk melakukan perhitungan *cosine similarity* dan melakukan perangkingan terhadap kecocokan setiap isi dokumen dengan query yang dimasukkan. Modul ini juga hanya memiliki satu fungsi yakni:

Fungsi	Spesifikasi
<code>GetRank(query)</code>	Fungsi <i>GetRank</i> menerima argumen berupa <b>string</b> <i>query</i> yang dimasukkan oleh pengguna. Dalam fungsi <i>GetRank</i> , dilakukan pengolahan data dengan memanfaatkan fungsi yang telah direalisasikan pada modul <i>terms.py</i> dan <i>tabel.py</i> . Kemudian, hasil olahan data tersebut dihitung dengan menggunakan <i>cosine similarity</i> dari <i>query</i> yang dimasukkan pengguna terhadap isi dari tiap dokumen. Hasil akhir dari fungsi ini adalah <b>array of tuple</b> yang telah diurutkan berdasarkan nilai <i>cosine similarity</i> , terurut mengecil. Isi dari <b>tuple</b> adalah nama file, banyak kata dalam file, hasil nilai <i>cosine similarity</i> serta kalimat pertama yang ada pada setiap file.

#### D. Modul *main.py*

modul *main.py* adalah modul program utama yang akan di-*execute* untuk menjalankan program yang telah dibuat. Modul ini melakukan import dari semua modul-modul sebelumnya, serta mengimport library *flask* yang merupakan *web framework* dengan basis bahasa Python. Dalam modul ini kita mengenal istilah *view function*, yakni fungsi yang akan merespons terhadap *request* yang dibuat melalui aplikasi web yang sudah dijalankan. Fungsi-fungsi tersebut ialah:

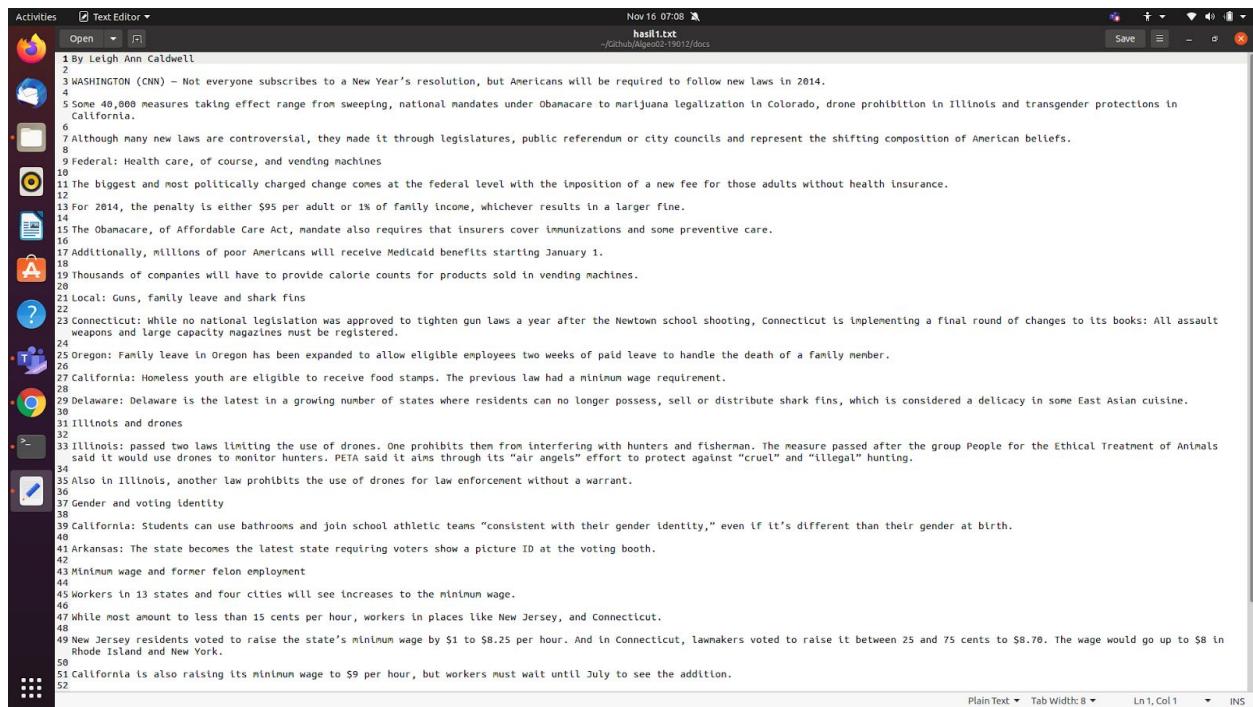
View Function	Spesifikasi
<code>HomePage()</code>	Merupakan <i>view function</i> (selanjutnya akan disebut sebagai <i>fungsi</i> ) yang akan membuka laman awal dari aplikasi web yang telah dibuat. Dalam laman awal ini, terdapat pilihan-pilihan untuk melakukan <i>upload file</i> , memasukkan <i>query</i> , melihat laman perihal, dan pindah ke laman untuk melakukan <i>web scraping</i> .
<code>SearchResultsPage(query)</code>	Fungsi yang akan dijalankan ketika pengguna memasukkan query dan telah menekan tombol <i>search</i> . Akan menampilkan semua dokumen yang ada pada folder, terurut mengecil berdasarkan kecocokan hasil <i>cosine similarity</i> . Pada laman ini, ditampilkan pula kalimat pertama dari setiap file, serta pengguna dapat melakukan klik ke nama file untuk melihat isi file secara menyeluruh.
<code>DisplayPage(filename)</code>	Fungsi yang dijalankan ketika pengguna melakukan klik ke salah satu nama file yang tersedia. Pengguna akan dialihkan ke laman baru yang berisi teks dari dokumen yang dipilih pengguna
<code>PerihalPage()</code>	Fungsi yang dijalankan ketika pengguna menekan tulisan perihal pada web. Akan ditampilkan laman perihal yang berisi deskripsi singkat mengenai tugas ini dan juga tentang kelompok yang mengerjakannya.
<code>WebScrapingPage()</code>	Fungsi yang dipanggil ketika pengguna memilih opsi <i>web scraping</i> . Pada laman ini, akan menerima input berupa alamat url laman yang ingin di <i>web scraping</i> dan nama file dari pengguna untuk

