

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



Kelompok 32:

13519070 - Mhd. Hiro Agayeff Muslion

13519072 - Hanif Arroisi Mukhlis

13519171 - Fauzan Yubairi Indrayadi

Asisten:

13518076 - Muhammad Ayyub Abdurrahman

## Bab 1

### Deskripsi Masalah

Membuat sebuah 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 2

### Teori Singkat

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  $w = (w_1, w_2, \dots, w_n)$  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:

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

### Bab 3

### Implementasi Program

Dibawah ini ditampilkan tabel dari seluruh modul, kelas dan fungsi yang digunakan pada search engine ini.

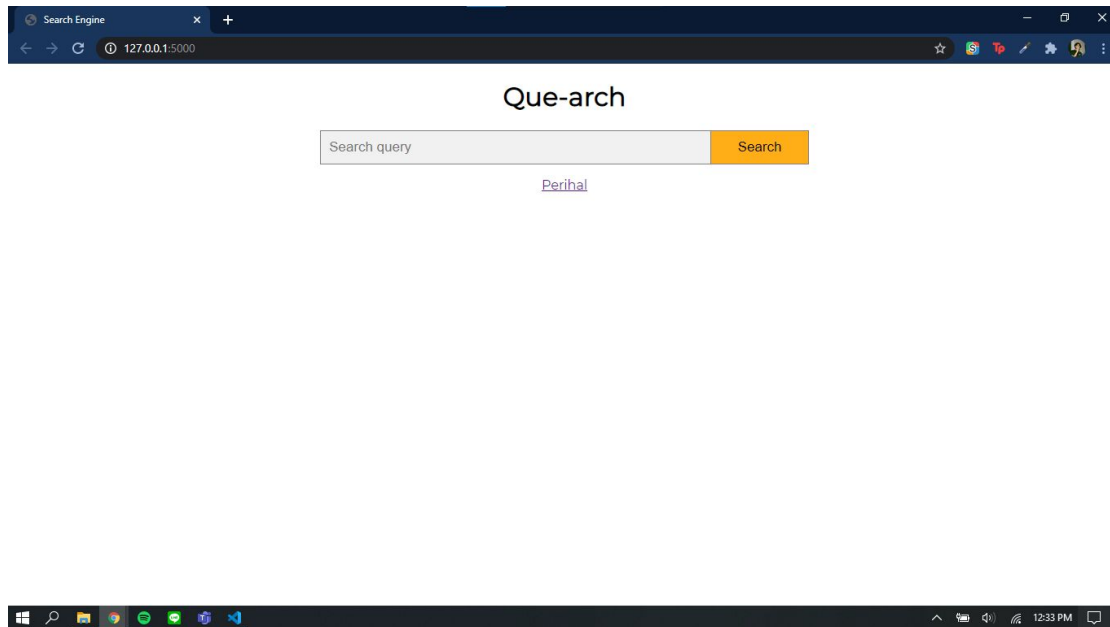
Nama	Tipe	Kegunaan
vektorkata	Modul	Mendefinisikan vektor kata dan operasi yang dapat dilakukan.
vektorkata.VektorKata	Kelas	Kelas abstraksi vektor kata. Menggunakan <i>set</i> sebagai basis implementasi.
vektorkata.similarity	Fungsi	Menghitung <i>cosine similarity</i> antara dua vektor. <i>Error</i> jika diberikan vektor kosong.
vektorkata.VektorBuilder	Kelas	Membangun vektor dari kumpulan kata. Menggunakan <i>stemming</i> .
vektorkata.load_vektor	Fungsi	Membaca vektor kata dari sebuah <i>stream</i> .
vektorkata.save_vektor	Fungsi	Menulis isi vektor kata ke sebuah <i>stream</i> .
html_getter	Modul	Membaca kata dari dokumen HTML.
html_getter.Parser	Kelas	Kelas spesialisasi yang akan memproses dokumen HTML.
html_getter.get_url	Fungsi	Membaca dokumen HTML dari URL dan membersihkan dari tanda baca, dll.
html_getter.get_url_unclean	Fungsi	Membaca dokumen HTML dari URL tanpa pembersihan tanda baca.
html_getter.get_file	Fungsi	Membaca dokumen HTML dari <i>file</i> dan membersihkan dari tanda baca, dll.
html_getter.get_file_unclean	Fungsi	Membaca dokumen HTML dari <i>file</i> tanpa pembersihan tanda baca.
__main__	Modul utama	Inisialisasi program.
__test__	Modul <i>unit test</i>	Tes implementasi vektor kata.

