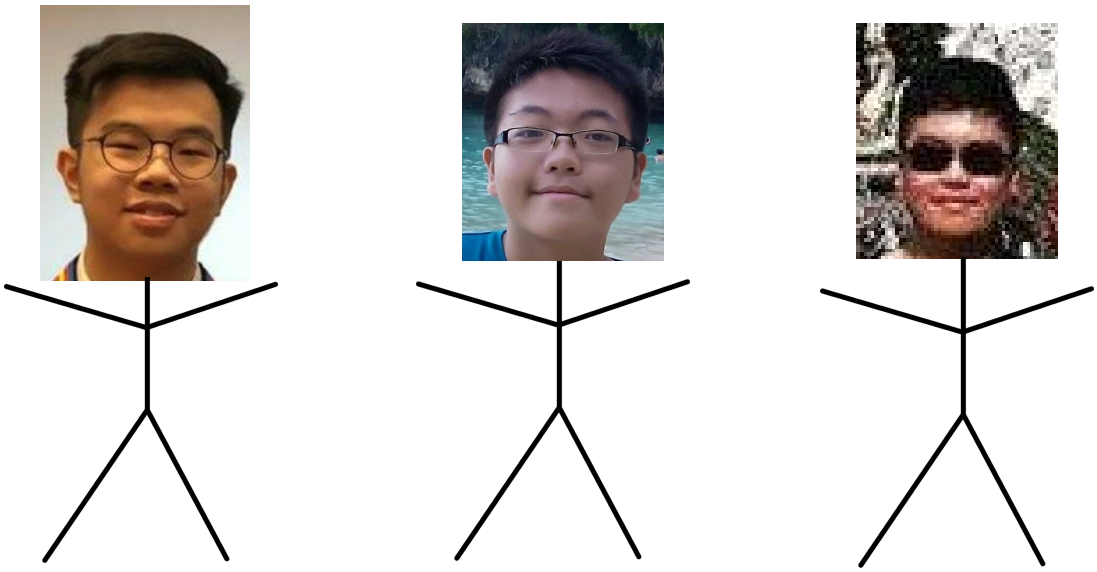
Laporan Tugas Besar 2 IF 2123 Aljabar Linier dan Geometri
Aplikasi Dot Product pada Sistem Temu-balik Informasi
Kelompok 65



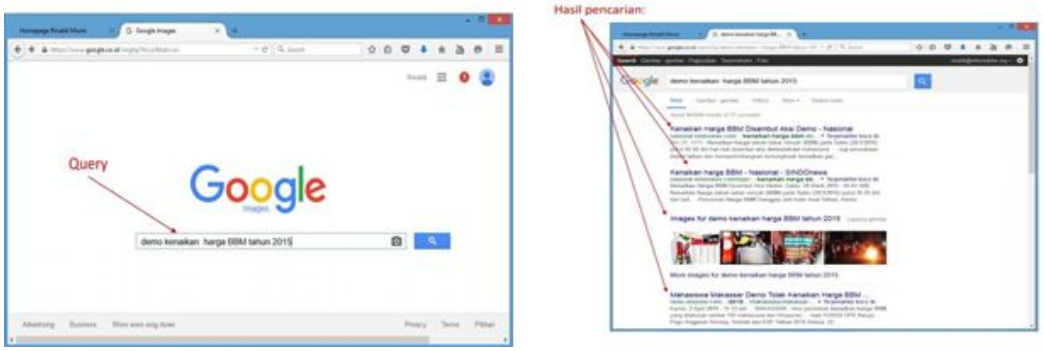
(Nama - NIM Anggota)
Girvin Junod - 13519096
David Owen Adiwiguna - 13519169
Leonard Matheus - 13519215

Semester 1 Tahun 2020/2021

BAB 1. DESKRIPSI MASALAH

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 = (w1, w2,..., wn) di dalam R^n, dimana nilai wi 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 2. TEORI SINGKAT

2.1. Retrieval Information

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.

