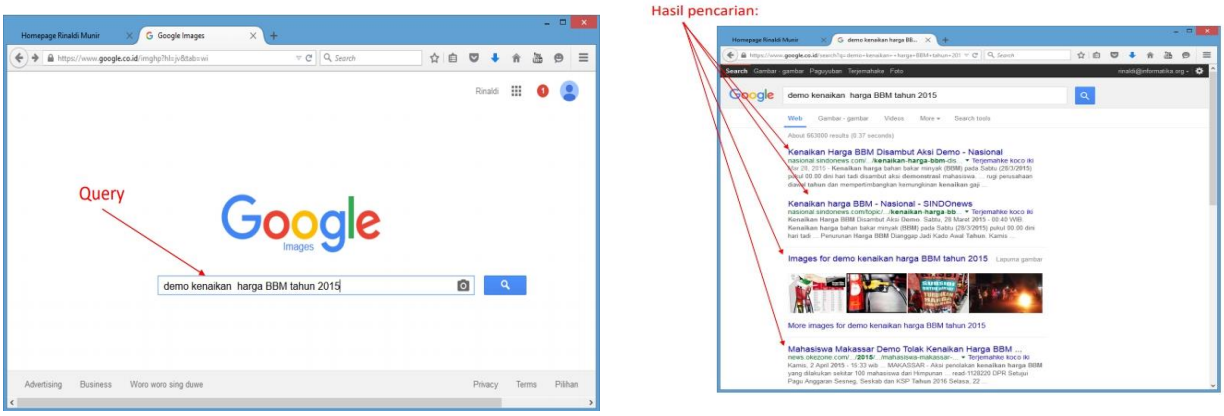
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung

Tugas Besar 2 IF 2123 Aljabar Linier dan Geometri
Aplikasi Dot Product pada Sistem Temu-balik Informasi
Semester I Tahun 2020/2021

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

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:

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

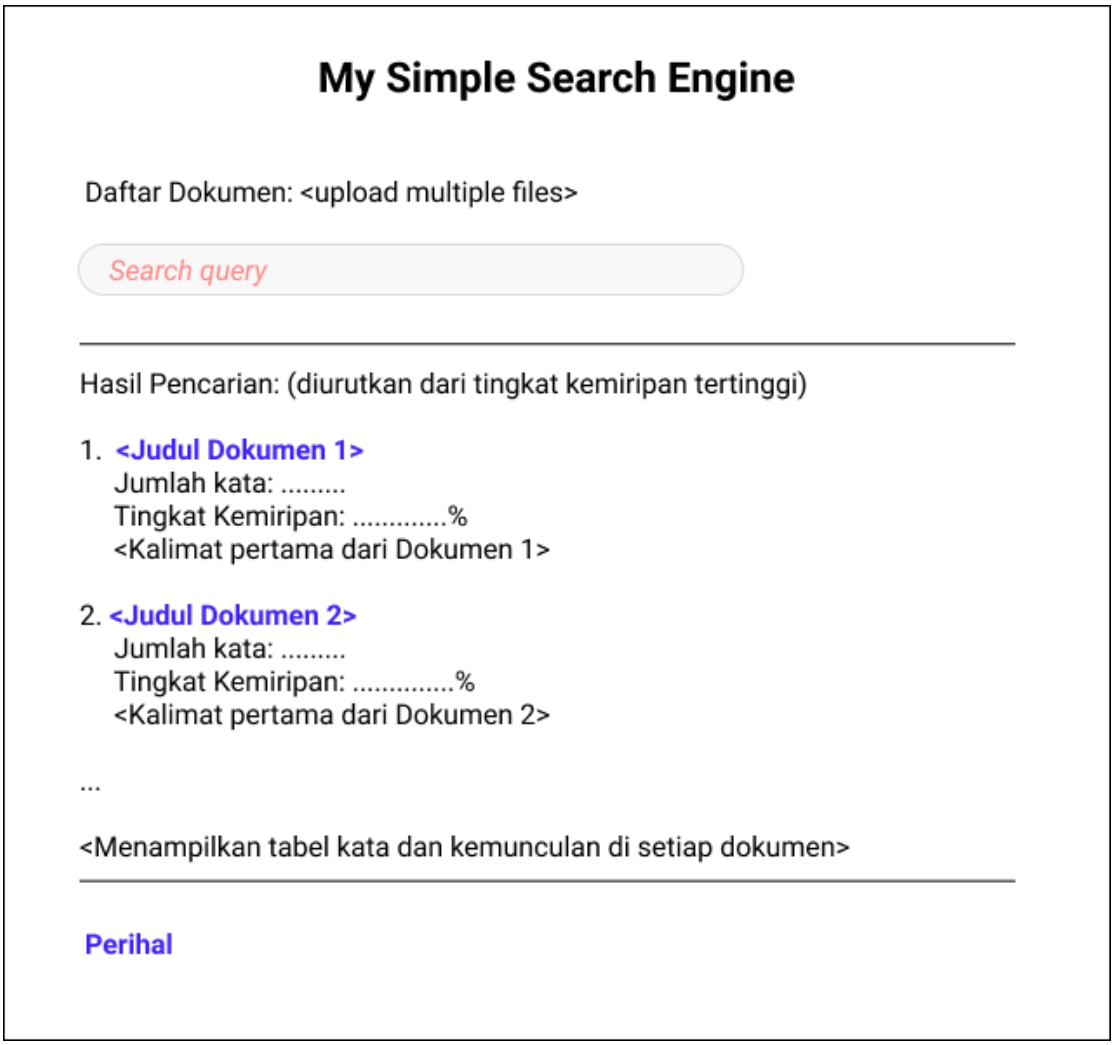
Pada kesempatan ini, kalian ditantang untuk membuat sebuah *search engine* sederhana dengan model ruang vector dan memanfaatkan cosine similarity.

