REPUBLIK INDONESIA
KEMENTERIAN HUKUM DAN HAK ASASI MANUSIA

SURAT PENCATATAN
CIPTAAN

Dalam rangka pelindungan ciptaan di bidang ilmu pengetahuan, seni dan sastra berdasarkan Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta, dengan ini menerangkan:

Nomor dan tanggal permohonan	:	EC002023105400, 6 November 2023
Pencipta		
Nama	:	**Lucia Dwi Krisnawati dan Aditya Wikan Mahastama**
Alamat	:	Kadirojo I RT008/RW002, Purwomartani, Kalasan, Sleman Daerah Istimewa Yogyakarta 55571, Kalasan, Sleman, DI Yogyakarta, 55571
Kewarganegaraan	:	Indonesia
Pemegang Hak Cipta		
Nama	:	**Lucia Dwi Krisnawati dan Aditya Wikan Mahastama**
Alamat	:	Kadirojo I RT008/RW002, Purwomartani, Kalasan, Sleman Daerah Istimewa Yogyakarta 55571, Kalasan, Sleman, DI Yogyakarta, 55571
Kewarganegaraan	:	Indonesia
Jenis Ciptaan	:	**Program Komputer**
Judul Ciptaan	:	**DWSiripTex Versi 1.0B**
Tanggal dan tempat diumumkan untuk pertama kali di wilayah Indonesia atau di luar wilayah Indonesia	:	1 Oktober 2023, di Yogyakarta
Jangka waktu pelindungan	:	Berlaku selama 50 (lima puluh) tahun sejak Ciptaan tersebut pertama kali dilakukan Pengumuman.
Nomor pencatatan	:	000538355

adalah benar berdasarkan keterangan yang diberikan oleh Pemohon.
Surat Pencatatan Hak Cipta atau produk Hak terkait ini sesuai dengan Pasal 72 Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta.

a.n. MENTERI HUKUM DAN HAK ASASI MANUSIA
Direktur Hak Cipta dan Desain Industri

Anggoro Dasananto
NIP. 196412081991031002

Disclaimer:
Dalam hal pemohon memberikan keterangan tidak sesuai dengan surat pernyataan, Menteri berwenang untuk mencabut surat pencatatan permohonan.

DWSiripTex

Buku Manual
Program Komputer – Versi 1.0B

DAFTAR ISI

1. Pendahuluan

1.1. Tentang Aplikasi DWSiripTex

DWSiripTex 1.0B adalah purwarupa program aplikasi yang akan digunakan untuk melakukan deteksi kemiripan tekstual dari dokumen-dokumen dengan dua bahasa yang berbeda. Pada purwarupa versi 1.0B ini, dokumen-dokumen tersebut menggunakan Bahasa Inggris dan Bahasa Indonesia.

DWSiripTex 1.0B menggunakan FAISS (Facebook AI Similarity Search) dengan *embedding* USE (Universal Sentence Encoder) dari Google. Dua mode FAISS dapat dipilih yaitu Flat2 (*brute force* atau *exhaustive search*) dan Inverted Index (diindeks kemudian dilakukan Inverted File Index (IVF) dan diklasterisasi berdasarkan *centroid*-nya).

1.2. Data Teknis

Nama Aplikasi	:	DWSiripTex
Versi	:	1.0B
Tanggal Rilis	:	01 Oktober 2023
Company	:	Fakultas Teknologi Informasi Universitas Kristen Duta Wacana
Pembuat	:	Lucia Dwi Krisnawati Aditya Wikan Mahastama

2. Petunjuk Penggunaan

2.1. Inisialisasi

Seluruh kode program terlampir harus dijalankan pada antarmuka Google Colaboratory. Inisialisasi dilakukan dengan menjalankan Cell 1 hingga 4 secara berurutan, dengan deskripsi sebagai berikut:

- Cell 1: Persiapan berupa impor pustaka-pustaka (*library*) yang diperlukan
- Cell 2: Load data dari *drive*. Data korpus dimaksud harus sudah diletakkan di Google Drive pada path yang disebutkan pada variabel **folder_path** dengan setting yang didefinisikan pada variabel **corpus_metadata**.
- Cell 3: Menentukan agar luaran proses disimpan pada file yang disebutkan pada variabel **folder_path_output.**
- Cell 4: Melakukan embedding dan membuat vektor dengan dimensi 512.

