LAPORAN TUGAS BESAR 2
IF2123 ALJABAR LINIER DAN GEOMETRI
APLIKASI DOT PRODUCT PADA SISTEM TEMU-BALIK INFORMASI
SEMESTER I TAHUN 2020/2021

Disusun oleh:
Ruhiyah Faradishi Widiaputri (13519034)
Melita (13519063)
Akifa Nabil Ufairah (13519179)

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG

2020

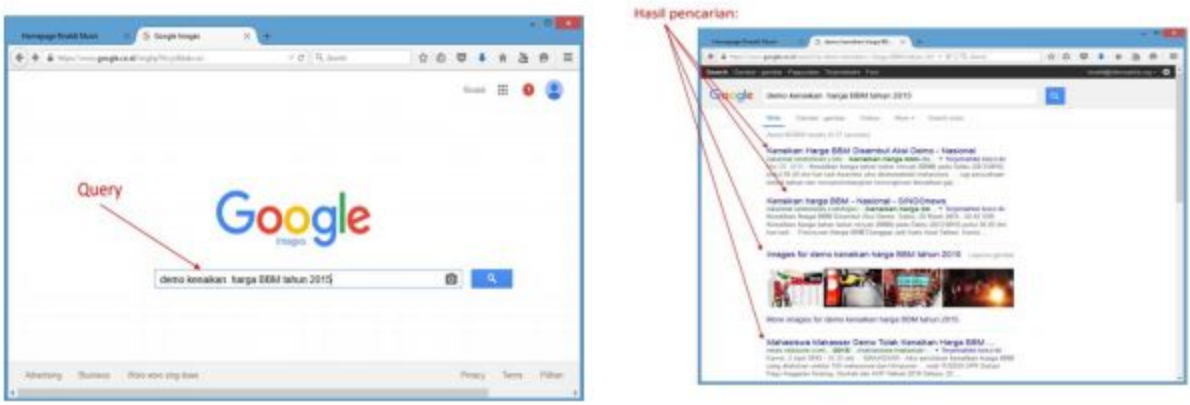
DAFTAR ISI

BAB I	**2**
BAB II	**3**
BAB III	**6**
BAB IV	**9**
BAB V	**15**
REFERENSI	**16**

BAB I
Deskripsi Masalah

1.1. 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 vektor 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

sumber: Aplikasi Dot Product pada Sistem Temu-balik Informasi by Rinaldi Munir

Ide utama dari sistem temu balik informasi adalah mengubah *search query* menjadi ruang vektor Setiap dokumen maupun *query* dinyatakan sebagai vektor $w = (w_1, w_2, ..., w_n)$ 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:

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

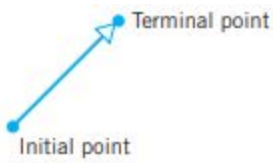
Pada kesempatan ini, kalian ditantang untuk membuat sebuah *search engine* sederhana dengan model ruang vektor dan memanfaatkan *cosine similarity*.

BAB II
Teori Singkat

2.1. Vektor

2.1.1. Pengertian vektor

Vektor merupakan kuantitas fisik yang memiliki besar dan arah. Vektor dilambangkan dengan huruf-huruf kecil yang dicetak tebal atau diberikan tanda panah (jika berupa tulisan tangan), misalkan **u**, **v**, **w**. Vektor di ruang 2D atau 3D dapat direpresentasikan dengan menggunakan panah.



Jika suatu vektor **v** mempunyai titik asal A dan titik terminal B, maka **v** = $\overrightarrow{AB}$. Vektor-vektor **u** = (u_1, u_2,...,u_n) dan **v** = (v_1, v_2,...,v_n) di R_n dikatakan sama (**u** = **v**) jika:

$$v_1 = w_1, \quad v_2 = w_2, \ldots, \quad v_n = w_n$$

2.1.2. Operasi dasar vektor

Jika **u** = (u_1, u_2,...,u_n) dan **v** = (v_1, v_2,...,v_n) adalah vektor-vektor di R_n, dan jika k adalah sembarang skalar maka didefinisikan:

$$\begin{aligned}
&\mathbf{v} + \mathbf{w} = (v_1 + w_1, v_2 + w_2, \ldots, v_n + w_n) \\
&k\mathbf{v} = (kv_1, kv_2, \ldots, kv_n) \\
&-\mathbf{v} = (-v_1, -v_2, \ldots, -v_n) \\
&\mathbf{w} - \mathbf{v} = \mathbf{w} + (-\mathbf{v}) = (w_1 - v_1, w_2 - v_2, \ldots, w_n - v_n)
\end{aligned}$$

2.1.2. Panjang vektor

Panjang dari suatu vektor **v** = (v_1,v_2,...,v_n) disebut juga norma dari **v**, dilambangkan dengan ||**v**|| didefinisikan sebagai berikut:

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

Jika **v** adalah vektor di Rn dan jika k adalah skalar maka:

- ||**v**|| ≥ 0
- ||**v**|| = 0 jika dan hanya jika **v** = **0**
- ||k**v**|| = |k|||**v**||

2.1.3. Dot Product

Jika **u** dan **v** adalah vektor-vektor tidak nol di R2 dan R3, dan jika θ adalah sudut di antara **u** dan **v** maka dot product (perkalian titik) dari **u** dan **v**, dilambangkan dengan **u** · **v** didefinisikan sebagai:

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta$$

Jika **u** = **0** atau **v** = **0**, maka **u** · **v** = 0.