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 search engine. Contoh lain dari Sistem Temu Kembali Informasi adalah sistem informasi perpustakaan

Permasalahan yang dihadapi pada Information Retrieval System (IRS) sama dengan permasalahan yang terdapat pada Data Retrieval System, yaitu jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah, dan data noise.

Tantangan Retrieval Information:

- Informasi dalam bentuk teks yang terstruktur.
- Jumlah data yang besar.
- Jumlah kemungkinan yang tinggi, memungkinkan semua kata dan frase.
- Hubungan antara konsep teks kompleks, contoh: “AOL bergabung dengan time-warner” & ”time-warner dibeli oleh AOL”.
- Kata ambigu dan kepekaan konteks, contoh: apple (perusahaan) atau apel (buah).
- Kesalahan data, contoh: kesalahan ejaan.

2.2. Vektor

Dalam aljabar vektor dikenal ruang vektor dengan n dimensi yang bernama Rn . Di dalam ruang tersebut terdapat elemen-elemen yang bernama vektor. Dalam ruang tersebut juga terdapat beberapa operasi yaitu perkalian skalar dan inner product (dot product).

Vektor adalah potongan/bagian/segmen dari suatu garis yang berarah. Besar sebuah vektor adalah panjangnya |v|. Arahnya ditunjukkan oleh arah anak panahnya. Sebuah vektor dianggap tidak berubah jika dipindahkan sejajar dengan dirinya. Vektor adalah benda 1 dimensi. Tetapi vektor dapat berada dalam ruang 3 dimensi.

2.2. Cosine Similarity

Cosine Similarity adalah ukuran kesamaan antara dua buah vektor dalam sebuah ruang dimensi yang didapat dari nilai cosinus sudut dari perkalian dua buah vektor yang dibandingkan karena cosinus dari 0 derajat adalah 1 dan kurang dari 1 untuk nilai sudut yang lain, maka nilai similarity dari dua buah vektor dikatakan mirip ketika nilai dari cosine similarity adalah 1.

Cosine similarity digunakan dalam ruang positif, dimana hasilnya dibatasi antara nilai 0 dan 1. Kalau nilainya 0 maka dokumen tersebut dikatakan mirip jika hasilnya 1 maka nilai tersebut dikatakan tidak mirip Perhatikan bahwa batas ini berlaku untuk sejumlah dimensi, dan Cosine similarity ini paling sering digunakan dalam ruang positif dimensi tinggi. Misalnya, dalam Information Retrieval, masing-masing kata/istilah (term) diasumsikan sebagai dimensi yang berbeda dan dokumen ditandai dengan vektor dimana nilai masing-masing dimensi sesuai dengan berapa istilah muncul dalam dokumen.

BAB 3. IMPLEMENTASI PROGRAM

Pada Tugas besar kali ini, kami menggunakan bahasa pemrograman Python dengan Flask sebagai alat bantu Routing untuk melakukan backend pada Website kami. Program kami dibuat secara modular sesuai fungsi dan spesifikasi masing-masing yang akan dijelaskan lebih lanjut di bawah ini.

Program kami terbagi dari 3 bagian besar yaitu:

1. App.py
    - Berisi Program utama beserta Routing yang ada.
2. Query.csv
    - Berfungsi untuk menyimpan term Table.
3. Templates
    - Berisi html dan tampilan css.

3.1. App.py

A. Fungsi tanpa Routing

1. Read File
    a. Berfungsi untuk membaca file yang tersimpan secara lokal.
    b. Memiliki parameter input alamat folder yang bersangkutan.
    c. Memiliki parameter output tuple kalimat yang telah dibaca dan siap diolah.
2. Read First
    a. Berfungsi untuk membaca kalimat pertama dari setiap file lokal.
    b. Memiliki parameter input alamat folder yang bersangkutan.
    c. Memiliki parameter output tuple kalimat yang awal untuk ditampilkan pada halaman website.
3. Clean Document
    a. Berfungsi untuk melakukan stemming dan cleaning pada stopwords.
    b. File yang kami gunakan berbahasa Inggris sehingga metode clean yang digunakan menggunakan pustaka NLTK.
    c. Memiliki parameter input alamat file dokumen yang bersangkutan.
    d. Memiliki parameter output tuple kalimat yang telah di-clean.