menyimpan hasil dari *web scraping*

## BAB IV

# EKSPERIMEN

Berikut contoh dokumen yang digunakan.



The screenshot shows a Linux desktop with a text editor window titled 'hasil.txt' open. The window displays a list of news items, each preceded by a line number. The news items are:

- 1 By Leigh Ann Caldwell
- 2 WASHINGTON (CNN) – Not everyone subscribes to a New Year’s resolution, but Americans will be required to follow new laws in 2014.
- 4 Some 40,000 measures taking effect range from sweeping, national mandates under Obamacare to marijuana legalization in Colorado, drone prohibition in Illinois and transgender protections in California.
- 6
- 7 Although many new laws are controversial, they made it through legislatures, public referendum or city councils and represent the shifting composition of American beliefs.
- 8
- 9 Federal: Health care, of course, and vending machines
- 10
- 11 The biggest and most politically charged change comes at the federal level with the imposition of a new fee for those adults without health insurance.
- 12
- 13 For 2014, the penalty is either \$95 per adult or 1% of family income, whichever results in a larger fine.
- 14
- 15 The Obamacare, of Affordable Care Act, mandate also requires that insurers cover immunizations and some preventive care.
- 16
- 17 Additionally, millions of poor Americans will receive Medicaid benefits starting January 1.
- 18
- 19 Thousands of companies will have to provide calorie counts for products sold in vending machines.
- 20
- 21 Local: Guns, family leave and shark fins
- 22
- 23 Connecticut: While no national legislation was approved to tighten gun laws a year after the Newtown school shooting, Connecticut is implementing a final round of changes to its books: All assault weapons and large capacity magazines must be registered.
- 24
- 25 Oregon: Family leave in Oregon has been expanded to allow eligible employees two weeks of paid leave to handle the death of a family member.
- 26
- 27 California: Homeless youth are eligible to receive food stamps. The previous law had a minimum wage requirement.
- 28
- 29 Delaware: Delaware is the latest in a growing number of states where residents can no longer possess, sell or distribute shark fins, which is considered a delicacy in some East Asian cuisine.
- 30
- 31 Illinois and drones
- 32
- 33 Illinois: passed two laws limiting the use of drones. One prohibits them from interfering with hunters and fishermen. The measure passed after the group People for the Ethical Treatment of Animals said it would use drones to monitor hunters. PETA said it aims through its “air angels” effort to protect against “cruel” and “illegal” hunting.
- 34
- 35 Also in Illinois, another law prohibits the use of drones for law enforcement without a warrant.
- 36
- 37 Gender and voting identity
- 38
- 39 California: Students can use bathrooms and join school athletic teams “consistent with their gender identity,” even if it’s different than their gender at birth.
- 40
- 41 Arkansas: The state becomes the latest state requiring voters show a picture ID at the voting booth.
- 42
- 43 Minimum wage and former felon employment
- 44
- 45 Workers in 13 states and four cities will see increases to the minimum wage.
- 46
- 47 While most amount to less than 15 cents per hour, workers in places like New Jersey, and Connecticut.
- 48
- 49 New Jersey residents voted to raise the state’s minimum wage by \$1 to \$8.25 per hour. And in Connecticut, lawmakers voted to raise it between 25 and 75 cents to \$8.70. The wage would go up to \$8 in Rhode Island and New York.
- 50
- 51 California is also raising its minimum wage to \$9 per hour, but workers must wait until July to see the addition.
- 52

