

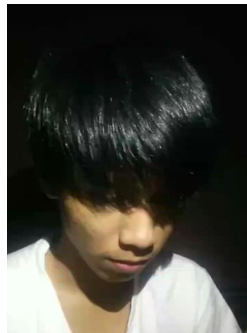
TUGAS BESAR 2 IF2123 ALJABAR LINEAR DAN GEOMETRI
Aplikasi Dot Product pada Sistem Temu-balik Informasi
Semester I Tahun 2020/2021

LAPORAN TUGAS

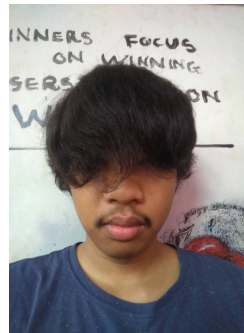
Diajukan sebagai laporan dari tugas besar dua mata kuliah Aljabar Linier dan Geometri
IF2123 pada Semester I Tahun Akademik 2020-2021

oleh

Juan Louis Rombetasik 13519075
Aria Bachrul Ulum Berlian 13519115



13519075



13519115

Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
2020

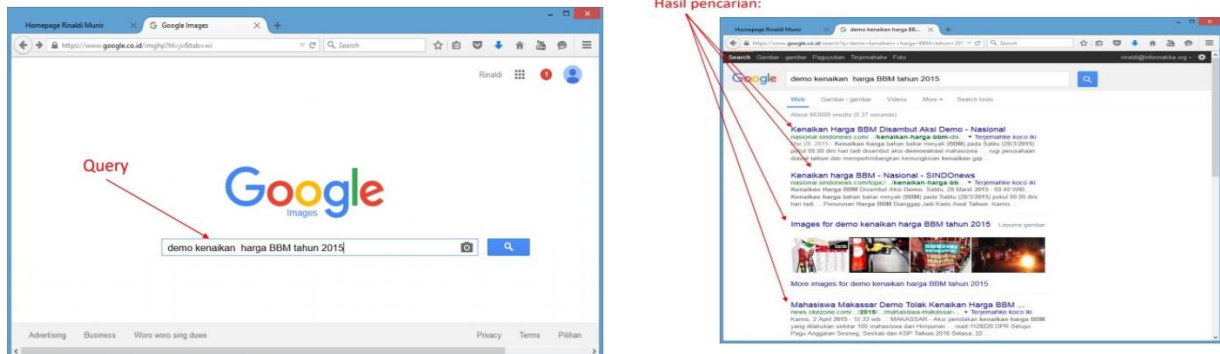
BAB I

Deskripsi Masalah

I. 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 $w = (w_1, w_2, \dots, w_n)$ 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 vektor dan memanfaatkan cosine similarity.

BAB II

Teori Singkat

1 Vektor

Vektor adalah suatu besaran yang memiliki nilai dan arah. Vektor biasanya ditulis dalam cetak tebal, seperti **a** atau **b**.

Salah satu cara pengoperasian vektor adalah dengan dot product yaitu:

$$a \cdot b = \sum_{i=1}^n a_i b_i \quad \mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta,$$

Dimana:

a = vektor pertama

b = vektor kedua

n = dimensi dari ruang vektor

a_i = komponen dari vektor **a**

b_i = komponen dari vektor **b**

$\|\mathbf{a}\|$ = norma vektor a

$\|\mathbf{b}\|$ = norma vektor b

θ = sudut antara dua vektor

2 Sistem Temu Balik Informasi

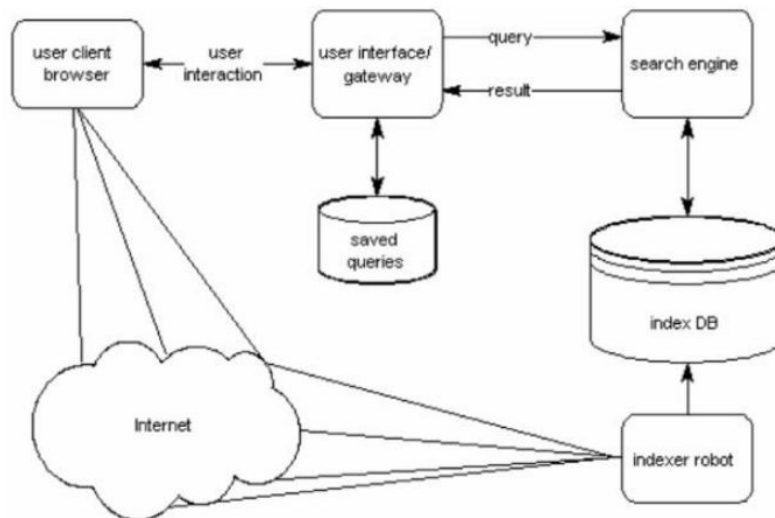
Temu Balik Informasi atau yang lebih sering dikenal sebagai *Information retrieval* dalam dunia informatika adalah salah satu contoh implementasi aplikasi dot vektor dalam kehidupan sehari-hari.

