## **APLIKASI DOT PRODUCT PADA SISTEM**

## **TEMU-BALIK INFORMASI**

## **LAPORAN TUGAS BESAR**

## Diajukan Untuk Memenuhi Tugas IF 2123 Aljabar Linier dan Geometri

## oleh

## **LOUIS RIEMENN 13519016**

## **MUHAMMAD RIFKY MUTHAHHARI 13519123**

**KEVIN KATSURA D. SITANGGANG 13518216**

**Sebuah gambar berisi bangunan, orang, luar ruangan, trotoar

Deskripsi dihasilkan secara otomatisSebuah gambar berisi orang, pria, pakaian, berdiri

Deskripsi dihasilkan secara otomatisSebuah gambar berisi orang, pria, berdiri, memegang

Deskripsi dihasilkan secara otomatis**

## **TEKNIK INFORMATIKA**

## **INSTITUT TEKNOLOGI BANDUNG**

## **BANDUNG**

## **2020**

## **BAB I**

## **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?

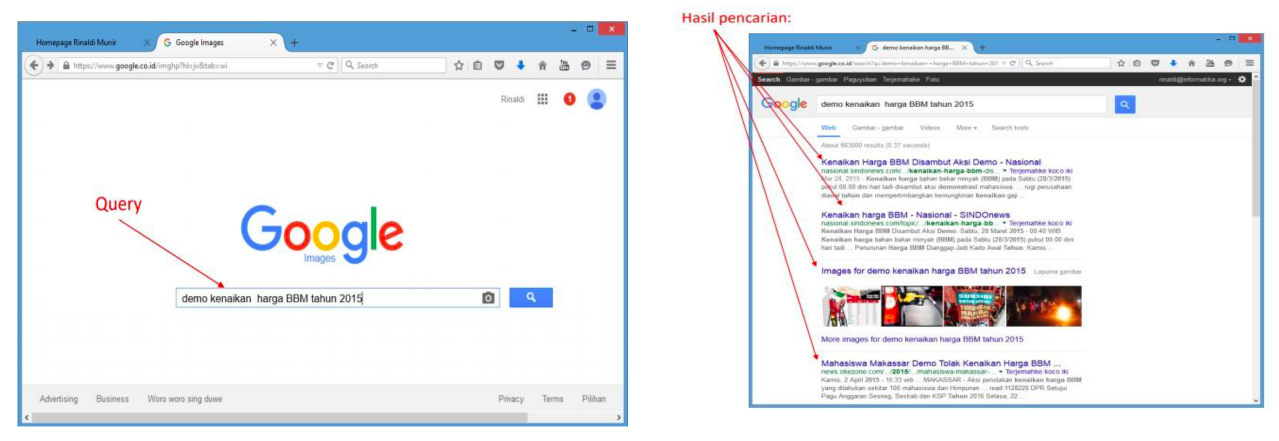
Sebagaimana yang telah diajarkan di dalam kuliah pada materi vector di ruang

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.

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 Rn, 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:



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.

Sebuah gambar berisi teks

Deskripsi dihasilkan secara otomatis

*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.

Sebuah gambar berisi meja

Deskripsi dihasilkan secara otomatis

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 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.

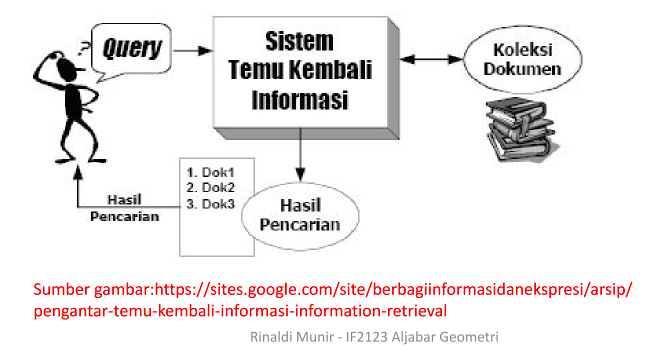
## **BAB II**

## **TEORI SINGKAT**

**2.1. Temu-balik Informasi**

• Temu-balik informasi (information retrieval): menemukan kembali (retrieval) informasi yang relevan terhadap kebutuhan pengguna dari suatu

kumpulan informasi secara otomatis.