2.2. *Indexing* dan Pencarian Dokumen – FAISS Flat L2 Index

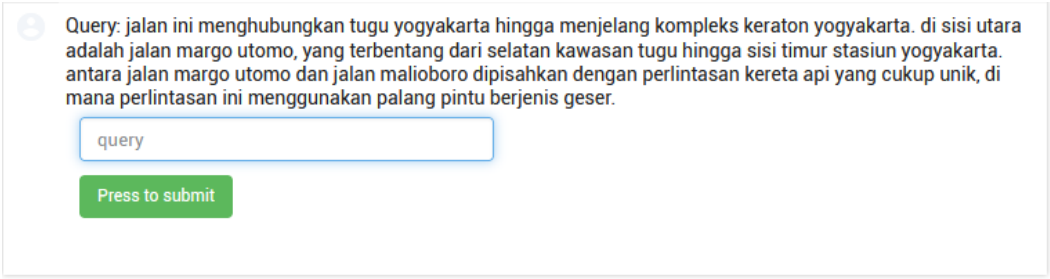
Untuk menggunakan *indexing* Faiss dengan *Flat L2 Index*, dimulai dengan menjalankan Cell 5, yang akan memberikan luaran berupa informasi hasil *indexing*

```
embeddings dimension :  512

Adding Indonesian embeddings to index

Adding English embeddings to index
<faiss.swigfaiss_avx2.IndexIDMap; proxy of <Swig Object of type
'faiss::IndexIDMapTemplate< faiss::Index > *' at 0x7a396c57b5d0>
>
```

Untuk mendeteksi kemiripan tekstual, pertama kali dimasukkan kueri berupa kalimat atau paragraf berisi teks yang akan diperiksa. Kueri dapat dalam Bahasa Indonesia atau Bahasa Inggris dimasukkan pada panel input pada Cell 6. Kueri akan ditampilkan pada panel tersebut, sebagaimana ditunjukkan oleh Gambar 1.



Gambar 1. Panel untuk masukan kueri

Jumlah hasil dokumen yang dikehendaki dimasukkan panel pada Cell 8 seperti ditunjukkan oleh Gambar 2, yaitu mengubah *slider* untuk variabel **num_results**.

Gambar 2. Panel untuk memasukkan jumlah hasil

Hasil deteksi dokumen akan ditampilkan di bawah Cell 8 dengan menampilkan jarak kemiripan serta id dokumen.

```
L2 distance: [2.917644966772548e-13, 0.2706255316734314,
0.5572736859321594, 0.6193249225616455, 0.6793750524520874,
0.7366632223129272, 0.8248614072799683, 0.8785932064056396,
0.9848678112030029, 1.0181164741516113]

 paper ID: [10042, 20042, 10041, 20041, 10081, 10045, 20045,
10082, 10043, 20043]
```

Jika Cell 9 dijalankan, akan menampilkan kalimat-kalimat dalam Bahasa Indonesia dan Inggris yang mirip yang ditemukan oleh kode pada Cell 8 secara urut berdasarkan hasil pencarian, sebagai contoh ditunjukkan di bawah ini.

```
this is index results [10042, 20042, 10041, 20041, 10081, 10045,
20045, 10082, 10043, 20043]
[10042] jalan ini menghubungkan tugu yogyakarta hingga menjelang
kompleks keraton yogyakarta. di sisi utara adalah jalan margo
utomo, yang terbentang dari selatan kawasan tugu hingga sisi
timur stasiun yogyakarta. antara jalan margo utomo dan jalan
malioboro dipisahkan dengan perlintasan kereta api yang cukup
unik, di mana perlintasan ini menggunakan palang pintu berjenis
geser.
[20042] this road connects tugu yogyakarta to the approach of the
yogyakarta palace complex. on the north side is jalan margo
utomo, which runs from the south of the tugu area to the east
side of yogyakarta station. jalan margo utomo and jalan malioboro
are separated by a unique railroad crossing, which uses a sliding
doorstop.
```