Information Retrieval ilmu yang mempelajari prosedur-prosedur dan metode-metode untuk menemukan kembali informasi yang tersimpan dari berbagai sumber (resources) yang relevan atau koleksi sumber informasi yang dicari atau dibutuhkan. Dengan tindakan index (indexing), panggilan (searching), pemanggilan data kembali (recalling).

Indexing atau pengindeksan merupakan proses membangun basis data indeks dari koleksi dokumen. Indexing dilakukan terhadap dokumen sebelum pencarian dilakukan. Dewasa ini, sistem pengindeksan secara manual mulai digantikan oleh sistem pengindeksan otomatis. Sistem pengindeksan otomatis dapat dilakukan dengan implementasi Natural Language Processing (NLP). NLP bertujuan untuk memahami arti yang diberikan dalam bahasa alami dan memberikan respon yang sesuai, misalnya dengan melakukan suatu aksi tertentu atau menampilkan data tertentu. Adapun tahapan dari pengindeksan adalah sebagai berikut:

1. Parsing dokumen yaitu proses pengambilan kata-kata dari kumpulan dokumen. 2. Stoplist yaitu proses pembuangan kata buang seperti: tetapi, yaitu, sedangkan, dan sebagainya. 3. Stemming yaitu proses penghilangan/ pemotongan dari suatu kata menjadi bentuk dasar. 4. Term Weighting dan Inverted File yaitu proses pemberian bobot pada istilah.

Berikut adalah skema umum cara kerja sistem temu balik informasi, dan algoritma text processing.



Gambar 2. Cara kerja sistem temu balik informasi

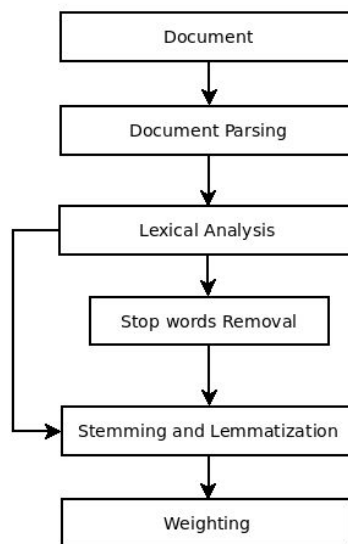


Fig. Processing of Textual data

Gambar 3. Cara text processing

Pembobotan term dilakukan dengan menghitung frekuensi kemunculan istilah term/kata di dalam suatu dokumen. Dalam proyek ini, metode pembobotan hanya dilakukan dengan **raw term frekuensi**. Sehingga **tidak ada** pembobotan lebih kepada istilah yang lebih penting seperti contohnya: nama orang, tempat dan sebagainya. Ada banyak cara pembobotan suatu term (TF-IDF) tapi kami tidak akan membahasnya disini karena kami nilai tidak terlalu relevan dengan mata kuliah yang sedang diajarkan.

Dalam proyek ini, kemunculan suatu term pada suatu dokumen disimpan dalam sebuah vektor(array), yaitu vektor dokumen[1,2,3...N]. Vektor tersebut bernilai jumlah kemunculan semua term terhadap dokumen tersebut, bernilai nol jika suatu term tidak muncul dalam dokumen tersebut, bernilai satu jika kemunculan term hanya satu kali dalam dokumen, dan seterusnya.

$$tf(t,d) = f_{t,d}$$

tf = term frequency

t = term

d = dokumen

$f_{t,d}$ = raw count term terhadap dokumen

Setelah didapat vektor dokumen dan vektor query, nilai kesamaan antara dua dokumen bisa dicari dengan mencari nilai sudut kosinus antara dua vektor, yakni vektor query (**Q**) dan vektor dokumen (**D**). Pengurutan dokumen dilakukan dengan mengurutkan nilai kesamaan (kosinus similaritas) dari yang terbesar.