• IR tidak sama dengan pencarian di dalam basisdata (database)

• IR umumnya digunakan pada pencarian informasi yang isinya tidak terstruktur

• Informasi terstruktur: tabel-tabel di dalam basisdata (database)

Sebuah gambar berisi meja

Deskripsi dihasilkan secara otomatis

• Informasi tak-terstruktur:

- dokumen (isinya bergantung pembuatnya)

- laman web (webpage)



• Aplikasi IR : search engine

Sebuah gambar berisi teks

Deskripsi dihasilkan secara otomatis

Sebuah gambar berisi teks

Deskripsi dihasilkan secara otomatis

Sebuah gambar berisi teks

Deskripsi dihasilkan secara otomatis

**2.2. INFORMATION RETRIEVAL DENGAN MODEL RUANG VEKTOR**

• Salah satu model IR adalah model ruang vektor

• Model ini menggunakan teori di dalam aljabar vector

• 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 = (w1, w2, ..., wn) di dalam Rn.

• wi = bobot setiap kata i di dalam query atau dokumen

• Nilai wi dapat menyatakan jumlah kemunculan kata tersebut dalam

dokumen (term frequency)

**Contoh**: Misalkan terdapat tiga buah kata (T1, T2, dan T3), dua buah dokumen (D1 dan D2) serta sebuah query Q. Masing-masing dinyatakan sebagai vector:

**D1** = (2, 3, 5), **D2** = (3, 7, 1), **Q** = (0, 0, 2)

**D1** = (2, 3, 5) artinya dokumen D1 mengandung 2 buah kata T1, 3 buah

kata T2, dan 5 buah kata T3.

Contoh: Misalkan *T1 = Menteri, T2 = minta, T3 = Korupsi*

**D1** = Menteri olahraga meminta maafatas perbuatan korupsi. Menteri tersebut terlibat korupsi anggaran. Meminta-minta komisi termasuk korupsi. Korupsi sudah mandarah daging di Indonesia. Korupsi sudah menjadi budaya.

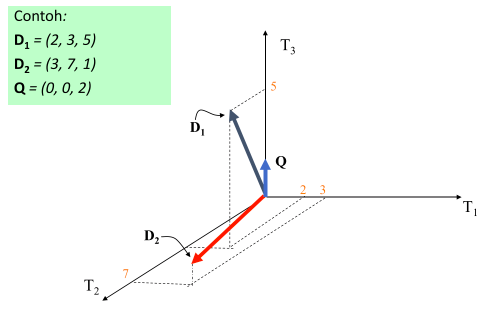
**D2** = (3, 7, 1) artinya dokumen D2 mengandung 3 buah kata T1, 7 buah kata T2, dan satu buah kata T3.

Contoh: *D2 = Gubernur Jabar meminta waktu ketemu Menteri Sosial. Dia meminta Pak Menteri mengunjungi panti. Permintaan yang wajar. Sekretaris Gubernur mengirim surat permintaan kepada Menteri tersebut. Apakah meminta-minta termasuk perbuatan korupsi? Tidak selalu, bukan? Meminta waktu saja.*

**Q** = (0, 0, 2) artinya query Q hanya mengandung 2 buah kata T3.

Contoh: Q = *Korupsi besar atau kecil tetap saja korupsi.*

**2.3. REPRESENTASI GRAFIK VEKTOR**

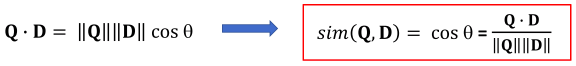
****

• Penentuan dokumen mana yang relevan dengan 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.

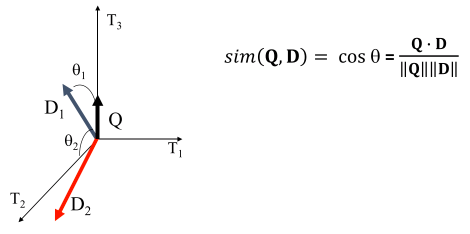
• Kesamaan (sim) antara dua vektor Q = (q1, q2, …, qn) dan D = (d1, d2, …, dn) diukur dengan rumus cosinus similarity yang merupakan bagian dari rumus perkalian titik (dot product) dua buah vektor:



dengan **Q D** adalah perkalian titik yang didefinisikan sebagai

Sebuah gambar berisi teks

Deskripsi dihasilkan secara otomatis



• Jika cos  = 1, berarti  = 0, vektor Q dan D berimpit, yang berarti dokumen D sesuai dengan query Q.