[10041] jalan malioboro jawa: ꦗꦭꦤ꧀ꦩꦭꦶꦪꦧꦺꦫꦺ, translit. dalan maliabara adalah nama salah satu kawasan jalan dari tiga jalan di kota yogyakarta yang membentang dari tugu yogyakarta hingga ke persimpangan titik nol kilometer yogyakarta. secara keseluruhan, kawasan malioboro terdiri atas jalan margo utomo, jalan malioboro, dan jalan margo mulyo. jalan ini merupakan poros garis imajiner kraton yogyakarta. itu terletak sumbu utara-selatan di garis antara kraton yogyakarta dan gunung merapi. ini sendiri penting bagi banyak penduduk lokal, orientasi utara-selatan antara istana dan gunung berapi menjadi penting.

[20041] malioboro street javanese: ꦗꦭꦤ꧀ꦩꦭꦶꦪꦧꦺꦫꦺ, translit. dalan maliabara is the name of one of the three streets in yogyakarta city that stretches from tugu yogyakarta to the intersection of yogyakartas zero kilometer point. overall, the malioboro area consists of jalan margo utomo, jalan malioboro, and jalan margo mulyo. this road is the axis of the yogyakarta kraton imaginary line. it lies north-south axis on the line between kraton yogyakarta and mount merapi. this in itself is important for many locals, the north-south orientation between the palace and the volcano being important.

[10081] tugu yogyakarta jawa: ꦠꦸꦒꦸꦔꦪꦺꦴꦒꦾꦏꦂꦠ, translit. tugu ngayogyåkartå adalah sebuah tugu atau monumen yang sering dipakai sebagai simbol atau lambang dari kota yogyakarta. tugu yang terletak di perempatan jalan jenderal sudirman dan jalan margo utomo ini, mempunyai nilai simbolis yang merupakan garis yang bersifat magis yang menghubungkan pantai parangtritis dan panggung krapyak di kabupaten bantul, keraton yogyakarta di kota yogyakarta dan gunung merapi di kabupaten sleman.
[10045] jalan malioboro sebenarnya hanya terbentang dari sisi selatan rel kereta api, di depan hotel grand inna hingga berakhir di pasar beringharjo sisi timur. dari titik ini, nama jalan berubah menjadi jalan margo mulyo hingga titik nol kilometer yogyakarta. jalan malioboro menjadi batas antara kemantren gedongtengen dan kemantren danurejan, di mana sisi barat malioboro adalah wilayah dari kemantren gedongtengen, dan sisi timur malioboro adalah wilayah dari kemantren danurejan. sedangkan seluruh sisi jalan margo utomo adalah wilayah dari kemantren jetis, dan sisi jalan margo mulyo adalah wilayah dari kemantren gondomanan.
[20045] malioboro street actually only stretches from the south side of the railroad tracks, in front of the grand inna hotel to end at pasar beringharjo on the east side. from this point, the street name changes to jalan margo mulyo until yogyakartas zero kilometer point. malioboro street becomes the boundary between gedongtengen kemantren and danurejan kemantren, where the west side of malioboro is the territory of gedongtengen kemantren, and the east side of malioboro is the territory of danurejan kemantren. while the entire side of margo utomo street is the area of kemantren jetis, and the side of margo mulyo street is the area of kemantren gondomanan.
[10082] monumen ini dibangun oleh sri sultan hamengkubuwono i pada tahun 1755. dikenal sebagai tugu golong-gilig, dan dibangun dalam semangat persatuan rakyat. di puncak tugu berbentuk bulat golong dan tiangnya berbentuk silindris gilig, demikianlah namanya. ketinggian monumen tersebut adalah 25 meter. dibangun di garis imajiner yogyakarta yang menghubungkan laut selatan, keraton ngayogyakarta hadiningrat, dan gunung merapi. pada saat

```
bertapa, konon sultan yogyakarta saat itu menggunakan tugu ini
sebagai patokan untuk menghadap ke puncak gunung merapi.
[10043] pada masa lalu, perlintasan ini dapat dilintasi oleh
kendaraan umum sebagai penghubung jalan margo utomo menuju
malioboro. namun karena meningkatnya volume kendaraan yang
melintas, membuat perlintasan ini hanya boleh dilintasi oleh
kendaraan-kendaraan kecil seperti becak atau sepeda, sedangkan
kendaraan lain harus memutar terlebih dahulu ke arah timur
melewati jembatan kewek, kemudian berbelok ke arah barat melalui
jalan abu bakar ali, barulah sampai di jalan malioboro.
[20043] in the past, this crossing could be crossed by public
transportation to connect jalan margo utomo to malioboro.
however, due to the increasing volume of vehicles passing
through, this crossing can only be crossed by small vehicles such
as pedicabs or bicycles, while other vehicles must first turn
east through kewek bridge, then turn west through jalan abu bakar
ali, then arrive at malioboro street.
```