The text editor window has a menu bar with 'Open', 'Save', and 'Print' options. The status bar at the bottom shows 'Plain Text', 'Tab Width: 8', 'Ln 1, Col 1', and 'INS'.



Berikut query dan hasil query yang didapat:

# 1) Query = seluruh isi salah satu dokumen

Activities Google Chrome Nov 16 07:10

Laporan Tubes 2 Ajabar L... 127.0.0.1:5000/search-in... x +

← → 127.0.0.1:5000/search-in%20the%20final%2C%20of%20furious%20days%20of%20his%20reelection%20campaign%2C%20President%20Donald%20Trump%20often%20turned%20his%20public%20rallies%20into%20personal%20therapy%20s... ☆

Apps Gmail YouTube Maps Gmail YouTube Maps ABSEN BOSS... The Data Scien... nim finder ter... Yaumul hisab Github DSC ITB

## My Simple Search Engine

Upload dokumen (.txt):  No file chosen

In the final, furious days of his

1. [sample.txt](#)  
Jumlah Kata: 669  
Tingkat Kemiripan: 100.0%  
In the final, furious days of his reelection campaign, President Donald Trump often turned his public rallies into personal therapy sessions, at which the embattled and embittered President rued what might have been.
2. [hasil1.txt](#)  
Jumlah Kata: 371  
Tingkat Kemiripan: 14.98%  
By Leigh Ann Caldwell WASHINGTON (CNN) — Not everyone subscribes to a New Year's resolution, but Americans will be required to follow new laws in 2014.
3. [hasil2.txt](#)  
Jumlah Kata: 358  
Tingkat Kemiripan: 0.54%  
SALT LAKE CITY — Governor Gary Herbert told FOX 13 he anticipates making recommendations this week for Utahns on how best to handle the upcoming holiday season, starting with Thanksgiving.

Term	Query	sample.txt	hasil1.txt	hasil2.txt
final	2	2	1	0
furious	1	1	0	0
days	2	2	0	2
reelection	3	3	0	0
campaign	2	2	0	0
president	10	10	0	0

Activities Google Chrome Nov 16 07:12

Laporan Tubes 2 Ajabar L... 127.0.0.1:5000/search-in... x +

← → 127.0.0.1:5000/search-in%20the%20final%2C%20of%20furious%20days%20of%20his%20reelection%20campaign%2C%20President%20Donald%20Trump%20often%20turned%20his%20public%20rallies%20into%20personal%20therapy%20s... ☆