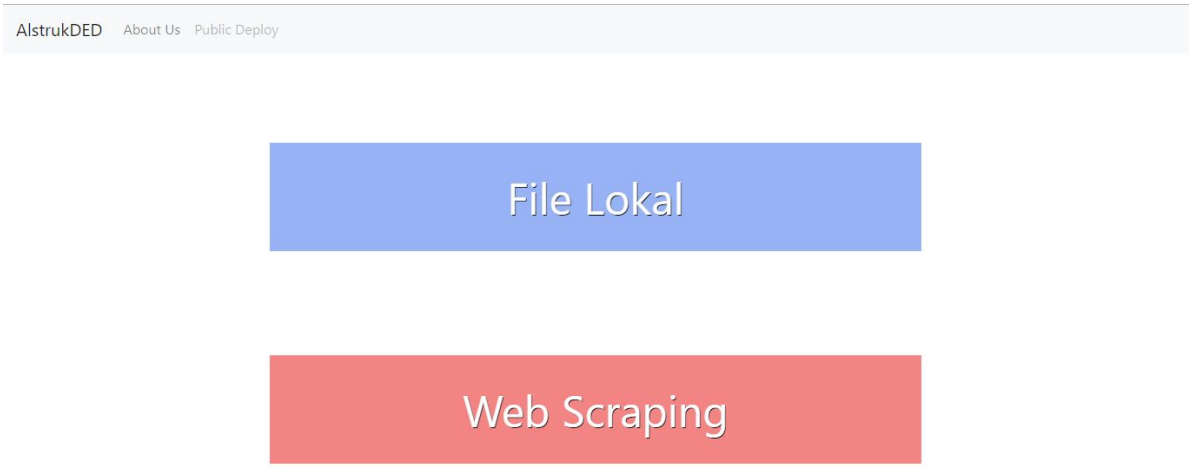
4. Banyak Kata
    a. Berfungsi untuk menghitung banyaknya kata yang ada pada dokumen (Tentunya setelah mengalami proses cleaning).
    b. Memiliki parameter input kalimat pada dokumen dan angka jumlah file keseluruhan.
    c. Memiliki parameter output tuple integer banyak kata pada keseluruhan dokumen.
5. Query Table
    a. Pusat Operasi pengolahan kata-kata unik terjadi pada fungsi ini.
    b. Berguna untuk mengolah kata-kata yang muncul pada suatu dokumen dan mengolahnya menjadi matriks query sesuai dengan judul-judul yang bersesuaian.
    c. Memiliki parameter input query, kalimat yang sudah di-clean, dan jumlah file secara keseluruhan.
    d. Memiliki parameter output berupa matriks.
6. Similar
    a. Proses *Cosine Similarity* secara manual terjadi pada fungsi ini.
    b. Cara kerjanya yaitu melakukan iterasi perkalian *dot product* dan melakukan analisis mirip dengan perhitungan Cos sudut tertentu (Skala 0 s.d. 1).
    c. Memiliki parameter input matriks yang dihasilkan pada Query Table.
    d. Memiliki output berupa tupel nilai kesamaan dan banyak query yang bersesuaian.
7. Term Table
    a. Berfungsi untuk menulis matriks yang bersesuaian tadi sesuai judulnya ke dalam suatu csv yang akan terintegrasi dengan pustaka Pandas.
    b. Memiliki parameter input matriks dan judul yang dihasilkan pada Query Table.
8. Read Web (**Bonus**)
    a. Melakukan web scraping sesuai kategori yang telah dipilih pada halaman web.
    b. Membaca keseluruhan elemen yang mengandung tag <p> dan <a> untuk kemudian diolah mirip seperti dokumen lokal pada umumnya.
    c. Melakukan pembersihan stemming sekaligus pembentukan tupel kalimat web.
    d. Memiliki parameter input kategori untuk penentuan link yang akan dituju.
    e. Memiliki parameter output berupa judul, kalimat, dan jumlah link.