Dari formula di atas kita dapat menentukan besarnya sudut yang dibentuk oleh 2 vektor **u** dan **v**, yaitu dari

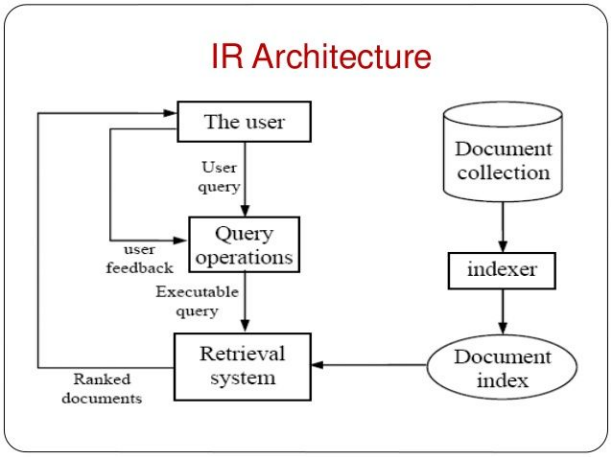
$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}$$

Jika **u** = (u_1, u_2,...,u_n) dan **v** = (v_1, v_2,...,v_n) adalah vektor-vektor di Rn, maka dot product dari **u** dan **v** juga dapat dinyatakan sebagai berikut:

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$$

2.2. Sistem Temu Balik Informasi

Sistem Temu Kembali Informasi (STKI) atau *Information Retrieval System* (IRS) digunakan untuk menemukan kembali *(retrieve)* informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Sistem temu balik informasi umumnya digunakan pada pencarian informasi yang isinya tidak terstruktur, seperti dokumen dan laman web.



Gambar: arsitektur sistem temu balik informasi

Salah satu penerapan dari sistem temu balik informasi ini adalah pada *search engine*. Sistem temu balik informasi membantu pengguna dalam menemukan informasi yang dibutuhkannya. Selain itu sistem temu balik informasi ini juga digunakan dalam filter spam pada email program untuk mengklasifikasikan email mana yang merupakan spam dan mana yang bukan spam.

Sistem temu balik informasi mempunyai kemampuan untuk menampilkan, menyimpan, mengatur, dan mengakses *item-item* informasi. Untuk melakukan pencarian dibutuhkan *query*, yaitu sekumpulan *keyword*.

2.3. Cosine Similarity

Cosine similarity adalah salah satu cara untuk menentukan seberapa besar kesamaan antara 2 buah dokumen. Model ini menggunakan teori di dalam aljabar vektor.

Misalkan terdapat n kata berbeda sebagai kamus kata (*vocabulary*) atau indeks kata (*term index*). Kata-kata tersebut membentuk ruang vektor berdimensi n. Setiap dokumen maupun *query* dinyatakan sebagai vektor **w** = (w_1, w_2,..., w_n) di dalam R_n dengan w_i adalah jumlah kemunculan setiap kata *i* di dalam dokumen.

Secara matematis, jika **d** adalah vektor yang kita peroleh dari suatu dokumen dan **q** adalah vektor yang kita dapatkan dari *query* seperti di atas, *cosine similarity* mengukur kosinus dari sudut yang dibentuk di antara 2 vektor **d** dan **q.**
Dengan kata lain kesamaan (sim) antara dua vektor **Q** = (q_1, q_2, …, q_n) dan **D** = (d_1, d_2, …, d_n) dapat ditentukan dengan:

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

Dengan **Q**,**D** adalah perkalian titik yang didefinisikan sebagai:

$$\mathbf{Q} \cdot \mathbf{D} = q_1 d_1 + q_2 d_2 + \ldots + q_n d_n$$

BAB III
Implementasi Program

3.1. Struktur Class yang Didefinisikan

1. Flask
2. StemmerFactory
3. StopWordRemoverFactory
4. Counter
5. BeautifulSoup

3.2. Garis Besar Program

Berikut adalah *working directory* program *search engine* pada *website* yang kami buat.

- doc
- src
  - static
    - images
    - styles
      - styles.css
    - uploads
  - templates
    - about.html
    - index.html
    - upload.html
  - index.py
  - index2.py
- test
- README.md
- screenshot.png
- screenshot2.png

Terdapat file index.py dan index2.py yang merupakan *backend* dari *website*. Pada program *file* index.py dan index2.py kami membuat 13 buah fungsi dan prosedur, yaitu:

1. getdata()
   - Fungsi ini mengumpulkan data dengan melakukan *web scraping* pada *website* alodokter.
2. countWordsArticles(df)
   - Fungsi untuk menghitung banyak kata pada setiap artikel dan mengembalikan sebuah array yang menyimpan banyak kata dari setiap artikel yang telah dikumpulkan sebelumnya.
3. clean_text(text)
   - Fungsi yang menerima parameter string ini akan melakukan proses *cleaning data* berupa penghapusan karakter-karakter yang tidak diperlukan seperti angka dan tanda baca menggunakan bantuan modul string, serta menghapus *stopwords* dan melakukan *stemming* menggunakan *library* sastrawi.