Apps Gmail YouTube Maps Gmail YouTube Maps ABSEN BOSS... The Data Scien... nim finder ter... Yaumul hisab Github DSC ITB

president	10	10	0	0
donald	3	3	0	0
trump	24	24	0	0
often	1	1	0	0
turned	2	2	0	0
public	2	2	1	2
rallies	1	1	0	0
personal	1	1	0	0
therapy	1	1	0	0
sessions	1	1	0	0
embattled	1	1	0	0
embittered	1	1	0	0
rued	1	1	0	0
might	1	1	0	0
four	3	3	1	0
five	1	1	0	0
months	1	1	0	0
ago	4	4	0	0
started	1	1	0	0
whole	1	1	0	0
thing	2	2	0	0
plague	1	1	0	0
came	2	2	0	0
made	3	3	1	0
coming	1	1	0	2
erie	2	2	0	0
told	1	1	0	2
late	1	1	0	0
october	1	1	0	0

## 2) Query = donald trump

The screenshot shows a web browser window with the address bar displaying '127.0.0.1:5000/search-donaldtrump'. The page title is 'My Simple Search Engine'. Below the title, there is an 'Upload dokumen (.txt):' section with a 'Choose File' button and a 'Submit' button. The search results are displayed as follows:

1. [sample.txt](#)  
Jumlah Kata: 669  
Tingkat Kemiripan: 39.58%  
In the final, furious days of his reelection campaign, President Donald Trump often turned his public rallies into personal therapy sessions, at which the embattled and embittered President rued what might have been.

2. [hasil1.txt](#)  
Jumlah Kata: 371  
Tingkat Kemiripan: 0.0%  
By Leigh Ann Caldwell WASHINGTON (CNN) — Not everyone subscribes to a New Year's resolution, but Americans will be required to follow new laws in 2014.

3. [hasil2.txt](#)  
Jumlah Kata: 358  
Tingkat Kemiripan: 0.0%  
SALT LAKE CITY — Governor Gary Herbert told FOX 13 he anticipates making recommendations this week for Utahns on how best to handle the upcoming holiday season, starting with Thanksgiving.

Term	Query	sample.txt	hasil1.txt	hasil2.txt
donald	1	3	0	0
trump	1	24	0	0

• [Perihal](#)

## 3) Query = hahaha

The screenshot shows a web browser window with the address bar displaying '127.0.0.1:5000/search-hahaha'. The page title is 'My Simple Search Engine'. Below the title, there is an 'Upload dokumen (.txt):' section with a 'Choose File' button and a 'Submit' button. The search results are displayed as follows:

1. [sample.txt](#)  
Jumlah Kata: 669  
Tingkat Kemiripan: 0.0%  
In the final, furious days of his reelection campaign, President Donald Trump often turned his public rallies into personal therapy sessions, at which the embattled and embittered President rued what might have been.

2. [hasil1.txt](#)  
Jumlah Kata: 371  
Tingkat Kemiripan: 0.0%  
By Leigh Ann Caldwell WASHINGTON (CNN) — Not everyone subscribes to a New Year's resolution, but Americans will be required to follow new laws in 2014.

3. [hasil2.txt](#)  
Jumlah Kata: 358  
Tingkat Kemiripan: 0.0%  
SALT LAKE CITY — Governor Gary Herbert told FOX 13 he anticipates making recommendations this week for Utahns on how best to handle the upcoming holiday season, starting with Thanksgiving.

Term	Query	sample.txt	hasil1.txt	hasil2.txt
hahaha	1	0	0	0

• [Perihal](#)