B. Fungsi Routing (Terhubung dengan Flask)

9. Menu Utama (‘/’)
    a. Membuka halaman utama untuk memilih apakah akan mencari dokumen dari lokal atau web.
    b. Redirect menuju menu_utama.html.
10. About us (‘data/about_us’)
    a. Berisi data kelompok, cara pengerjaan.
    b. Redirect menuju aboutus.html.
11. Home(‘/local’)
    a. Berisi search box untuk mencari txt secara lokal.

    b. Redirect menuju index.html apabila search query tidak ada isi.
    c. Redirect ke result.html apabila hasilnya ada.
12. Web Home (‘/web’)
    a. Berisi search box untuk mencari web secara global.
    b. Redirect menuju webindex.html apabila search query tidak ada isinya.
    c. Redirect menuju webresult.html apabila hasilnya ada.
13. Web Result (‘/websearch/<kat>/<res>’)
    a. Fungsi Utama dalam pencarian hasil query pada routing.
    b. Memiliki parameter input berupa kategori pilihan berita dan input query.
    c. Hasil akan ditampilkan pada webresult.html dan akan langsung tertuju ke website yang diinginkan.
14. Upload Form(‘/upload’)
    a. Fungsi untuk melakukan upload pada website.
    b. File yang diupload akan secara otomatis masuk ke folder test.
    c. Redirect ke upload.html.
15. Result (‘/search/<res>’)
    a. Fungsi Utama dalam pencarian hasil query pada routing.
    b. Memiliki parameter input berupa input query.
    c. Hasil akan ditampilkan pada result.html dan akan langsung tertuju ke txt yang diinginkan secara lokal.
16. Table (‘/table’)
    a. Berfungsi untuk membaca csv hasil search lokal
    b. Redirect ke website termtable.html
    c. Menggunakan pustaka pandas
17. Web Table (‘/webtable’)
    a. Berfungsi untuk membaca csv hasil search website
    b. Redirect ke website webtermtable.html
    c. Menggunakan pustaka pandas
18. Uploaded file (‘test/<filename>’)
    a. Berfungsi untuk membaca file txt yang sudah diupload hasil hyperlink
    b. Memiliki parameter input nama file yang akan dibaca.
19. Upload file (‘/’)
    a. Membaca upload file berupa format txt
    b. Menyimpan hasil upload file ke folder test
    c. Redirect ke local

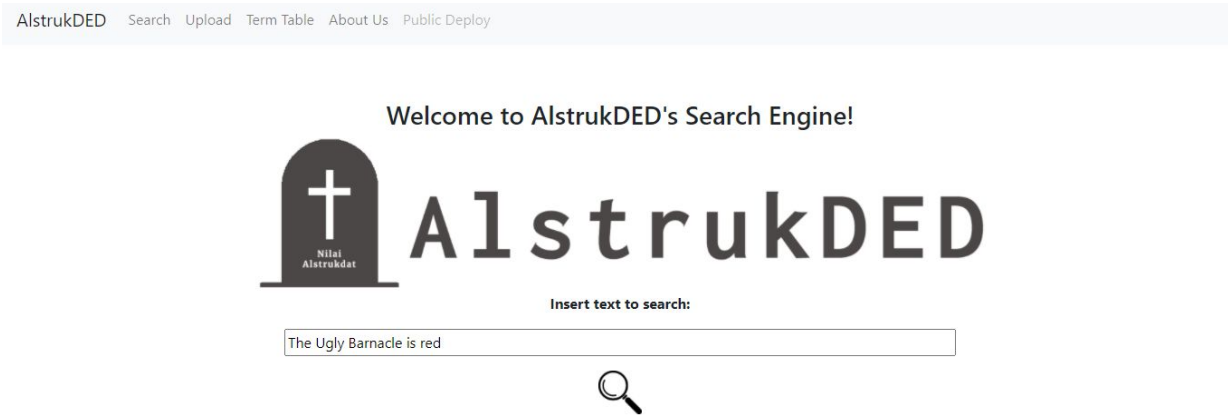
BAB 4. EKSPERIMEN



Gambar 2. Menu utama website

Pada menu utama akan ada pilihan file lokal dan web scraping. Klik file lokal untuk ke search engine lokal dan klik web scraping untuk ke search engine web scraping.