2.3. *Indexing* dan Pencarian Dokumen – FAISS's Inverted file with exact post-verification

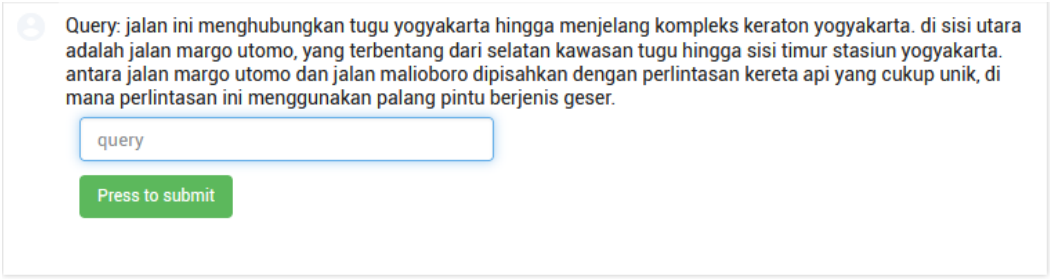
Untuk menggunakan *indexing* Faiss dengan *Inverted file with exact post-verification*, dimulai dengan menjalankan Cell 10, dengan terlebih dulu memasukkan jumlah **ncell** seperti ditunjukkan oleh Gambar 3.

Gambar 3. Panel untuk memasukkan jumlah ncell

Setelah dijalankan, akan ditampilkan luaran berupa informasi hasil *indexing:*

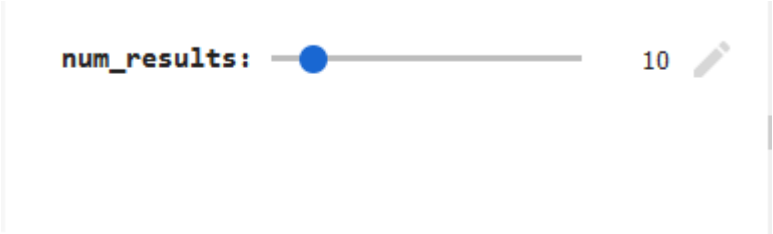
```
embeddings dimension :  512
```

Untuk mendeteksi kemiripan tekstual, pertama kali dimasukkan kueri berupa kalimat atau paragraf berisi teks yang akan diperiksa. Kueri dapat dalam Bahasa Indonesia atau Bahasa Inggris dimasukkan pada panel input pada Cell 11. Kueri akan ditampilkan pada panel tersebut, sebagaimana ditunjukkan oleh Gambar 4.



Gambar 4. Panel untuk masukan kueri

Jumlah hasil dokumen yang dikehendaki dimasukkan panel pada Cell 12 seperti ditunjukkan oleh Gambar 5, yaitu mengubah *slider* untuk variabel **num_results**.