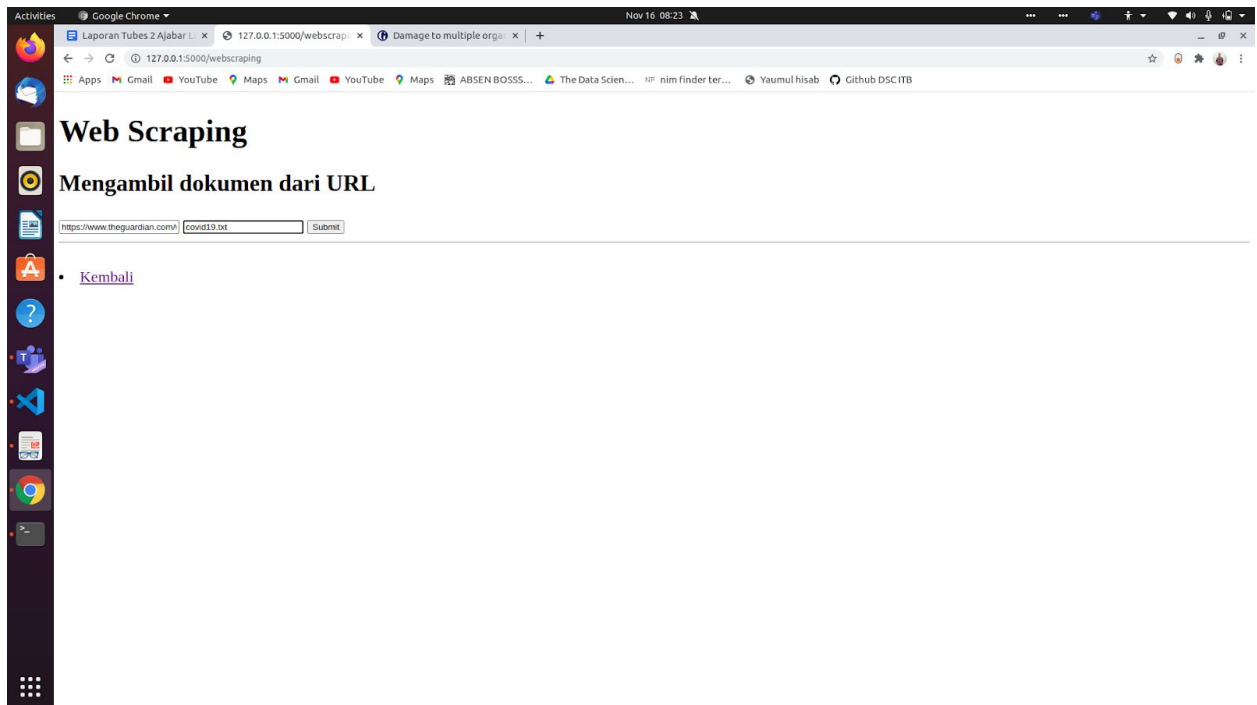
Berikut adalah implementasinya dengan web scraping

### 1) Masukkan url dan nama file yang diinginkan. Pastikan menggunakan .txt. URL:

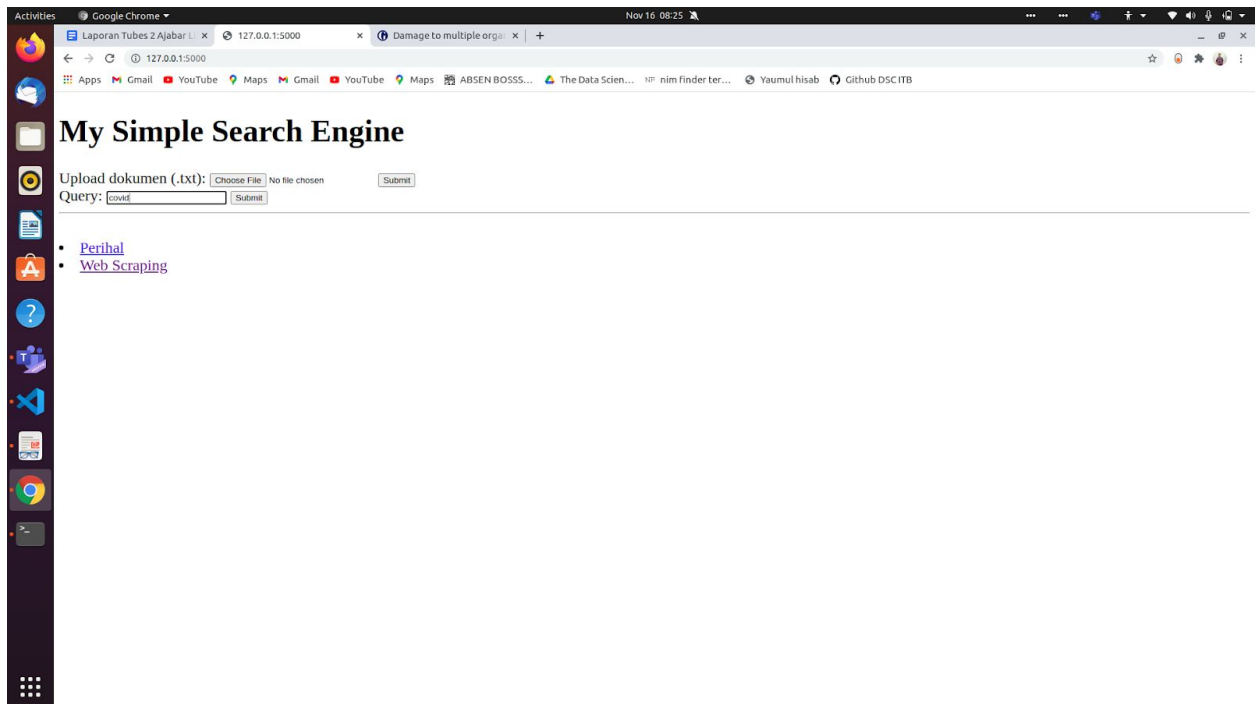
<https://www.theguardian.com/world/2020/nov/15/damage-to-multiple-organs-recorded-in->