4.1 Search Engine Web Lokal

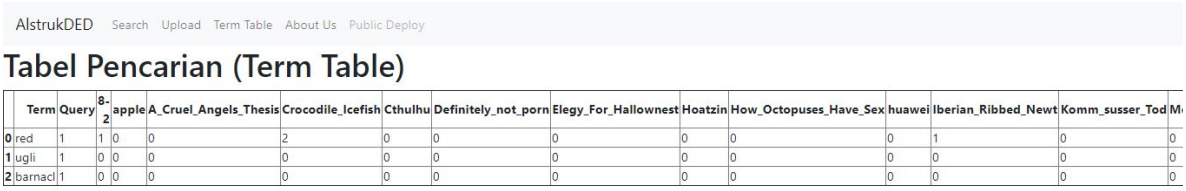


Gambar 3. Halaman untuk search engine untuk file lokal

Ini adalah halaman search file lokal. Untuk search lokal ini kamu memilih untuk menggunakan bahasa inggris untuk file-file lokal dan query. Ketik query yang ingin di cari di search bar itu, contohnya di sini diisi dengan query “The Ugly Barnacle is red”. Lalu klik tombol search yang berupa magnifying glass atua tekan enter di keyboard. Tunggu hingga loadingnya selesai.

Gambar 4. Halaman search result file lokal

Setelah loading selesai, maka akan ditampilkan halaman berupa hasil search query. Ditunjukkan file-file lokal dari yang memiliki tingkat kemiripan tertinggi. Dalam kasus ini, itu adalah file "The_Ugly_Barnacle", lalu file "Transistor", dan selanjutnya sampai file lokal terakhir.

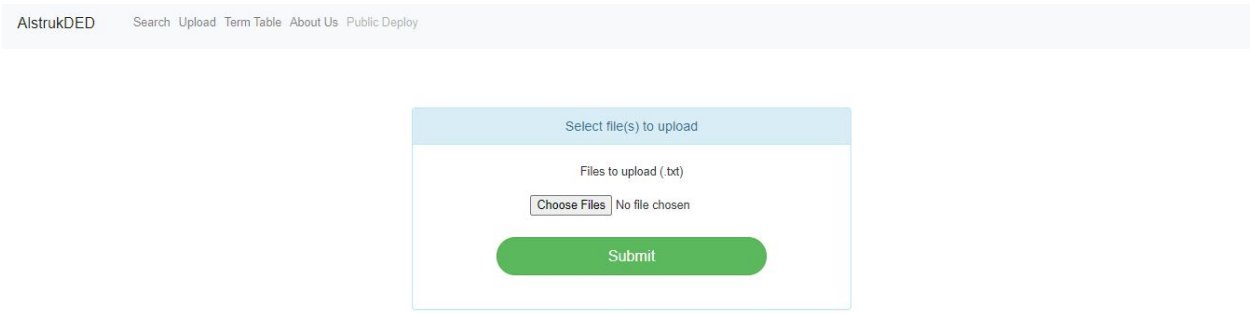


	Term	Query	8-2	apple	A_Cruel_Angels_Thesis	Crocodile_Icefish	Cthulhu	Definitely_not_porn	Elegy_For_Hallownest	Hoatzin	How_Octopuses_Have_Sex	huawei	Iberian_Ribbed_Newt	Komm_susser_Tod	M
0	red	1	1	0	0	2	0	0	0	0	0	0	1	0	0
1	ugli	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	barnacl	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Gambar 5. Term table file lokal