4. clean_articles(Articles)
    - Fungsi yang menerima *input* sebuah *array* yang berisi artikel-artikel yang digunakan pada program ini, lalu melakukan pembersihan data pada setiap artikel pada *array* tersebut dengan memanggil kembali fungsi clean_text(text)
5. nilaidot(vec,q_vec)
    - Fungsi yang menerima 2 buah *array* yang merepresentasikan vektor sebagai parameternya dan mengembalikan hasil perkalian dot dari kedua vektor tersebut.
6. panjangvektor(vector)
    - Fungsi yang menerima sebuah *array* yang merepresentasikan vektor sebagai parameter *input* dan mengembalikan panjang dari vektor tersebut.
7. get_sorted_sim(q,df)
    - Fungsi ini terlebih dahulu akan mengubah *query* q ke dalam bentuk vektor, lalu menghitung nilai similarity dari *query* q dengan setiap artikel yang terdapat pada df dengan memanggil fungsi-fungsi sebelumnya. Setelah mendapat nilai *similarity* dari *query* dengan tiap artikel yang disimpan dalam sebuah *array* sim, *array* ini akan *di-sorting* mulai dari nilai *similarity* terbesar hingga terkecil. Kemudian, fungsi akan mengembalikan *array* yang sudah di-*sorting* ini.
8. getdatafile()
    - Fungsi ini dipanggil saat mengumpulkan data dari *file* .txt.
9. get_unique_words(articles)
    - Fungsi untuk mencari kata-kata unik dalam artikel (len>=2).
10. vectorize(articles,unique_words)
    - Fungsi untuk membuat DataFrame dari artikel.
11. vektorquery(query,unique_words)
    - Fungsi untuk mengubah *query* menjadi vektor.
12. listterm(qkata)
    - Fungsi untuk membuat *array* berisi *query.*
13. kolterm(kata, arrterm)
    - Fungsi untuk menghitung kata pada *query.*

Selain fungsi di atas, juga terdapat 4 fungsi dari website ini yang dibuat dengan menggunakan Web Framework Flask, yaitu:

1. index()

    Fungsi ini akan dipanggil saat halaman /index diakses. Saat *method* yang terbaca adalah POST, yaitu saat user memasukan input *query* maka fungsi ini akan melakukan perhitungan *similarity* untuk menampilkan *output* yang sesuai dengan menggunakan fungsi-fungsi yang sebelumnya sudah dijabarkan. Kemudian index() akan mengembalikan data yang akan digunakan pada *file* index.html untuk ditampilkan ke *website.*

2. about()

   Fungsi akan dipanggil saat halaman /about diakses, yaitu saat pengguna mengklik pada tulisan perihal di halaman indeks.
3. upload_form()

   Fungsi dipanggil saat halaman /upload diakses.
4. upload_file()

   Fungsi dipanggil saat *submit* file .txt.

Selanjutnya pada folder templates, terdapat 3 *file* html, yaitu index.html, upload.html, dan about.html yang masing-masing akan ditampilkan sesuai dengan halaman *website* yang diakses. Secara garis besar kedua html ini terbagi menjadi 2 bagian yaitu head dan body. Bagian head berisi title dan link ke stylesheets dari web page terkait. Sedangkan pada bagian body, terdapat informasi yang ingin ditampilkan ke *website*.

Pada folder static, terdapat *resources* tambahan yang kami gunakan pada *website*, yang terbagi menjadi 2 folder, yaitu images dan styles. Pada folder images terdapat *image file* yang ditampilkan pada *html file*. Sedangkan pada folder style terdapat file styles.css yang digunakan sebagai stylesheet di *file* html. Selain itu, pada static ada satu folder lagi bernama uploads yang menyimpan data .txt yang pernah di-*upload* ke *website.*

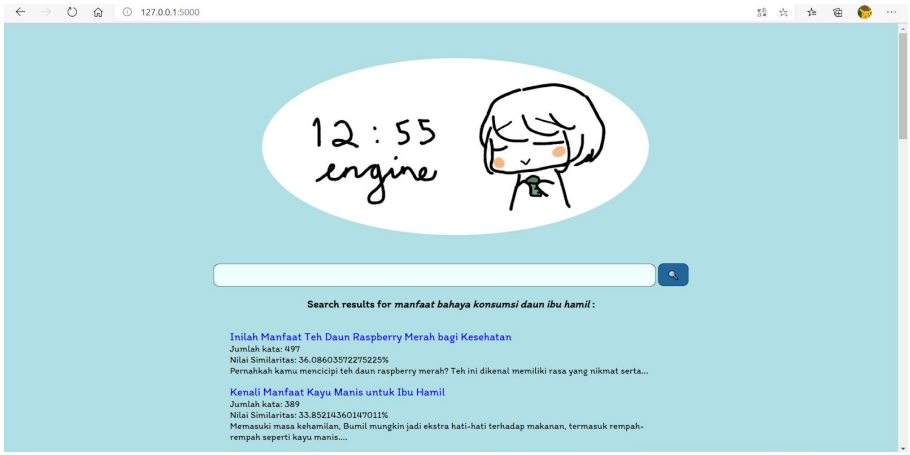
BAB IV
Eksperimen

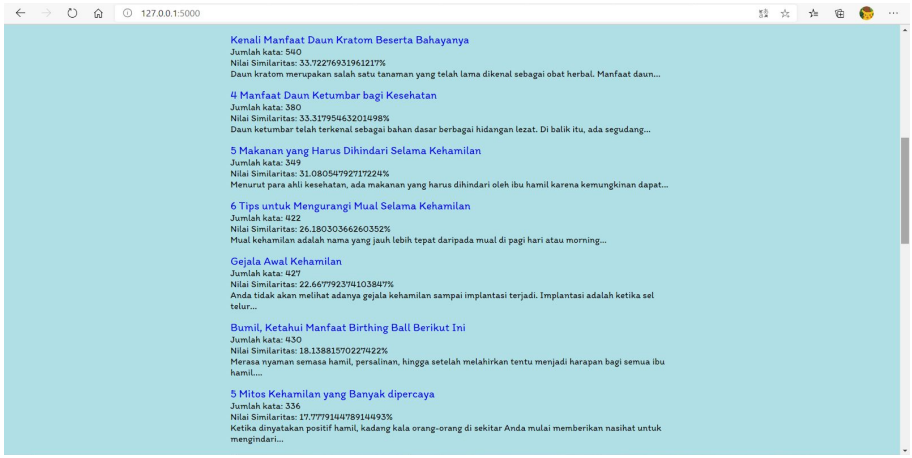
4.1. Eksperimen dengan Dokumen dari *Web Scraping*