Gambar 5. Panel untuk memasukkan jumlah hasil

Hasil deteksi dokumen akan ditampilkan di bawah Cell 12 dengan menampilkan jarak kemiripan serta id dokumen.

```
L2 distance: [2.917644966772548e-13, 0.2706255316734314,
0.5572736859321594, 0.6193249225616455, 0.6793750524520874,
0.7366632223129272, 0.8248614072799683, 0.8785932064056396,
0.9848678112030029, 1.0181164741516113]

 paper ID: [10042, 20042, 10041, 20041, 10081, 10045, 20045,
10082, 10043, 20043]
```

Jika Cell 13 dijalankan, akan menampilkan kalimat-kalimat dalam Bahasa Indonesia dan Inggris yang mirip yang ditemukan oleh kode pada Cell 12 secara urut berdasarkan hasil pencarian, sebagai contoh ditunjukkan di bawah ini.

```
this is index results [10042, 20042, 10041, 20041, 10081, 10045,
20045, 10082, 10043, 20043]
[10042] jalan ini menghubungkan tugu yogyakarta hingga menjelang
kompleks keraton yogyakarta. di sisi utara adalah jalan margo
utomo, yang terbentang dari selatan kawasan tugu hingga sisi
timur stasiun yogyakarta. antara jalan margo utomo dan jalan
malioboro dipisahkan dengan perlintasan kereta api yang cukup
unik, di mana perlintasan ini menggunakan palang pintu berjenis
geser.
[20042] this road connects tugu yogyakarta to the approach of the
yogyakarta palace complex. on the north side is jalan margo
utomo, which runs from the south of the tugu area to the east
side of yogyakarta station. jalan margo utomo and jalan malioboro
are separated by a unique railroad crossing, which uses a sliding
doorstop.
```

[10041] jalan malioboro jawa: ꦗꦭꦤ꧀ꦩꦭꦶꦪꦧꦺꦫꦺ, translit. dalan maliabara adalah nama salah satu kawasan jalan dari tiga jalan di kota yogyakarta yang membentang dari tugu yogyakarta hingga ke persimpangan titik nol kilometer yogyakarta. secara keseluruhan, kawasan malioboro terdiri atas jalan margo utomo, jalan malioboro, dan jalan margo mulyo. jalan ini merupakan poros garis imajiner kraton yogyakarta. itu terletak sumbu utara-selatan di garis antara kraton yogyakarta dan gunung merapi. ini sendiri penting bagi banyak penduduk lokal, orientasi utara-selatan antara istana dan gunung berapi menjadi penting.

[20041] malioboro street javanese: ꦗꦭꦤ꧀ꦩꦭꦶꦪꦧꦺꦫꦺ, translit. dalan maliabara is the name of one of the three streets in yogyakarta city that stretches from tugu yogyakarta to the intersection of yogyakartas zero kilometer point. overall, the malioboro area consists of jalan margo utomo, jalan malioboro, and jalan margo mulyo. this road is the axis of the yogyakarta kraton imaginary line. it lies north-south axis on the line between kraton yogyakarta and mount merapi. this in itself is important for many locals, the north-south orientation between the palace and the volcano being important.