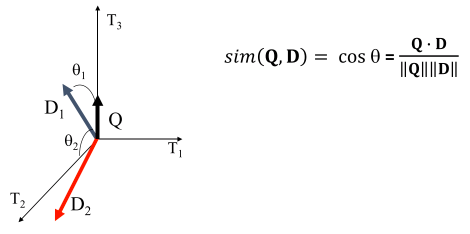
• Jadi, nilai cosinus yang besar (mendekati 1) mengindikasikan bahwa dokumen cenderung sesuai dengan query.

• Setiap dokumen di dalam koleksi dokumen dihitung kesamaannya dengan *query* dengan rumus cosinus di atas.

• Selanjutnya hasil perhitungan di-ranking berdasarkan nilai cosinus dari besar ke kecil sebagai proses pemilihan dokumen yang yang “dekat” dengan query.

• Pe-ranking-an tersebut menyatakan dokumen yang paling relevan hingga yang kurang relevan dengan query.

• Nilai cosinus yang besar menyatakan dokumen yang relevan, nilai cosinus yang kecil menyatakan dokumen yang kurang relevan dengan query.



Sebuah gambar berisi meja

Deskripsi dihasilkan secara otomatis

## **BAB III**

## **IMPLEMENTASI PROGRAM**

**3.1. index.html**

|  |  |
| --- | --- |
| **CLASS / ID / NAME** | **FUNGSI** |
| navbar navbar-expand-lg navbar-light container-fluid bg-dark my-6 | Navigator bar sebagai wadah Home button dan About Us. |
| navbar-toggler  navbar-toggler-icon  collapse navbar-collapse  navbarNavAltMarkup  navbar-nav  nav-link active text-info font-weight-bold  sr-only  nav-link text-info | Tombol untuk menampilkan Home button dan About Us. |
| container-fluid  main-section | Bagian tengah atau main dari page. |
| ext-title text-light | Wadah tulisan “Documents Search Engine” |
| form-group my-2 mx-8 text-center container  row  form-inline my-2  main-search | Wadah Search Engine. |
| form-control mr-sm-2 rounded-pill col-sm-11 | Tempat memasukkan input kata. |
| btn btn-outline-info my-2 my-sm-0 rounded-pill col-sm-1 | Tombol search. |
| container-fluid text-light bg-dark | Pemisah bagian tengah dan bawah. |
| second-section bg-dark text-light | Bagian bawah web utama. |
| container-fluid  upload-section  bg-info text-light  fas fa-file-upload fa-4x  custom-message  selected-file-span | Wadah upload file dan tulisan “choose a file”. |
| form-control-file  fileHandler | Tombol choose a file. |
| file-submit  bg-info text-light | Tombol upload. |

**3.2. aboutus.html**

|  |  |
| --- | --- |
| CLASS / ID / NAME | FUNGSI |
| navbar navbar-expand-lg navbar-light container-fluid bg-dark my-6 | Navigator bar sebagai wadah Home button dan About Us. |
| navbar-toggler-icon  navbar-toggler  collapse navbar-collapse  navbarNavAltMarkup  navbar-nav  nav-link text-info  sr-only  active nav-link text-info font-weight-bold | Tombol untuk menampilkan Home button dan About Us. |
| container-fluid bg-dark  story-section | Container main of aboutus. |
| continer px-2  text-title text-light text-left | Container tulisan “About This Website”. |
| container  row  card bg-info col-md-12 col-xs-12 col-lg-12 col-sm-12  card-body | Container isi dari About this Webiste. |
| card-title | Judul “Overview dan Seach Engine with Cosine Similarity(h5 tag)”. |
| card-text | Text (Isi Overview/Cosine Similarity). |