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

BAB III

Implementasi Program

Program ini adalah aplikasi web yang berbasis python django. Views dari django akan menyimpan file yang diunggah apabila *request method*-nya adalah post. Semua file yang diunggah akan disimpan dalam folder media, tepatnya di src/truckFinder/media. Apabila didapati *request method*-nya adalah get, maka akan dicek apakah ada query atau tidak. Bila tidak ada query, maka akan menampilkan kolom pencarian dan tombol upload dokumen saja di template. Bila ada query, views akan memanggil fungsi truckFinder dengan argumen query. Data yang dikembalikan dari fungsi truckFinder akan dikirimkan ke dalam template. Data tersebut adalah tuple yang berisi listDok, termArray, vektorQuery, vektorDokumen, dan tableFrekuensi. listDok adalah array dua dimensi dengan barisnya adalah nomor dokumen dan kolomnya adalah nama file, jumlah kata, tuple yang berisi judul dan kalimat pertama, list isi dokumen yang sudah distem, dan nilai dari kosinus similiaritasnya. Baris listDok terurut berdasarkan kosinus similiaritasnya. termArray adalah array berisi term dari semua dokumen dan query. vektorQuery mengembalikan array yang berisi jumlah kemunculan kata dalam query relatif terhadap term. vektorDokumen adalah array dua dimensi dimana barisnya adalah nomer dokumen dan kolomnya berisi jumlah kemunculan kata dalam dokumen relatif terhadap term. tableFrekuensi adalah tabel dengan kolom term, vektorQuery, vektorDokumen.

Pada algoritma search engine truckFinder, kami mengimplementasi teknik information retrieval. Semua kata dan dokumen akan di stemming dan dilakukan penghilangan stopwords(bahasa). Pembuatan vektor query dan vektor dokumen dilakukan dengan perhitungan jumlah kata kemunculan (term frequency raw). Ranking dokumen dilakukan dengan mengurutkan dari kosinus similaritas yang terbesar hingga terkecil.

Fungsi antara yang dipanggil dalam fungsi truckFinder digunakan untuk membantu implementasi algoritma. Fungsi antara tersebut adalah fungsi webscrap, file_peek, file_read, dan stemming. Fungsi webscrap menerima lokasi file dan mengembalikan string yang berisi dokumen bersih dari file html. Fungsi file_peek dan file_read menerima masukkan yang sama dengan webscrap. Fungsi file_peek mengembalikan judul dan kalimat pertama dokumen. Fungsi file_read mengembalikan isi dokumen. Fungsi stemming menerima string berisi dokumen dan mengembalikan array kata unik yang telah distem dan dihilangkan stopwordsnya.

\

Perbedaan utama pada algoritma search-engine kami(truckFinder) adalah kami tidak memakai database untuk semua term dan dokumen dan tidak memakai query ekspansi untuk memudahkan pembuatan program. Pembuatan term dilakukan secara lokal sewaktu ada query yang di-inputkan. Algoritma ini menghilangkan banyak kesusahan karena tidak perlu membuat

database, tetapi kekurangannya adalah program memakan waktu yang cukup lama karena pembuatan data frame lokal untuk setiap kali panggilan.

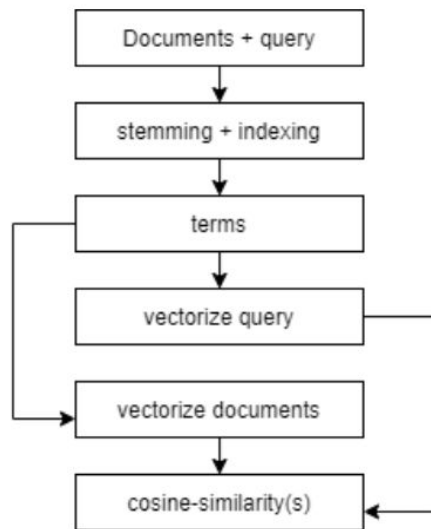


Fig. truckFinder Algorithm
(Information retrieval)

Gambar 4. Cara kerja truckFinder

BAB IV Eksperimen

1. Query: Sangkuriang Memakan Ikan

127.0.0.1:8000/?q=sangkuriang+memakan+ikan

Cari dokumen...