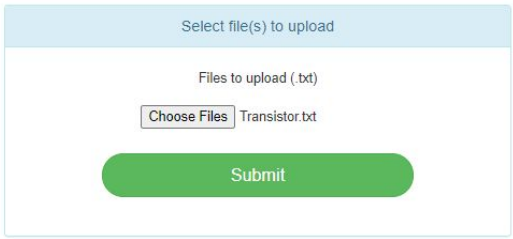
Term table dari search lokal tadi dapat dilihat dari tab Term Table. Dapat dilihat termnya adalah dari query yang telah di-clean, makanya tidak ada kata 'The' dan 'is'. Kata 'ugly' dan 'barnacle' juga terpotong atau berubah karena adanya stemming. Term table akan berubah setiap kali ada search dengan file lokal.

Untuk melakukan pencarian lagi klik tab search atau klik link home di halaman search result. Untuk kembali ke menu utama klik tulisan AlstrukDED.



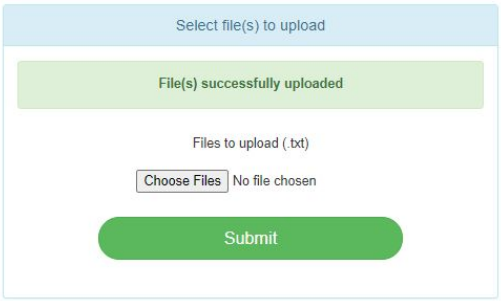
Gambar 6. Halaman upload file lokal

Untuk mengupload file ke search file lokal ini, klik tab upload. Lalu klik choose file dan pilih file .txt yang ingin di upload ke pencarian. Lalu klik submit.



Gambar 7. Contoh file (Transistor.txt) yang sudah terpilih untuk diupload

Untuk contoh ini, file Transistor.txt akan diupload.



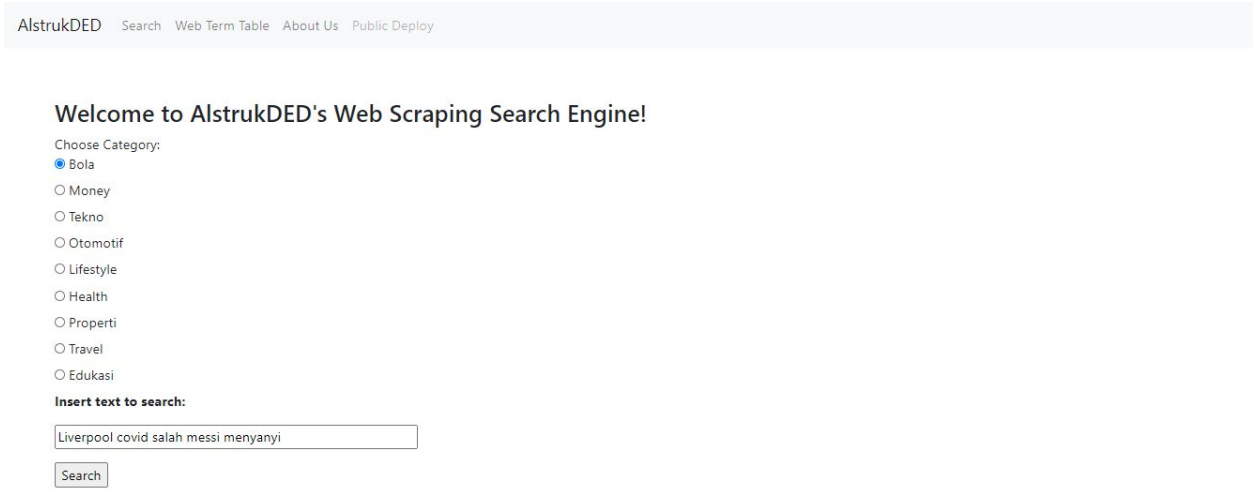
Gambar 8. Contoh upload yang berhasil

Jika upload berhasil, maka akan keluar tulisan File(s) successfully uploaded seperti ini.