PENGGUNAAN PROGRAM

Berikut ini adalah input yang akan dimasukkan pengguna untuk eksekusi program.

1. **Search query**, berisi kumpulan kata yang akan digunakan untuk melakukan pencarian
2. **Kumpulan dokumen**, dilakukan dengan cara mengunggah multiple file ke dalam web browser.

Tampilan layout dari aplikasi web yang akan dibangun adalah sebagai berikut.



Gambar 2. Tampilan layout dari aplikasi web search engine yang dibangun.

Perihal: link ke halaman tentang program dan pembuatnya (Konsep singkat *search engine* yang dibuat, How to Use, About Us).

Catatan: Teks yang diberikan warna **biru** merupakan hyperlink yang akan mengalihkan halaman ke halaman yang ingin dilihat. Apabila menekan *hyperlink* <Judul Dokumen 1>, maka akan diarahkan pada sebuah halaman yang berisi *full-text* terkait dokumen 1 tersebut (seperti *Search Engine*).

Anda dapat menambahkan menu lainnya, gambar, logo, dan sebagainya. Tampilan Front End dari website dibuat semenarik mungkin selama mencakup seluruh informasi pada layout yang diberikan di atas.

Data uji berupa dokumen-dokumen yang akan diunggah ke dalam web browser. Format dan extension dokumen dibebaskan selama bisa dibaca oleh web browser (misalnya adalah dokumen dalam bentuk file *txt* atau file *html*). Minimal terdapat 15 dokumen berbeda.

Tabel term dan banyak kemunculan term dalam setiap dokumen akan ditampilkan pada web browser dengan layout sebagai berikut.

Term	**Query**	**D1**	**D2**	**...**	**D3**
Term1					
Term2					
...					
TermN					

Untuk menyederhanakan pembuatan search engine, terdapat hal-hal yang perlu diperhatikan dalam eksekusi program ini.

1. Silahkan lakukan stemming dan penghapusan *stopwords* pada setiap dokumen
2. Tidak perlu dibedakan antara huruf-huruf besar dan huruf-huruf kecil.
3. *Stemming* dan penghapusan stopword dilakukan saat **penyusunan vektor**, sehingga halaman yang berisi *full-text* terkait dokumen tetap seperti semula.
4. Penghapusan karakter-karakter yang tidak perlu untuk ditampilkan (jika menggunakan *web scraping* atau format dokumen berupa html)
5. Bahasa yang digunakan dalam dokumen adalah bahasa Inggris atau bahasa Indonesia (pilih salah satu)

Petunjuk: silahkan gunakan library sastrawi atau nltk untuk stemming kata dan penghapusan stopwords

SARAN PENGERJAAN

Anda disarankan untuk membuat program testing pada **backend** terlebih dahulu untuk menguji keberhasilan dari perhitungan cosine similarity tersebut.

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 f*ramework* 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.

PROSEDUR PENGERJAAN

1. Tugas dikerjakan secara berkelompok yang terdiri dari 3 orang dan **tidak boleh sama dengan anggota kelompok tubes sebelumnya.** Kelompok dipilih secara mandiri. Daftarkan kelompok anda via tautan http://bit.ly/PendataanTubesAlgeo1 sebelum Minggu, 1 November 2020 pukul 23.59 WIB. **Perhatikan bahwa jika tidak mendaftar maka anda akan dimasukkan ke kelompok baru secara acak.**
2. Tugas ini dikumpulkan hari Senin, 16 November 2020 paling lambat pukul 12.55 WIB. Setelah deadline tersebut., silahkan menghubungi asisten **maksimal** Jumat, 20 November 2020 pukul 22.00 WIB untuk demo program yang telah dibuat.
3. Tanya jawab dilakukan dengan mengisi sheet QnA pada http://bit.ly/PendataanTubesAlgeo1
4. **Dilarang keras menyalin program dari sumber lain (buku, internet, program kakak tingkat, program kelompok lain)**