Upload dokumen

Petualangan Ikan Terbang
Ada ikan, bisa terbang | Jumlah kata: 11 | Tingkat Kemiripan : 31%

Dampak Buruk Junk Food Untuk Kesehatan Tubuh
Junk Food disebut makanan instan atau makanan cepat saji yang kini telah berkembang pesat di persaingan perusahaan makanan di Indonesia

Cerita sangkuriang
Cerita Sangkuriang bermula ketika raja yang bernama Sungging Perbangkara sedang pergi berburu ke hutan

Diet Berakhir Jeruji
Adalah Joe, yang hanya bisa mengejar tukang bakso dengan pandangannya yang pilu

Matahari Pun Tak Bosan
Ku bangkit setelah lama ambil posisi jongkok menyaksikan kejadian yang menimpa embun

Tinggal Kenangan
Pagi itu sangatlah cerah, mentari pagi muncul memancarkan sinar cerah dengan semangat 67 eh semangat 45 maksudnya

Fenomena Lelaki Tulen Berpenampilan Bagai Seorang Wanita
Bagi Anda yang mengikuti perkembangan berita masa kini khususnya berita di kalangan selebriti pastinya sudah sering dibahas lelaki seperti wanita

JUAN LOUIS & ARIA

Semut yang Pindah Rumah
"Maju

Surat Cinta dan Sebatang Coklat
Aku mengingat dari balik pohon beringin, agak jauh dari gadis itu

Telur Ceplok Penentu Sifat
Sebagian orang percaya bahwa zodiak adalah penentu sifat seseorang

Tabel Terms

Term	Query	Petualangan Ikan Terbang	Dampak Buruk Junk Food Untuk Kesehatan Tubuh	Cerita sangkuriang	Diet Berakhir Jeruji	Matahari Pun Tak Bosan	Tinggal Kenangan	Fenomena Lelaki Tulen Berpenampilan Bagai Seorang Wanita	Fenomena Pelakor (Perebut Laki Orang) atau Perebut Suami Orang	Giant Musul Setiap Antag
ikan	1	2	0	0	0	0	0	0	0	0
makan	1	0	13	1	4	1	1	0	0	0
sangkuriang	1	0	0	25	0	0	0	0	0	0

HOME CONCEPT HOW TO USE ABOUT US

JUAN LOUIS & ARIA

Dokumen yang tampil paling atas adalah Petualangan Ikan Terbang dengan tingkat kemiripan 31% dan jumlah kata 11. Urutan kedua ditempati oleh Dampak Buruk Junk Food dengan tingkat kemiripan 29% dengan jumlah kata 276. Diikuti oleh Cerita Sangkuriang dengan tingkat kemiripan 27% dengan jumlah kata 653. Hal ini dapat dipahami karena meskipun jumlah kemunculan kata 'makan' di dokumen Dampak Buruk ada 12, namun jumlah kata yang terdapat

2. Query: Kenangan manis di hari itu jadi alasan untuk kembali

halaman 9, IF2123

Mobilku adalah mobil tercepat di muka bumi ini

Petualangan Ikan Terbang

Ada ikan, bisa terbang, dan ada petualangan

n pilan ang	Fenomena Pelakor (Perebut Laki Orang) atau Perebut Suami Orang	Matahari Pun Tak Bosan	Pengaruh Pendidikan Terhadap Kualitas Anak Pedalaman	Kebakaran Hutan di Indonesia	Dampak Buruk Junk Food Untuk Kesehatan Tubuh	Surat Cinta dan Sebatang Coklat	Diet Berakhir Jeruji	Semut yang Pindah Rumah	Telur Ceplok Penentu Sifat	Mobil Pilihan	Petualangan Ikan Terbang
1	0	1	1	0	1	0	0	0	0	0	
0	3	0	1	2	3	4	0	0	0	0	
3	4	2	2	1	2	1	1	2	0	0	
0	1	0	0	0	1	0	1	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	

HOME CONCEPT HOW TO USE ABOUT US

JUAN LOUIS & ARIA

Dokumen Cerita Sangkuriang menempati posisi pertama dalam tingkat relevansi dengan query. Cerita Sangkuriang memiliki 2 kata ‘hari’ dan 13 kata ‘jadi’ sehingga tingkat relevansinya mencapai 11%.