Gambar 9. File yang diupload berhasil masuk ke search engine

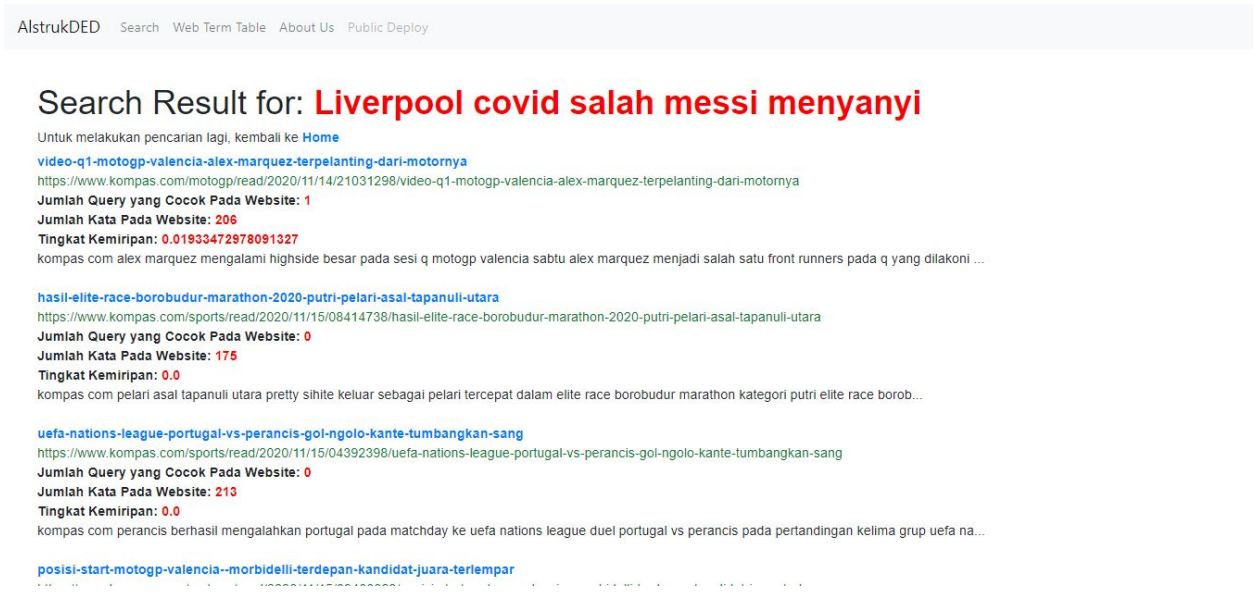
Dapat dilihat bahwa file Transistor.txt sudah ada dan sudah bisa di search.

4.2 Search Engine Web Scraping



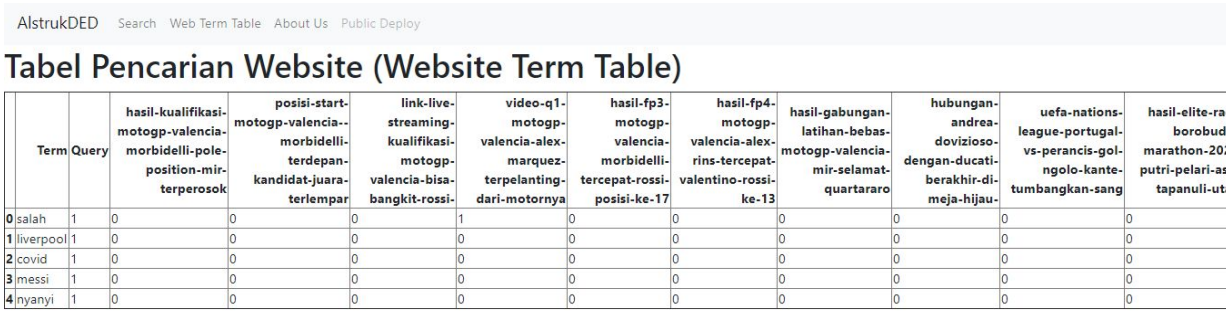
Gambar 10. Halaman search engine web scraping

Ini adalah halaman search dengan web scraping. Dalam ini digunakan web kompas.com sebagai sumber web scraping. Dapat dipilih kategori berita yang ingin di search dari web kompas.com. Karena dari kompas.com, maka bahasa untuk search web scraping ini adalah bahasa Indonesia. Ketik query yang ingin di search di search bar. Dalam contoh ini, querynya adalah “Liverpool covid salah messi menyanyi”. Lalu klik tombol search atau tekan enter. Tunggu loading search.