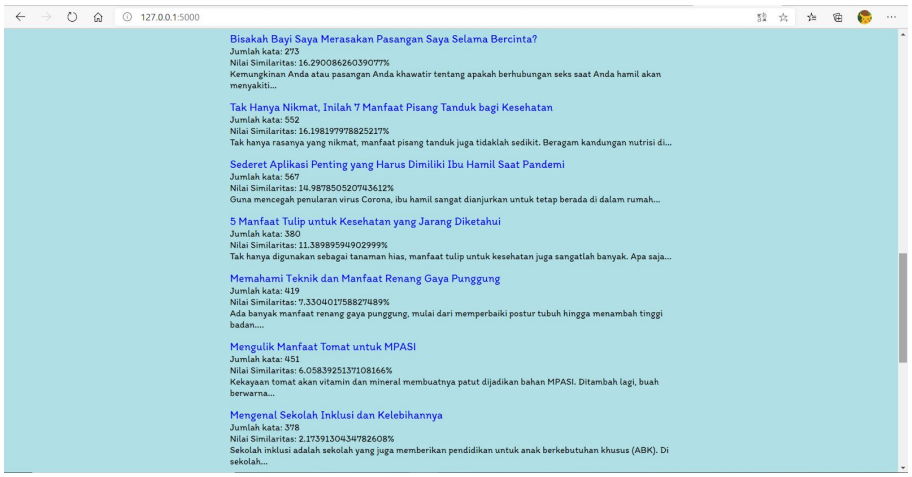
4.1.1. Kasus 1:

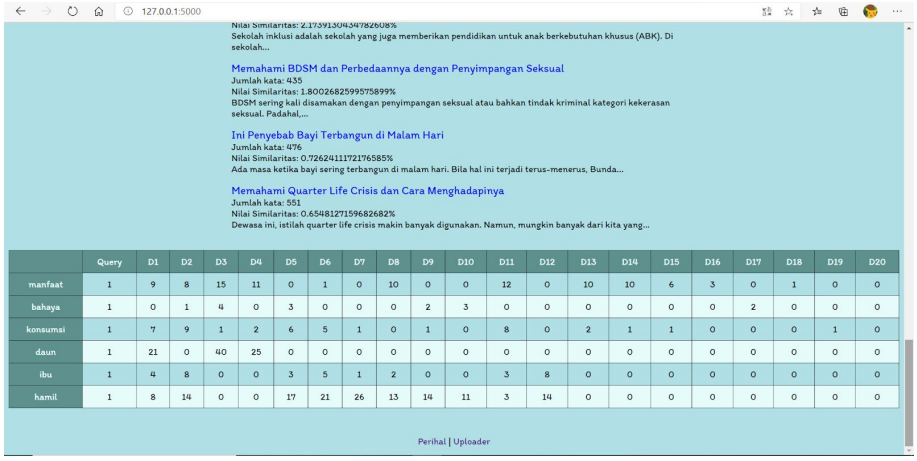
Query = manfaat dan bahaya konsumsi daun-daunan untuk ibu hamil

Hasil:







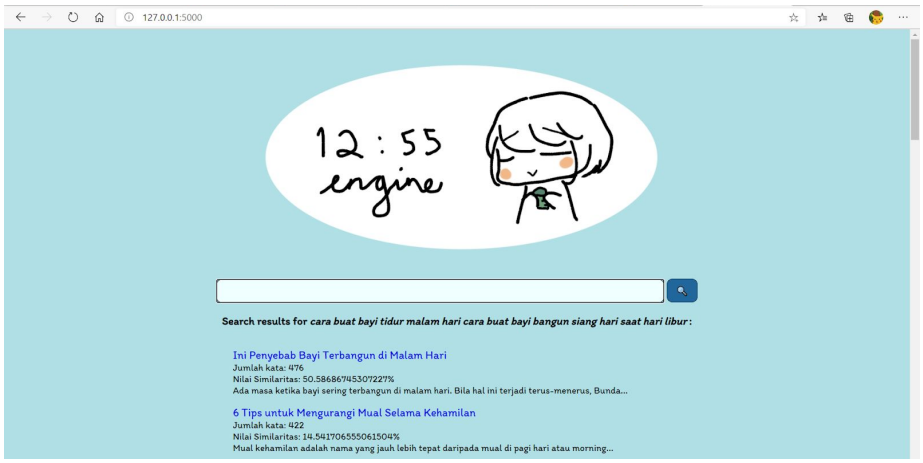


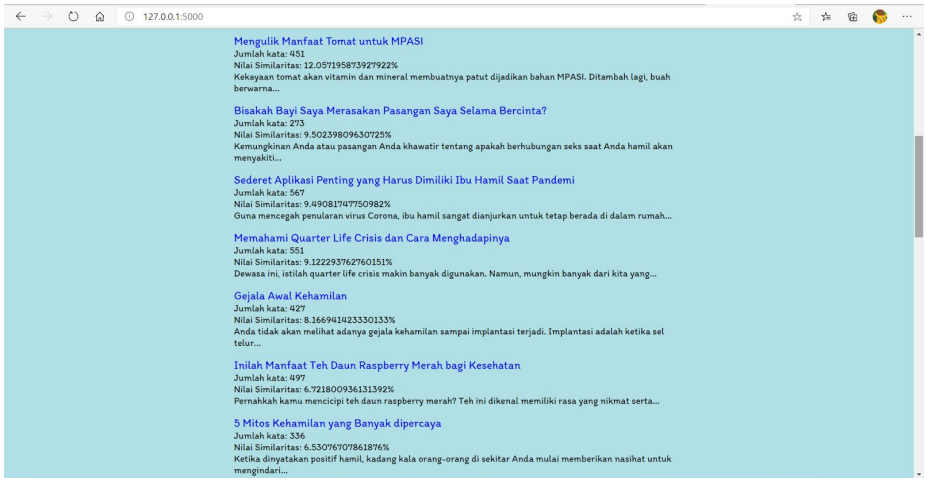
	Query	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
manfaat	1	9	8	15	11	0	1	0	10	0	0	12	0	10	10	6	3	0	1	0	0
bahaya	1	0	1	4	0	3	0	0	0	2	3	0	0	0	0	0	0	2	0	0	0
konsumsi	1	7	9	1	2	6	5	1	0	1	0	8	0	2	1	1	0	0	0	1	0
daun	1	21	0	40	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ibu	1	4	8	0	0	3	5	1	2	0	0	3	8	0	0	0	0	0	0	0	0
hamil	1	8	14	0	0	17	21	26	13	14	11	3	14	0	0	0	0	0	0	0	0

4.1.2. Kasus 2:

Query = cara membuat bayi tidur pada malam hari dan cara membuat bayi terbangun pada siang hari pada saat hari libur

Hasil:





127.0.0.1:5000

Bumil, Ketahui Manfaat Birthing Ball Berikut Ini
Jumlah kata: 430
Nilai Similaritas: 5.576707335582519%
Merasa nyaman semasa hamil, persalinan, hingga setelah melahirkan tentu menjadi harapan bagi semua ibu hamil....

5 Makanan yang Harus Dihindari Selama Kehamilan
Jumlah kata: 349
Nilai Similaritas: 4.633638292702506%
Menurut para ahli kesehatan, ada makanan yang harus dihindari oleh ibu hamil karena kemungkinan dapat...

Tak Hanya Nikmat, Inilah 7 Manfaat Pisang Tanduk bagi Kesehatan
Jumlah kata: 552
Nilai Similaritas: 4.489247660743688%
Tak hanya rasanya yang nikmat, manfaat pisang tanduk juga tidaklah sedikit. Beragam kandungan nutrisi di...

Memahami Teknik dan Manfaat Renang Gaya Punggung
Jumlah kata: 419
Nilai Similaritas: 3.8415393827692665%
Ada banyak manfaat renang gaya punggung, mulai dari memperbaiki postur tubuh hingga menambah tinggi badan....

Memahami BDSM dan Perbedaannya dengan Penyimpangan Seksual
Jumlah kata: 435
Nilai Similaritas: 3.4592835938508664%
BDSM sering kali disamakan dengan penyimpangan seksual atau bahkan tindak kriminal kategori kekerasan seksual. Padahal,...

6 Fakta Eyelash Extension yang Penting untuk Diketahui
Jumlah kata: 517
Nilai Similaritas: 3.3181031838528585%
Eyelash extension dilakukan agar bulu mata tampak lebih panjang dan cantik. Namun, sebelum memutuskan untuk...

5 Manfaat Tulip untuk Kesehatan yang Jarang Diketahui
Jumlah kata: 380
Nilai Similaritas: 3.191725268112872 3%
Tak hanya digunakan sebagai tanaman hias, manfaat tulip untuk kesehatan juga sangatlah banyak. Apa saja...

127.0.0.1:5000

5 Manfaat Tulip untuk Kesehatan yang Jarang Diketahui
Jumlah kata: 380
Nilai Similaritas: 3.1917252681128723%
Tak hanya digunakan sebagai tanaman hias, manfaat tulip untuk kesehatan juga sangatlah banyak. Apa saja...

4 Manfaat Daun Ketumbar bagi Kesehatan
Jumlah kata: 380
Nilai Similaritas: 2.948368258934485%
Daun ketumbar telah terkenal sebagai bahan dasar berbagai hidangan lezat. Di balik itu, ada segudang...

Kenali Manfaat Kayu Manis untuk Ibu Hamil
Jumlah kata: 389
Nilai Similaritas: 1.6262043771744308%
Memasuki masa kehamilan, Bumil mungkin jadi ekstra hati-hati terhadap makanan, termasuk rempah-rempah seperti kayu manis....

Kenali Manfaat Daun Kratom Beserta Bahayanya
Jumlah kata: 540
Nilai Similaritas: 1.0799929584688661%
Daun kratom merupakan salah satu tanaman yang telah lama dikenal sebagai obat herbal. Manfaat daun...

Mengenal Sekolah Inklusi dan Kelebihannya
Jumlah kata: 378
Nilai Similaritas: 0.6962093643699441%
Sekolah inklusi adalah sekolah yang juga memberikan pendidikan untuk anak berkebutuhan khusus (ABK). Di sekolah...

127.0.0.1:5000

Kenali Manfaat Daun Kratom Beserta Bahayanya
Jumlah kata: 540
Nilai Similaritas: 1.0799929584688661%
Daun kratom merupakan salah satu tanaman yang telah lama dikenal sebagai obat herbal. Manfaat daun...

Mengenal Sekolah Inklusi dan Kelebihannya
Jumlah kata: 378
Nilai Similaritas: 0.6962093643699441%
Sekolah inklusi adalah sekolah yang juga memberikan pendidikan untuk anak berkebutuhan khusus (ABK). Di sekolah...