## **3.3 index.js**

|  |  |
| --- | --- |
| **FUNGSI** | **KETERANGAN** |
| isMember(member) | Memberikan nilai kebenaran true jika member merupakan anggota setKata (setKata merupakan array) |
| loaderSelected | Fungsi untuk menghapus keterangan “choose file” ketika file telah dimasukkan. |
| showResultHeader | Menampilkan kalimat “Your Search On (Kata yang dicari)” ketika tombol search di klik. |
| getMatriksTerm | Membuat matriks berisi term |
| getMatriksDocuments | Membuat matriks berisi term untuk setiap dokumen |
| getCosineSimilarity | Menghitung kesamaan kosinus |
| printResult | Menuliskan hasil ke html |

## **BAB IV**

## **EKSPERIMEN**

4.1 Tampilan Utama Website

Sebuah gambar berisi elektronik, orang, komputer, depan

Deskripsi dihasilkan secara otomatis

4.2. Tekan tombol “Choose a file”. Setelah itu masukkan file .txt yang akan jadi sumber.

Sebuah gambar berisi teks

Deskripsi dihasilkan secara otomatis

4.3. Ketika file telah dipilih, maka akan terlihat nama file yang telah dipilih di bawah tombol Choose a file sebelumnya. Kata “Choose a file” pada tombol tersebut juga akan hilang

Sebuah gambar berisi orang, cuplikan layar, komputer, monitor

Deskripsi dihasilkan secara otomatis

4.4. Maka pada local akan muncul file yang di upload pada folder test.

Sebuah gambar berisi cuplikan layar, monitor, layar, duduk

Deskripsi dihasilkan secara otomatis

4.5. Setelah menekan tombol search, maka info mengenai kata yang di input pada file yang diupload akan tertera. Pada result akan tersedia : jumlah kata pada file.txt, similarity isi file dengan kata input, dan kalimat pertama pada file.

Sebuah gambar berisi teks

Deskripsi dihasilkan secara otomatis

4.6. Ketika ADT STACK.txt ditekan (yang bertulisan biru), maka halaman akan berpindah ke isi dari ADT STACK.txt tsb

Sebuah gambar berisi teks

Deskripsi dihasilkan secara otomatis

4.7. Pada Navigator akan ada tombol about us yang membawa kita pada halaman seperti di bawah ini.

Sebuah gambar berisi cuplikan layar, monitor, layar, computer

Deskripsi dihasilkan secara otomatis

## **BAB V**

## **PENUTUP**

## **5.1 Kesimpulan**

## Dari tugas besar berjudul “Aplikasi Dot Product pada Sistem Temu-balik Informasi”, maka website yang kami buat dapat menerima input file dari pengguna, dan menampilkan informasi seperti jumlah kata pada file, similarity, dan first sentence sesuai kata yang diinput.

## **5.2 Saran**

## Berikut ini adalah saran-saran yang dapat kami berikan untuk tugas besar ini.

## 1. Dalam pembuatan program, sebaiknya membuat function dan procedure kecil dalam file yang berbeda sehingga dapat dimanfaatkan kapan saja pada kondisi lain. Selain dapat dimanfaatkan kapan saja, penggunaan function dan procedure ini dapat memperindah dan memperjelas struktur kode.

## **5.3 Refleksi**

1. Dalam pengerjaan tugas sebaiknya tidak dikerjakan mendekati deadline pengumpulan tugas.
2. Dalam pengerjaan tugas sebaiknya dilakukan secara mencicil sehingga tidak membebani aktivitas perkuliahaan lainnya.
3. Jika mengerjakan tugas dengan waktu sisa yang terbatas sebaiknya dapat belajar lebih cepat dan efisien.
4. Pembagian tugas harus seimbang dan tidak timpang supaya tidak membebani.

## **REFERENSI**

Informatika.stei.itb.ac.id. (2020). Tugas Besar 2 IF 2123 Aljabar Linier dan Geometri. Diakses pada 1 November 2020, dari <http://informatika.stei.itb.ac.id/~rinaldi.munir/AljabarGeometri/2020-2021/Tubes2-Algeo-2020.pdf>.

Informatika.stei.itb.ac.id. (2020). Aplikasi Dot Product pada IR. Diakses pada 13 November 2020, dari <http://informatika.stei.itb.ac.id/~rinaldi.munir/AljabarGeometri/2020-2021/Algeo-12-Aplikasi-dot-product-pada-IR.pdf>.