3. Query: Tak sengaja lewat depan rumahmu, ku melihat ada tenda biru dihiasi indahnya janur kuning. Hati bertanya pernikahan siapa. Tak percaya, namun ini terjadi. Kau bersanding duduk di pelaminan. Air mata jatuh tak tertahankan. Kau khianati cinta suci ini. Tanpa undangan diriku kau lupakan. Tanpa utusan diriku kau tinggalkan.

www.oploverz.in

Laporan tubes2 alge...

Algeo02-19075/con...

Truck-kun Finder

Mamet 2020 - Goog...

Meet - jco-gpsq

127.0.0.1:8000/?q=Tak+sengaja+lewat+depan+rumahmu%2Cku+melihat+ada+tenda+biru+di+hiasi+indahny+janur+kuning.+Hati+bertanya+per...

Apps Bab 2.pdf Altium Designer 20... Black Clover - HQ C... Membuat Hash Stri... Menyamakan pass... [ULASAN] 5 Point P... Operasi File - Pytho...

TRUCK-KUN

Cari dokumen...

Upload dokumen

Kebakaran Hutan di Indonesia

Indonesia adalah suatu negara dengan 4 Jumlah kata: 320 | Tingkat Kemiripan : 17%

Matahari Pun Tak Bosan

Ku bangkit setelah lama ambil posisi jongkok menyaksikan kejadian yang menimpa embun

Tinggal Kenangan

Pagi itu sangatlah cerah, mentari pagi muncul memancarkan sinar cerah dengan semangat 67 eh semangat 45 maksudnya

Fenomena Lelaki Tulen Berpenampilan Bagi Seorang Wanita

Bagi Anda yang mengikuti perkembangan berita masa kini khususnya berita di kalangan selebriti pastinya sudah sering dibahas lelaki seperti wanita

Surat Cinta dan Sebatang Coklat

Aku mengintip dari balik pohon beringin, agak jauh dari gadis itu

Pengaruh Pendidikan Terhadap Kualitas Anak Pedalaman

Bukan rahasia lagi bila anak-anak yang tinggal di daerah pedalaman sangat sulit mendapatkan kehidupan yang layak seperti anak-anak pada umumnya

Cerita sangkuriang

HOME CONCEPT HOW TO USE ABOUT US

JUAN LOUIS & ARIA

www.oploverz.in x Laporan tubes2 alge... x Algeo02-19075/con... x Truck-kun Finder x Mamet 2020 - Goo... x Meet - jco-gpsq... x

127.0.0.1:8000/?q=Tak+sengaja+lewat+depan+rumahmu%2C+ku+meliha... 127.0.0.1:8000/?q=Tak+sengaja+lewat+depan+rumahmu%2C+ku+meliha... 127.0.0.1:8000/?q=Tak+sengaja+lewat+depan+rumahmu%2C+ku+meliha...

Apps Bab 2.pdf Altium Designer 20... Black Clover - HQ C... Membuat Hash Stri... Menyamakan pass... [ULASAN] 5 Point P... Operasi File - Pytho...

Petualangan Ikan Terbang

Ada ikan, bisa terbang, dan ada petualangan

Tabel Terms

Term	Query	Kebakaran Hutan di Indonesia	Matahari Pun Tak Bosan	Tinggal Kenangan	Fenomena Lelaki Tulen Berpenampilan Bagi Seorang Wanita	Surat Cinta dan Sebatang Coklat	Pengaruh Pendidikan Terhadap Kualitas Anak Pedalaman	Cerita sangkuriang	Giant Musuh Setiap Antagonis	Diet Berakhir Jeruji
air	1	0	1	1	0	4	1	2	0	0
biru	1	0	0	1	0	0	0	0	0	0
cinta	1	0	1	2	0	3	0	0	0	0
depan	1	0	0	2	0	1	0	0	0	7
diri	2	5	0	0	2	1	0	1	1	3
duduk	1	1	1	0	0	1	0	0	0	0
hati	1	0	1	6	0	0	0	3	0	1
hias	1	0	0	0	0	0	0	0	0	0

HOME CONCEPT HOW TO USE ABOUT US

JUAN LOUIS & ARIA

Type here to search

16:06 14/11/2020

www.oploverz.in x Laporan tubes2 alge... x Algeo02-19075/con... x Truck-kun Finder x Mamet 2020 - Goo... x Meet - jco-gpsq... x

127.0.0.1:8000/?q=Tak+sengaja+lewat+depan+rumahmu%2C+ku+meliha... 127.0.0.1:8000/?q=Tak+sengaja+lewat+depan+rumahmu%2C+ku+meliha... 127.0.0.1:8000/?q=Tak+sengaja+lewat+depan+rumahmu%2C+ku+meliha...