app	Modul	Aplikasi Flask.
app.all_words	Fungsi	Menghasilkan daftar semua kata di <i>query</i> , termasuk <i>search term</i> .
app.to_list	Fungsi	Menghasilkan <i>list</i> yang berisi 1 jika kata ada di vektor, dan 0 jika sebaliknya. Berguna untuk pembentukan tabel.
app.sort_similarity	Fungsi	Membandingkan <i>search term</i> dengan daftar <i>query</i> dan menghasilkan <i>list</i> terurut tingkat <i>similarity</i> .

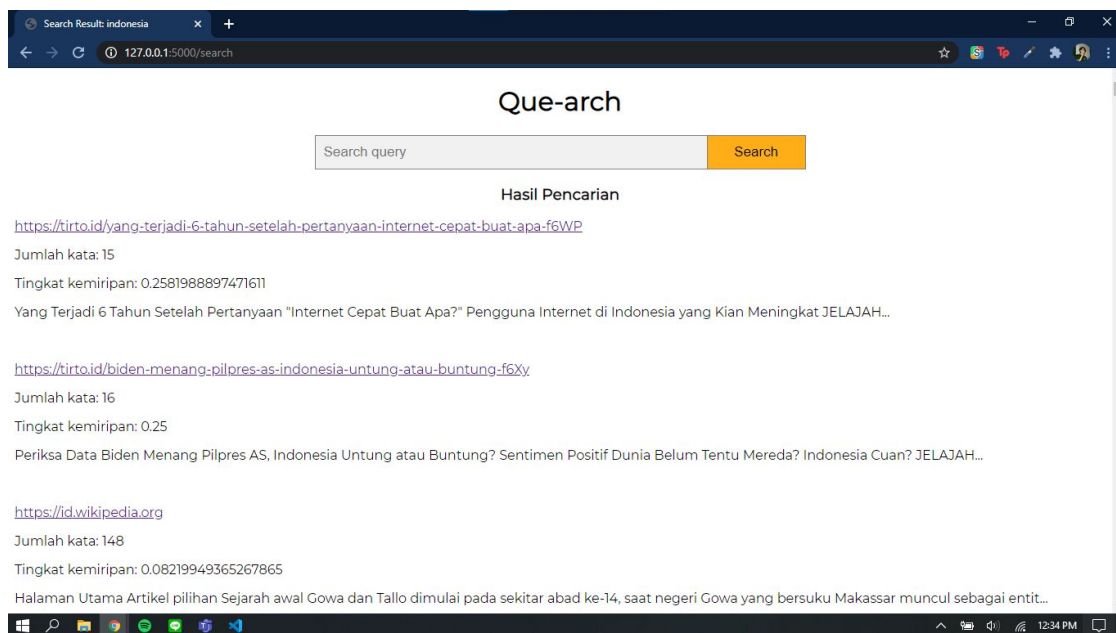
## Bab 4

### Eksperimen

#### Start page



#### Page hasil pencarian



## Tabel term

Search Result: indonesia

127.0.0.1:5000/search

in	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
inang	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
inangdan	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
incertae	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
indah	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
independen	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
india	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
indian	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
indikasi	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
individu	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
indo-eropa	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
indonesia	1	1	1	1	0	1	1	1	0	1	1	1	0	1	0	1
induk	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
industri	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
infeksi	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
infeksius	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
infektan	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
influenza	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
info	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
informasi	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0
ingat	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
inggris	0	0	0	0	0	0	1	0	1	1	0	0	0	0	1	1
inggrisadwal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1


## Perihal page

Perihal


127.0.0.1:5000/perihal

[Back](#)