LAPORAN

Laporan terdiri dari:

1. *Cover*: *Cover* laporan ada foto anggota kelompok (foto bertiga kalau ada, atau foto masing-masing, bebas gaya). Foto ini menggantikan logo "gajah" ganesha.
2. Bab 1: Deskripsi masalah (dapat meng-*copy paste* spesifikasi tugas ini).
3. Bab 2: Teori singkat mengenai retrieval information, vektor dan cosine similarity
4. Bab 3: Implementasi program meliputi struktur *class* yang didefinisikan (atribut dan method), garis besar program, dll.
5. Bab 4: Eksperimen. Bab ini berisi hasil eksekusi program terhadap contoh-contoh kasus berikut analisis hasil eksekusi tersebut
6. Bab 5: Kesimpulan, saran, dan refleksi (hasil yang dicapai, saran pengembangan, dan refleksi anda terhadap tugas ini).

7. Tuliskan juga referensi (buku, web), yang dipakai/diacu di dalam Daftar Referensi.

Keterangan laporan dan program:

a) Laporan ditulis dalam bahasa Indonesia yang baik dan benar, tidak perlu panjang tetapi tepat sasaran dan jelas.
b) Identitas per halaman harus jelas (misalnya : halaman, kode kuliah).
c) *Listing* program tidak perlu disertakan pada laporan.
e) Program disimpan di dalam *repository github* dengan nama repository Algeo02-xxxxx. Lima digit terakhir adalah NIM anggota terkecil. Didalam *repository* tersebut terdapat empat folder: bin, src, test dan doc yang masing-masing berisi
   - Folder *src* berisi *source code* dari program dan website
   - Folder *test* berisi dokumen uji.
   - Folder *doc* berisi laporan

Sertakan juga *readme* yang dibuat sebaik mungkin. Silahkan gunakan template https://github.com/ritaly/README-cheatsheet atau template lain sebagai referensi. Pastikan *repository* bersifat ***private*** dan telah mengundang asisten yang pembagiannya akan diumumkan kemudian.

PENGUMPULAN TUGAS

1. Yang diserahkan saat pengumpulan tugas adalah:
   a) Laporan (*soft copy*)
   b) Kode program
   Kedua komponen tersebut di-*commit* sebelum Senin 16 November 2020 pukul 12.55 WIB. Commit setelah waktu tersebut akan mendapat pengurangan nilai.
2. Program dapat dijalankan. Asisten pemeriksa hanya akan melakukan *setting* atau kompilasi sesuai dengan *readme* agar program dapat berjalan. Program yang tidak dapat dijalankan tidak akan diberi nilai.
3. Laporan dikumpulkan juga melalui google form melalui tautan http://bit.ly/LaporanAlgeo2. format penamaan file adalah sebagai berikut: [Nomor Kelompok]_LaporanTugasBesar2IF2123_[Nama Kelompok].pdf

PENILAIAN

Komposisi penilaian umum adalah sebagai berikut :

1. Program: 80 %
2. Laporan : 20 %

REFERENSI

Perhatikan bahwa referensi hanya ada sebagai sumber belajar, bukan sebagai standar apa yang kami ekspektasi dari kode kalian. Banyak sumber lain, jangan menutup diri ke hanya sumber sumber di bawah ini.

"Aplikasi Dot Product pada sistem temu balik aplikasi" by Rinaldi Munir
https://informatika.stei.itb.ac.id/~rinaldi.munir/AljabarGeometri/2020-2021/Algeo-12-Aplikasi-dot-product-pada-IR.pdf

"Create A Simple Search Engine Using Python" by Irfan Alghani Khalid
https://link.medium.com/yEtxO932Kab

"Implementing the TF-IDF Search Engine" by Kartheek Akella
https://link.medium.com/UcdqCt92Kab

“Create APIs with python and flask” by programming historian
https://programminghistorian.org/en/lessons/creating-apis-with-python-and-flask
Link ini untuk belajar teori tentang API

“How to create a react flask project” by Miguel Grinberg
https://blog.miguelgrinberg.com/post/how-to-create-a-react--flask-project
Contoh project kecil react + flask

“Removing Stop Words with NLTK Python”
https://www.geeksforgeeks.org/removing-stop-words-nltk-python/#:~:text=What%20are%20Stop%20words%3F,result%20of%20a%20search%20query

“Sastrawi Library”
https://pypi.org/project/Sastrawi/