[long-covid-cases](#). Nama file: covid19.txt



2) Klik kembali lalu masukkan query yang diinginkan. Query: covid.



### 3) Berikut hasilnya

The screenshot shows a web browser window with the title 'My Simple Search Engine'. The address bar shows the URL '127.0.0.1:5000/search-covid'. The search bar contains the query 'covid'. Below the search bar, there are six search results listed, each with a document name, word count, and similarity percentage. The results are as follows:

Term	Query	hasil2.txt	sample.txt	joe Biden.txt	covid19.txt	test.txt	hasil1.txt
covid		1	4	3	1	11	0

Below the table, there is a link labeled 'Perihal'.

Kemudian, ini dicoba dengan 15 dokumen, dengan webscraping dari laman the guardian sport serta query: lewis hamilton. Berikut hasilnya

Activities

Google Chrome

Nov 16 08:35

127.0.0.1:5000/search-lew... Sport news, comment an... Dominic Thiem beats Ste... Laporan Tubes 2 Ajabar...

127.0.0.1:5000/search-lewis%20hamilton

Apps Gmail YouTube Maps Gmail YouTube Maps ABSEN BOSS5... The Data Scien... nim finder ter... Yaumul hisab Github DSC ITB

Update

# My Simple Search Engine

Upload dokumen (.txt):

Query:

1. [coba7.txt](#)

Jumlah Kata: 418

Tingkat Kemiripan: 34.51%

Lewis Hamilton equalled Michael Schumacher's record of seven world championships on Sunday by winning the Turkish Grand Prix and said afterwards that he hoped his success as a black man in an almost exclusively white sport would act as inspiration to children everywhere.

2. [coba2.txt](#)

Jumlah Kata: 550

Tingkat Kemiripan: 0.0%

The English players suddenly disappeared in the final stages of the Indian Premier League.

3. [coba6.txt](#)

Jumlah Kata: 556

Tingkat Kemiripan: 0.0%

The highest compliment due to Dustin Johnson is that he did not allow the final round of the 84th Masters to become particularly interesting.

4. [coba9.txt](#)

Jumlah Kata: 429

Tingkat Kemiripan: 0.0%

So, this is the way the ATP Finals end.

5. [coba5.txt](#)

Jumlah Kata: 592

Tingkat Kemiripan: 0.0%

Ray Clemence always seemed to be in the right place at the right time Ray Clemence was not just one of the finest goalkeepers England has ever produced, he was a link back to the time when footballers used to have day jobs or summer occupations.

6. [test.txt](#)

Jumlah Kata: 20

Tingkat Kemiripan: 0.0%

The team behind our US election results tracker discuss how it came together, why readers around the world loved it, and how it came to be the most-viewed page ever on the Guardian's website

7. [coba4.txt](#)

Jumlah Kata: 269

Tingkat Kemiripan: 0.0%

Gareth Southgate insisted England could draw encouragement from their display against Belgium and the performance of Jack Grealish, despite their interest in this season's Nations League ending with a 2-0 defeat in Leuven.