	Query	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
cara	2	1	3	1	0	0	4	1	0	0	1	1	2	0	2	1	1	1	0	0	1
buat	2	1	1	1	0	4	6	0	1	0	3	0	4	2	0	4	1	1	1	0	0
bayi	2	22	0	8	8	5	0	0	4	5	2	2	0	0	0	0	0	0	0	0	0
tidur	1	14	4	0	0	0	0	2	0	0	0	0	0	1	0	0	3	0	0	4	0
malam	1	17	1	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0
hari	3	16	6	3	0	3	3	4	3	1	1	1	1	1	1	0	0	1	0	0	0
bangun	1	17	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
siang	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
saat	1	1	1	0	1	2	0	1	0	0	1	0	0	3	1	1	0	0	2	0	0
libur	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

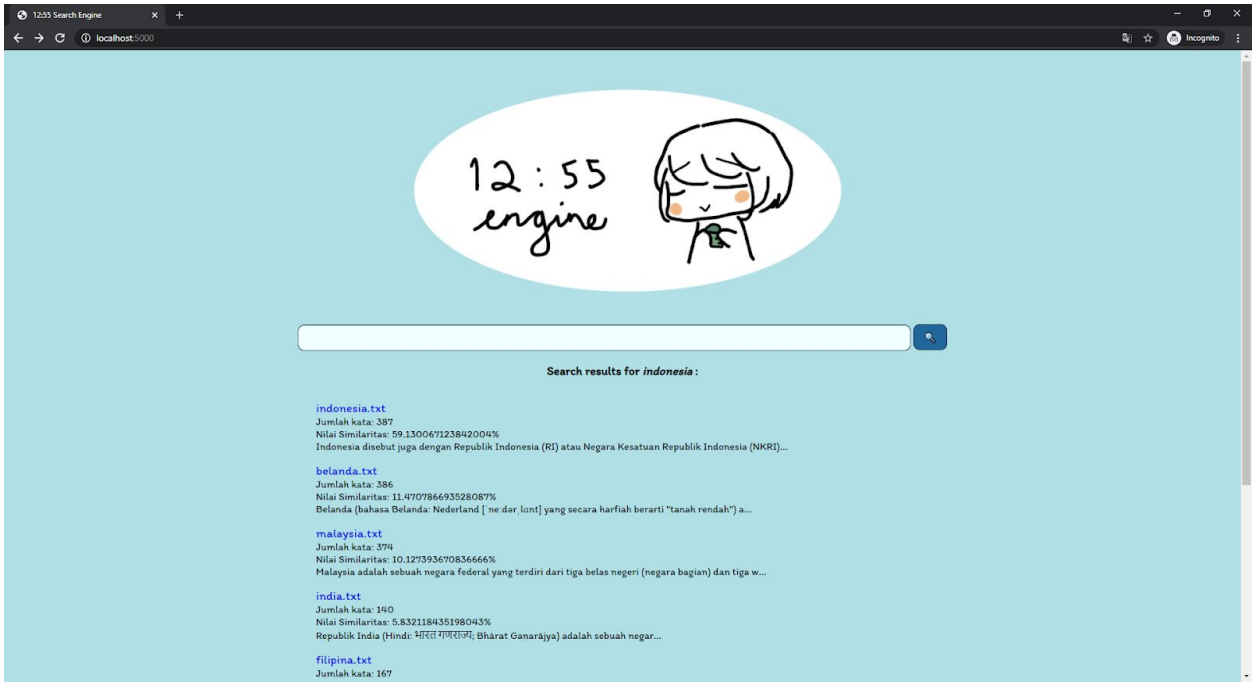
Perihal | Uploader

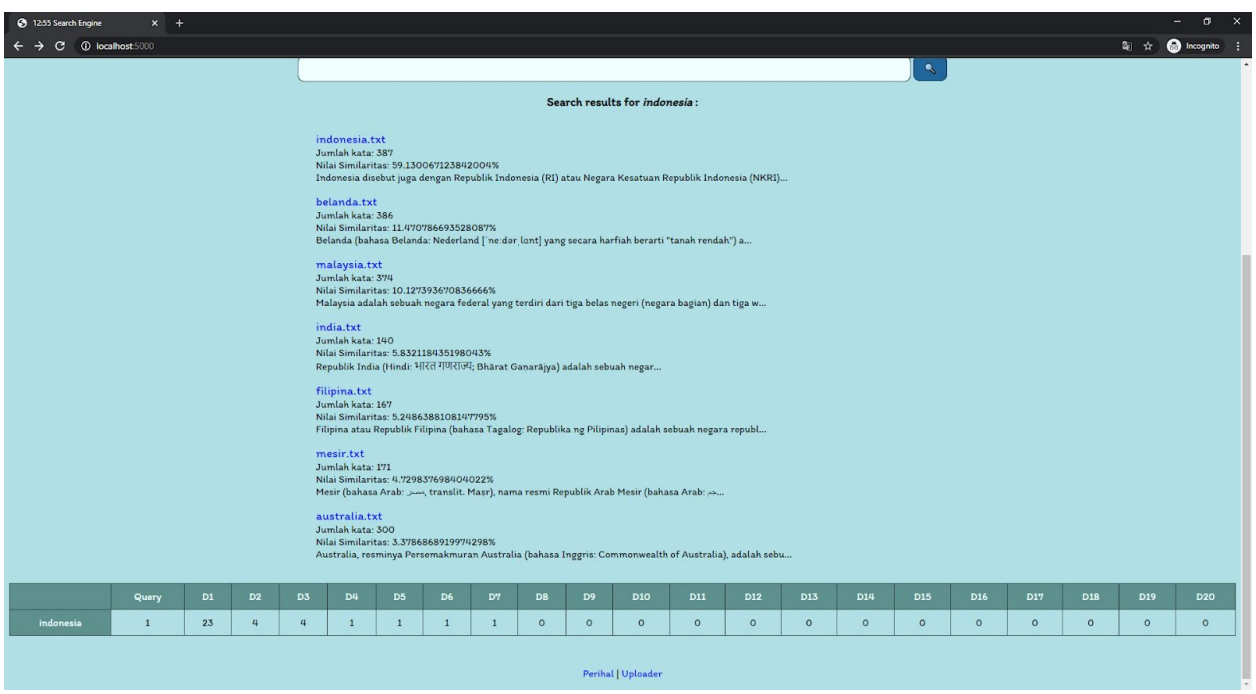
4.2. Eksperimen dengan Dokumen .txt

4.2.1. Kasus 1:

Query = indonesia

Hasil:



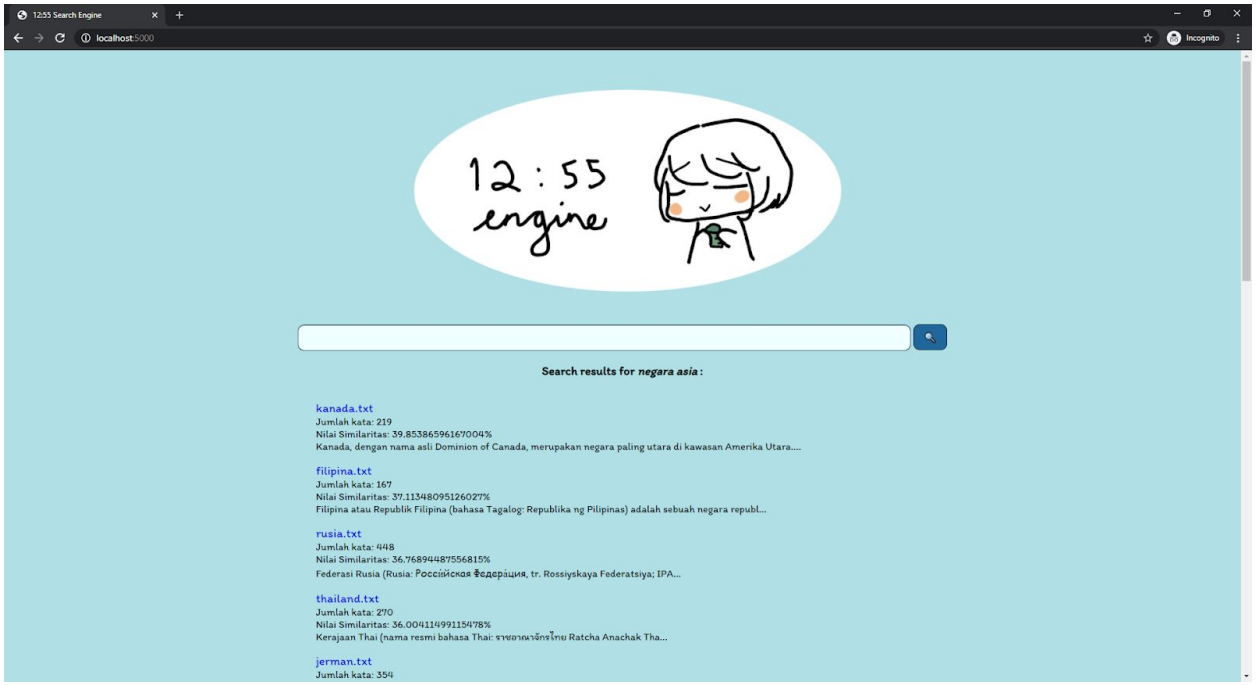


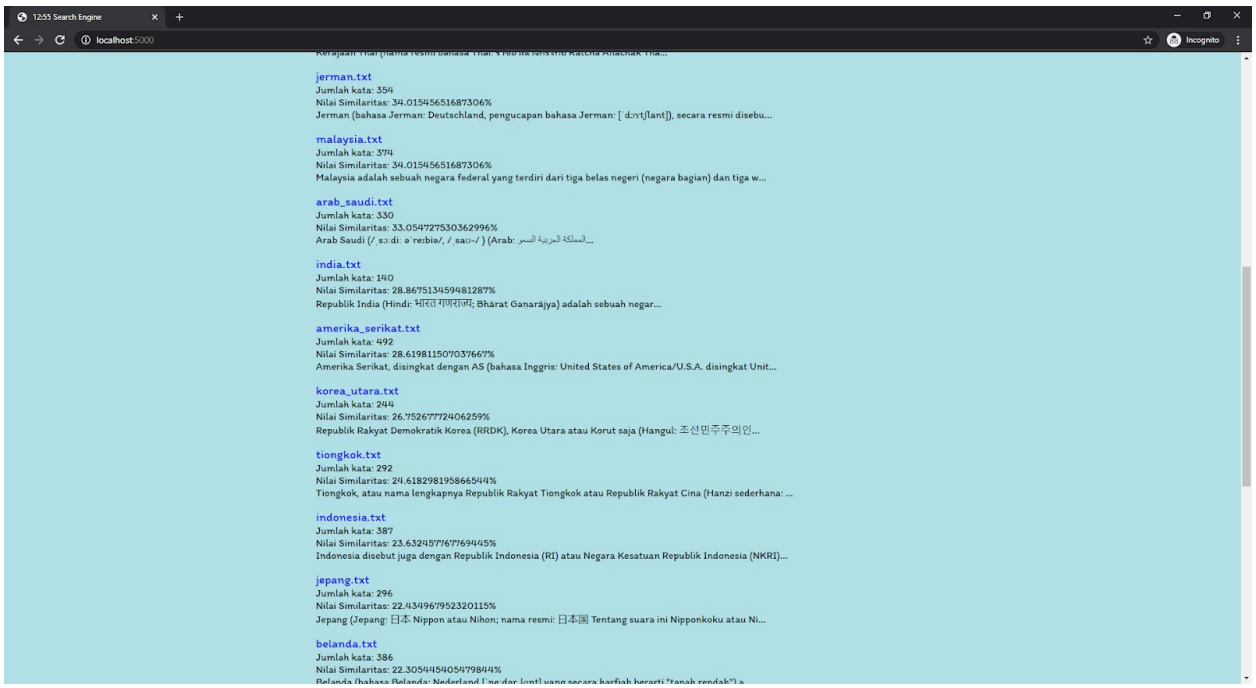
	Query	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
indonesia	1	23	4	4	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