Apps Bab 2.pdf Altium Designer 20... Black Clover - HQ C... Membuat Hash Stri... Menyamakan pass... [ULASAN] 5 Point P... Operasi File - Pytho...

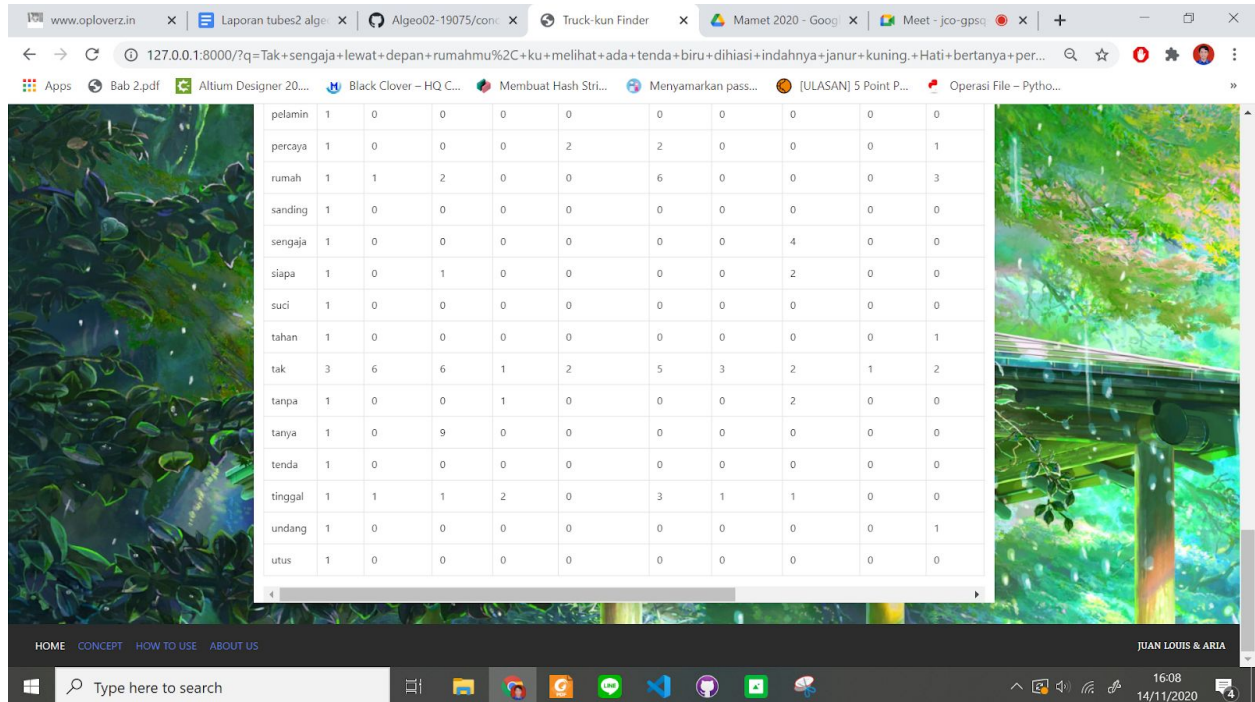
hias	1	0	0	0	0	0	0	0	0	0
indah	1	0	0	0	0	2	0	0	0	0
ini	1	1	1	0	0	0	0	0	2	1
jadi	1	2	4	3	5	2	2	13	4	1
janur	1	0	0	0	0	0	0	0	0	0
jatuh	1	0	0	3	0	1	0	2	0	0
kau	4	0	0	2	0	0	0	0	0	0
khianat	1	0	0	0	0	0	0	0	0	0
ku	1	0	2	0	0	0	0	0	0	0
kuning	1	0	1	0	0	0	0	0	0	0
lewat	1	0	1	1	0	0	0	0	0	1
lihat	1	0	2	3	0	5	0	2	0	3
lupa	1	0	0	1	0	0	0	0	0	0
mata	1	0	0	2	0	5	0	0	0	3
nikah	1	0	0	0	0	0	0	2	0	0
pelamin	1	0	0	0	0	0	0	0	0	0