8. [sample.txt](#)

Jumlah Kata: 669

Tingkat Kemiripan: 0.0%

In the final, furious days of his reelection campaign, President Donald Trump often turned his public rallies into personal therapy sessions, at which the embattled and embittered President rued what might have been.

9. [coba3.txt](#)

Jumlah Kata: 474

Tingkat Kemiripan: 0.0%

All runners but one disqualified in Fontwell race for going around hurdle Racing's seemingly fathomless ability to generate contentious sideshows was once more on show at Fontwell on Sunday, where the final race descended into farce, all the runners but one being disqualified for going around a hurdle they should have jumped.

Activities

Google Chrome

Nov 16 08:35

127.0.0.1:5000/search-le... Sport news, comment an... Dominic Thiem beats Ste... Laporan Tubes 2 Ajabar...

127.0.0.1:5000/search-lewis%20hamilton

Apps Gmail YouTube Maps Gmail YouTube Maps ABSEN BOSS5... The Data Scien... nim finder ter... Yaumul hisab Github DSC ITB

Update

8. [sample.txt](#)

Jumlah Kata: 669

Tingkat Kemiripan: 0.0%

In the final, furious days of his reelection campaign, President Donald Trump often turned his public rallies into personal therapy sessions, at which the embattled and embittered President rued what might have been.

9. [coba3.txt](#)

Jumlah Kata: 474

Tingkat Kemiripan: 0.0%

All runners but one disqualified in Fontwell race for going around hurdle Racing's seemingly fathomless ability to generate contentious sideshows was once more on show at Fontwell on Sunday, where the final race descended into farce, all the runners but one being disqualified for going around a hurdle they should have jumped.

10. [hasil1.txt](#)

Jumlah Kata: 371

Tingkat Kemiripan: 0.0%

By Leigh Ann Caldwell WASHINGTON (CNN) — Not everyone subscribes to a New Year's resolution, but Americans will be required to follow new laws in 2014.

11. [joebiden.txt](#)

Jumlah Kata: 382

Tingkat Kemiripan: 0.0%

The Democrat scored a 396-232 electoral college victory but president writes 'he won because the election was rigged' Sign up for the Guardian's First Thing newsletter On Sunday morning, Donald Trump tweeted about Joe Biden.

12. [coba1.txt](#)

Jumlah Kata: 441

Tingkat Kemiripan: 0.0%

'Nobody has ever done that to me': Kell Brook well beaten by Terence Crawford Terence Crawford started slow and finished fast, stopping Kell Brook with a barrage of punches in the fourth round Saturday night to retain his welterweight title in Las Vegas.

13. [hasil2.txt](#)

Jumlah Kata: 358

Tingkat Kemiripan: 0.0%

SALT LAKE CITY — Governor Gary Herbert told FOX 13 he anticipates making recommendations this week for Utahns on how best to handle the upcoming holiday season, starting with Thanksgiving.

14. [coba8.txt](#)

Jumlah Kata: 279

Tingkat Kemiripan: 0.0%

Eddie Jones has warned his players of the "massive step-up" they face against Ireland on Saturday, suggesting England are at a disadvantage because of the lack of high-quality opposition they have faced this autumn.

15. [covid19.txt](#)

Jumlah Kata: 343

Tingkat Kemiripan: 0.0%

Young and previously healthy people with ongoing symptoms of Covid-19 are showing signs of damage to multiple organs four months after the initial infection, a study suggests.