### The Developer



**Hiro Agayeff**  
13519070  
Teknik Informatika



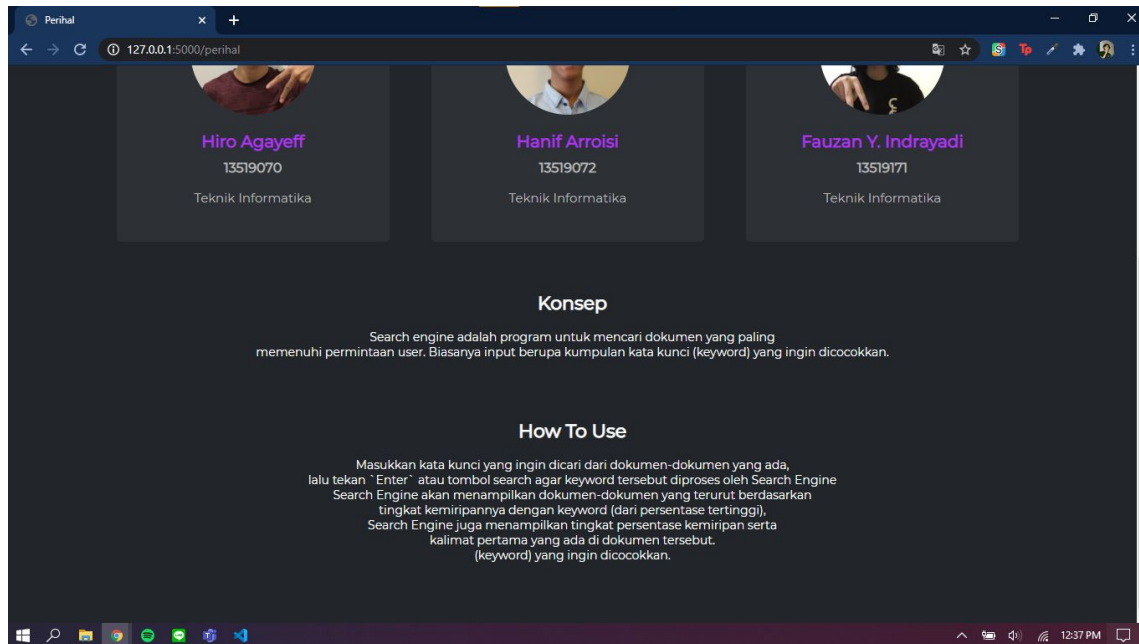
**Hanif Arroisi**  
13519072  
Teknik Informatika



**Fauzan Y. Indrayadi**  
13519171  
Teknik Informatika

### Konsep

Search engine adalah program untuk mencari dokumen yang paling



Klik pada dokumen





## **Bab 5**

### **Kesimpulan dan Saran**

#### **5.1 Kesimpulan**

Program “Que-arch” dapat menerima input string query berupa kata dasar maupun berimbuhan, lalu search engine akan memroses input query tersebut dan menampilkan output dokumen-dokumen dengan persentase kemunculan input query (term-frequency) dari dokumen-dokumen yang ada, jumlah kata pada masing-masing dokumen dan beberapa kata pertama dari masing-masing dokumen. Output disortir berdasarkan tingkat kemiripan kata-kata pada dokumen dengan input query (similarity measure) dari tingkat kemiripan tertinggi ke tingkat kemiripan terendah.

#### **5.2 Saran**

Sebaiknya terdapat pembekalan akan framework yang digunakan serta ditentukan dari awal jenis framework apa yang akan digunakan.

#### **5.3 Refleksi**

- Memahami materi ruang vektor lebih dalam
- Menambah pengetahuan implementasi ruang vektor pada level pemrograman
- Menambah pengetahuan pemrograman pada bahasa Python
- Menambah pengetahuan pemrograman pada bahasa Hypertext Markup Language (HTML)
- Menambah pengetahuan pemrograman pada bahasa Cascading Style Sheet (CSS)
- Time management dalam pelaksanaan tugas berkelompok

## REFERENSI

<https://www.tutorialspoint.com/css/index.htm>

<https://www.tutorialspoint.com/html/index.htm>

<https://informatika.stei.itb.ac.id/~rinaldi.munir/AljabarGeometri/2020-2021/Algeo-12-Aplikasi-dot-product-pada-IR.pdf>

<https://programminghistorian.org/en/lessons/creating-apis-with-python-and-flask>

<https://link.medium.com/yEtXO932Kab>