HOME CONCEPT HOW TO USE ABOUT US

JUAN LOUIS & ARIA

Type here to search

16:07 14/11/2020



petamin	1	0	0	0	0	0	0	0	0	0
percaya	1	0	0	0	2	2	0	0	0	1
rumah	1	1	2	0	0	6	0	0	0	3
sanding	1	0	0	0	0	0	0	0	0	0
sengaja	1	0	0	0	0	0	0	4	0	0
siapa	1	0	1	0	0	0	0	2	0	0
suci	1	0	0	0	0	0	0	0	0	0
tahan	1	0	0	0	0	0	0	0	0	1
tak	3	6	6	1	2	5	3	2	1	2
tanpa	1	0	0	1	0	0	0	2	0	0
tanya	1	0	9	0	0	0	0	0	0	0
tenda	1	0	0	0	0	0	0	0	0	0
tinggal	1	1	1	2	0	3	1	1	0	0
undang	1	0	0	0	0	0	0	0	0	1
utus	1	0	0	0	0	0	0	0	0	0

Meskipun query sangat panjang, dokumen Kebakaran Hutan di Indonesia yang menempati urutan pertama hanya mendapat tingkat kemiripan 17% dengan jumlah kata 320.

Analisis:

Memang terkadang hasil pencarian tidak sesuai seperti apa yang kita inginkan (relevan) dikarenakan truckFinder hanya memakai *Term Frequency raw*.

Pada beberapa kasus, misal terdapat dokumen yang berisikan banyak kata, walaupun setiap kata pada query terdapat di dokumen, tidak selalu dokumen tersebut ditampilkan di urutan teratas. Hal ini dikarenakan oleh rumus cosine similarity, dimana semakin banyak kata di dokumen, semakin kecil pula cosine similaritynya.

BAB V

Kesimpulan, Saran dan Refleksi

5.1 Kesimpulan

Aplikasi dot vektor bisa diterapkan dalam kehidupan sehari-hari yaitu contohnya adalah sistem temu balik informasi (Information Retrieval)

terkadang hasil pencarian tidak sesuai seperti apa yang kita inginkan (relevan) dikarenakan truckFinder hanya memakai *Term Frequency raw*. tidak ada pembobotan lebih kepada istilah yang lebih penting seperti contohnya: nama orang, tempat dan sebagainya
Pada beberapa kasus, misal terdapat dokumen yang berisikan banyak kata, walaupun setiap kata pada query terdapat di dokumen, tidak selalu dokumen tersebut ditampilkan di urutan teratas. Hal ini dikarenakan oleh rumus cosine similarity, dimana semakin banyak kata di dokumen, semakin kecil pula cosine similaritynya.

5.2 Saran

Program harus bisa menghitung lebih cepat dengan mengurangi looping dsb, atau membuat database dan query expansion

Saran untuk tubes kedepannya: Tubesnya bikin yang lebih ‘seru’ lagi! Tubesnya kalau bisa anggotanya dua orang juga boleh....

5.3 Refleksi

Kami belajar banyak dari tubes ini, bahwa tubes harus kelar hari-7 sebelum deadline supaya tubes benar benar kelar, kasarnya kami menyelesaikan tubes ini seminggu sebelum pengumpulan dan kami temukan berbagai macam bug contohnya adalah tanda petik (left double quotation (“/U+201C) dan right double quotation mark (”/U+201D)) yang tidak umum ("/U+0022) membuat error sistem stemming kami. Dari tubes ini kami menyadari bahwa dua orang bekerja efektif dengan pembagian tugas yang benar lebih baik daripada banyak orang tetapi tidak sinkron dan tidak efektif. Komunikasi adalah faktor penting dalam suatu sinkronisasi pekerjaan.

BAB VI

Daftar Referensi

<https://stackoverflow.com/>

<https://www.w3schools.com/>

<https://medium.com/@ksnugroho/dasar-text-preprocessing-dengan-python-a4fa52608ffe>

<https://docs.djangoproject.com/en/3.0/>

https://www.youtube.com/watch?v=hPXNP1NoVNQ&list=PLZS-MHyEIRo6p_RwsWntxMO5QAqIHHHld&ab_channel=KelasTerbuka

Fitri, Meisya. 2013. *Perancangan Sistem **Temu balik informasi** dengan Metode Pembobotan Kombinasi TF-IDF untuk pencarian Dokumen berbahasa indonesia*. Jurnal Sistem dan Teknologi Informasi (JUSTIN)