Term	Query	coba7.txt	coba2.txt	coba6.txt	coba9.txt	coba5.txt	test.txt	coba4.txt	sample.txt	coba3.txt	hasil1.txt	joebiden.txt	coba1.txt	hasil2.txt	coba8.txt	covid19.txt
lewis	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hamilton	1	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Perihal

Web Scraping

19 | Tubes Algeo IF 2123

# BAB V

## KESIMPULAN

### A. Kesimpulan

Kesimpulan dari laporan tugas besar ini adalah:

1. Pada tugas besar ini, telah dibuat program *information retrieval* atau sistem temu balik informasi yang memanfaatkan prinsip cosine similarity untuk menentukan tingkat relevansi suatu dokumen terhadap query yang dimasukkan oleh pengguna.
2. Dalam implementasinya, telah digunakan *web framework* dalam bahasa Python, yakni Flask. Dengan menggunakan *web framework* Flask, proses pengerjaan tugas besar ini menjadi terbantu karena Flask menyederhanakan persoalan pengintegrasian kode *back end* dengan tampilan *front end* yang merupakan antar muka yang digunakan oleh pengguna (*user*).
3. Dimanfaatkan pula berbagai macam library Python seperti *nltk*, *newspaper3k*, *re*, dan *string* yang membuat proses pembersihan dokumen serta query menjadi dapat dilakukan. Dengan dibersihkannya dokumen serta query, maka tingkat relevansi yang didakan antara query dengan dokumen terkait menjadi lebih baik.
4. Dengan bantuan library *newspaper3k*, program ini juga mampu melakukan *web scraping* dengan cara memasukkan *url* suatu laman berita, sehingga nanti program akan mampu untuk memberikan salinan konten berita dari laman terkait.

### B. Saran

Program yang kami buat dengan menggunakan bantuan *framework* Flask dalam bahasa Python ini, masih memiliki tampilan antar muka pengguna yang sangat sederhana. Hal ini dikarenakan Flask merupakan *framework* yang lebih berfokus di bagian *back end* saja. Kedepannya, agar dapat membuat tampilan antar muka yang lebih menarik, bisa digunakan *framework* tambahan yang memang akan memudahkan proses pengerjaan *front end* dalam laman web. *Framework* yang dimaksud salah satunya ialah *framework* React. Selain itu, cosine similarity yang digunakan sebagai basis sistem temu balik pada program kali ini masihlah merupakan implementasi yang tergolong dasar. Agar lebih memiliki algoritma sistem temu balik yang lebih canggih lagi, dapat diimplementasikan algoritma TF-IDF (*Term Frequency-Inverse Document Frequency*).

### C. Refleksi

Refleksi yang didapat dari pengerjaan tugas besar ini adalah:

1. Kahfi Soobhan Zulkifli (13519012)  
Belajar banyak hal tentang implementasi Python dalam search engine.
2. Muhammad Akram Al Bari (13519142)  
Tugas besar kali ini memberikan pengalaman baru berupa penggunaan *framework* Flask yang membuat saya lebih memiliki gambaran tentang bagaimana proses persiapan *back end* suatu laman website dikerjakan. Tugas

besar kali ini juga kembali mengajarkan saya untuk terus semangat dan melakukan eksplorasi terhadap ilmu-ilmu yang bisa diperjuangkan.

3. Jauhar Wibisono (13519160)

Pada tugas ini saya belajar menggunakan beberapa teknologi dan bahasa pemrograman baru - Flask dan HTML + Jinja, sungguh mengasyikkan.

## DAFTAR PUSTAKA

[1] flask.palletproject.com

[2] "Aplikasi Dot Product pada sistem temu balik aplikasi" oleh Rinaldi Munir  
<https://informatika.stei.itb.ac.id/~rinaldi.munir/AljabarGeometri/2020-2021/Algeo12-Aplikasi-dot-product-pada-IR.pdf>

[3] "Create A Simple Search Engine Using Python" oleh Irfan"Create A Simple  
<https://link.medium.com/yEtXO932Kab> Alghani Khalid

[4] "Removing Stop Words with NLTK Python"  
<https://www.geeksforgeeks.org/removing-stop-words-nltkpython/#:~:text=What%20are%20Stop%20words%3F,result%20of%20a%20search%20query>

[5] "Stemming and Lemmatization Python"  
<https://www.datacamp.com/community/tutorials/stemming-lemmatization-python>

[6] "The Flask Mega-Tutorial"  
<https://blog.miguelgrinberg.com/post/the-flask-mega-tutorial-part-i-hello-world>

[7] "How to read multiple text files in a folder with python"  
<https://stackoverflow.com/questions/57111243/how-to-read-multiple-text-files-in-a-folder-with-python>