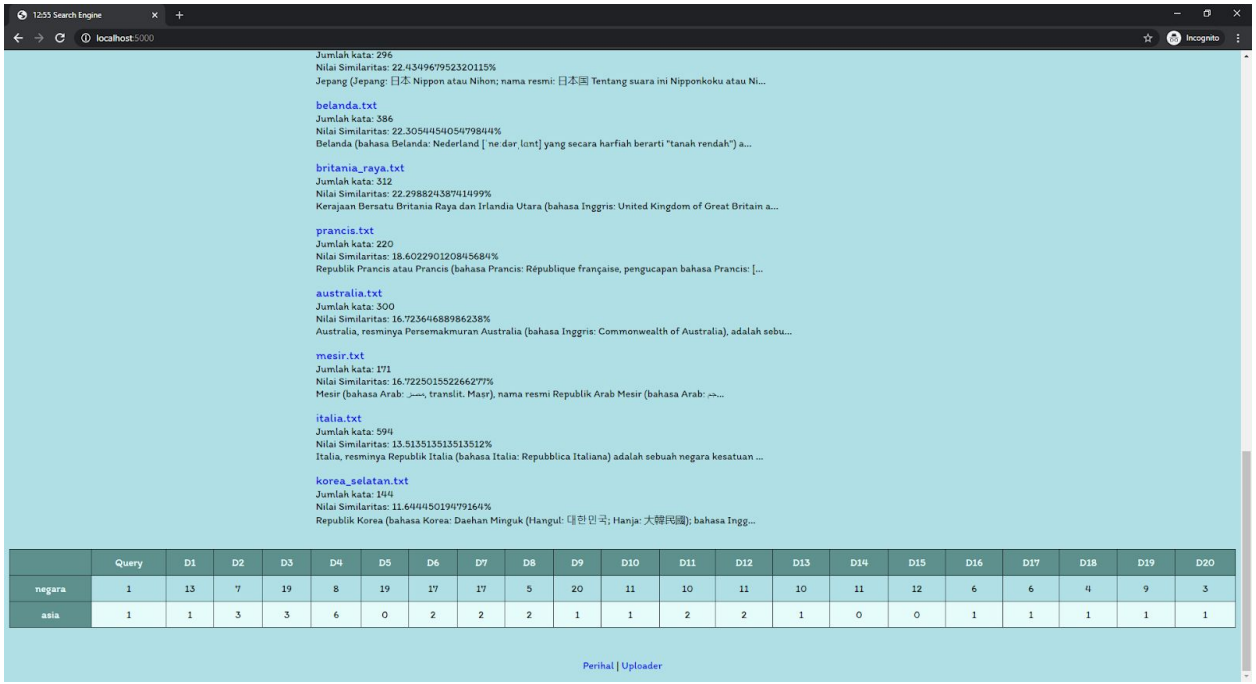
4.2.2. Kasus 2:

Query = negara di asia

Hasil:







	Query	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
negara	1	13	7	19	8	19	17	17	5	20	11	10	11	10	11	12	6	6	4	9	3
asia	1	1	3	3	6	0	2	2	2	1	1	2	2	1	0	0	1	1	1	1	1

BAB V
Kesimpulan, Saran, dan Refleksi

5.1. Kesimpulan

Search engine yang dibuat dapat mencari informasi menggunakan *cosine similarity,* baik dari *file* .txt ataupun *web scraping* dari situs alodokter. Hasil pencarian diurutkan berdasarkan nilai *similarity* terbesar sampai terkecil. Data yang ditampilkan adalah judul artikel berupa *link* menuju artikel tersebut, nilai *cosine similarity,* jumlah kata, dan deskripsi singkat artikel. Ditampilkan juga tabel yang berisi seluruh kata dalam *query* dan jumlah kata-kata tersebut di masing-masing dokumen.

5.2. Saran

Untuk pengembangan di masa depan, dapat digunakan *parser* lain agar proses *web scraping* menjadi lebih cepat. Selain itu, proses *file* dapat dilakukan langsung, sehingga pengguna tidak perlu me-*restart* aplikasi setelah meng-*upload file* baru.

5.3. Refleksi

Penanganan kasus *query* kosong sebaiknya tidak dilupakan. Penentuan *path* menggunakan Flask juga harus diperhatikan.

REFERENSI

https://informatikalogi.com/sistem-temu-kembali-informasi/ diakses 14 November 2020 pukul 22.14
https://image.slidesharecdn.com/tdminformationretrieval-150803041801-lva1-app6891/95/tdm-information-retrieval-13-638.jpg?cb=1438575616 diakses 14 November 2020 pukul 22.19
https://www.geeksforgeeks.org/what-is-information-retrieval/ diakses 14 November 2020 pukul 22.29
https://medium.com/dev-genius/get-started-with-multiple-files-upload-using-flask-e8a2f5402e20 diakses 15 November 2020 pukul 17.55
https://id.wikipedia.org/wiki/Daftar_negara_menurut_jumlah_penduduk diakses 15 November 2020 pukul 16.48