Gambar 11. Search Result Web Scraping

Setelah loading selesai maka akan muncul halaman search result untuk web scraping. Disini ditunjukkan artikel-artikel dari web kompas.com berdasarkan kategori dan query, artikel hasil pencarian diurutkan berdasarkan kemiripan ke query.



	Term	Query	hasil-kualifikasi-motogp-valencia-morbidelli-pole-position-mir-terperosok	posisi-start-motogp-valencia--morbidelli-terdepan-kandidat-juara-terlempar	link-live-streaming-kualifikasi-motogp-valencia-bisa-bangkit-rossi-	video-q1-motogp-valencia-alex-marquez-terpelanting-dari-motornya	hasil-fp3-motogp-valencia-morbidelli-tercepat-rossi-posisi-ke-17	hasil-fp4-motogp-valencia-alex-rins-tercepat-valentino-rossi-ke-13	hasil-gabungan-latihan-bebas-motogp-valencia-mir-selamat-quartararo	hubungan-andrea-dovizioso-dengan-ducati-berakhir-di-meja-hijau-	uefa-nations-league-portugal-vs-perancis-gol-ngolo-kante-tumbangkan-sang	hasil-elite-ra borobud marathon-20 putri-pelari-a tapanuli-ut
0	salah	1	0	0	0	1	0	0	0	0	0	0
1	liverpool	1	0	0	0	0	0	0	0	0	0	0
2	covid	1	0	0	0	0	0	0	0	0	0	0
3	messi	1	0	0	0	0	0	0	0	0	0	0
4	nyanyi	1	0	0	0	0	0	0	0	0	0	0

Gambar 12. Term Table Search Web Scraping

Term table untuk web scraping dapat dilihat di tab Web Term Table. Dapat dilihat termnya adalah query yang sudah di clean, oleh karena itu kata ‘menyanyi’ berubah menjadi kata ‘nyanyi’. Web Term Table ini akan berubah setiap kali ada search dengan web scraping.

Untuk melakukan search dengan web scraping lagi, klik tab search atau klik link home di halaman search result. Untuk kembali ke menu utama, klik tulisan AlstrukDED.

BAB 5. KESIMPULAN

5.1 Kesimpulan

Program ini berhasil kami buat tanpa campur tangan kelompok lain. Program berhasil mencari query pada file lokal maupun pada website dengan menggunakan Web Scraping dan berhasil menunjukkan tingkat similaritas query. Akan tetapi, website ini belum ter-deploy untuk publik.

5.2. Saran Pengembangan

Bisa ditambah fitur pencarian video, gambar, maupun pencarian audio untuk kemutakhiran pencarian. Bisa ditambah juga saran rekomendasi untuk tipografi yang salah, rekomendasi berita terbaru, dll. Stemming kata masih bisa dikembangkan lagi terlebih dari filter alphabet, stopwords, dan bagian2 akhir kata seperti dengan filter bahasa, perbedaan penulisan bahasa inggris UK US, dll.

5.3. Refleksi

Kami dapat menggunakan waktu untuk mengerjakan Tubes Algeo dengan baik. Bahkan, pada saat penutupan pembagian kelompok, backend kami sudah berjalan dengan optimal.

Daftar Pustaka

1. https://informatikalogi.com/sistem-temu-kembali-informasi/
2. https://informatika.stei.itb.ac.id/~rinaldi.munir/AljabarGeometri/2015-2016/Makalah-2015/Makalah-IF2123-2015-016.pdf
3. https://ejournal.unsrat.ac.id/index.php/informatika/article/view/13752/13332