[10081] tugu yogyakarta jawa: ꦠꦸꦒꦸꦔꦪꦺꦴꦒꦾꦏꦂꦠ, translit. tugu ngayogyåkartå adalah sebuah tugu atau monumen yang sering dipakai sebagai simbol atau lambang dari kota yogyakarta. tugu yang terletak di perempatan jalan jenderal sudirman dan jalan margo utomo ini, mempunyai nilai simbolis yang merupakan garis yang bersifat magis yang menghubungkan pantai parangtritis dan panggung krapyak di kabupaten bantul, keraton yogyakarta di kota yogyakarta dan gunung merapi di kabupaten sleman.
[10045] jalan malioboro sebenarnya hanya terbentang dari sisi selatan rel kereta api, di depan hotel grand inna hingga berakhir di pasar beringharjo sisi timur. dari titik ini, nama jalan berubah menjadi jalan margo mulyo hingga titik nol kilometer yogyakarta. jalan malioboro menjadi batas antara kemantren gedongtengen dan kemantren danurejan, di mana sisi barat malioboro adalah wilayah dari kemantren gedongtengen, dan sisi timur malioboro adalah wilayah dari kemantren danurejan. sedangkan seluruh sisi jalan margo utomo adalah wilayah dari kemantren jetis, dan sisi jalan margo mulyo adalah wilayah dari kemantren gondomanan.
[20045] malioboro street actually only stretches from the south side of the railroad tracks, in front of the grand inna hotel to end at pasar beringharjo on the east side. from this point, the street name changes to jalan margo mulyo until yogyakartas zero kilometer point. malioboro street becomes the boundary between gedongtengen kemantren and danurejan kemantren, where the west side of malioboro is the territory of gedongtengen kemantren, and the east side of malioboro is the territory of danurejan kemantren. while the entire side of margo utomo street is the area of kemantren jetis, and the side of margo mulyo street is the area of kemantren gondomanan.
[10082] monumen ini dibangun oleh sri sultan hamengkubuwono i pada tahun 1755. dikenal sebagai tugu golong-gilig, dan dibangun dalam semangat persatuan rakyat. di puncak tugu berbentuk bulat golong dan tiangnya berbentuk silindris gilig, demikianlah namanya. ketinggian monumen tersebut adalah 25 meter. dibangun di garis imajiner yogyakarta yang menghubungkan laut selatan, keraton ngayogyakarta hadiningrat, dan gunung merapi. pada saat

bertapa, konon sultan yogyakarta saat itu menggunakan tugu ini sebagai patokan untuk menghadap ke puncak gunung merapi.
[10043] pada masa lalu, perlintasan ini dapat dilintasi oleh kendaraan umum sebagai penghubung jalan margo utomo menuju malioboro. namun karena meningkatnya volume kendaraan yang melintas, membuat perlintasan ini hanya boleh dilintasi oleh kendaraan-kendaraan kecil seperti becak atau sepeda, sedangkan kendaraan lain harus memutar terlebih dahulu ke arah timur melewati jembatan kewek, kemudian berbelok ke arah barat melalui jalan abu bakar ali, barulah sampai di jalan malioboro.
[20043] in the past, this crossing could be crossed by public transportation to connect jalan margo utomo to malioboro. however, due to the increasing volume of vehicles passing through, this crossing can only be crossed by small vehicles such as pedicabs or bicycles, while other vehicles must first turn east through kewek bridge, then turn west through jalan abu bakar ali, then arrive at malioboro street.

3. Kode Sumber (*Source Code*) Program

INISIALISASI

Cell 1

```
!pip install "tensorflow-text==2.11.*"
!pip install bokeh
!pip install tqdm
!pip install faiss-cpu

import bokeh
import numpy as np
import os
import pandas as pd
import tensorflow.compat.v2 as tf
import tensorflow_hub as hub
from tensorflow_text import SentencepieceTokenizer
import sklearn.metrics.pairwise
```

Cell 2

```
from tqdm import tqdm
from tqdm import trange
from google.colab import drive

drive.mount('/content/drive')

module_url = 'https://tfhub.dev/google/universal-sentence-encoder-
multilingual/3'

embed = hub.load(module_url)

corpus_metadata = [('id', 'id_pars01.csv', 'Indonesian'), ('en',
'en_pars01.csv', 'English')]
lg2senteces_tmp = {}
language_to_sentences = {}
language_to_news_path = {}
texts={}
parID = {}
line_ele =[]
folder_path="/content/drive/MyDrive/Semester 7/ProyekEvaluasi/"
for language_code, csv_file, language_name in corpus_metadata:
  text_path = os.path.join(os.path.dirname(folder_path), csv_file)
  lg2senteces_tmp[language_code] = pd.read_csv(text_path, sep='\t',
header=0, encoding_errors="ignore" )
  #language_to_news_path[language_code] = text_path
```

```
  language_to_sentences[language_code] =
lg2senteces_tmp[language_code][1:]

  print('{:,} {}
paragraphs'.format(len(language_to_sentences[language_code]),
language_name))
  for kee, val in language_to_sentences.items():
    texts[language_code] = val['paragraphs']
    parID[language_code] = val['DocParID']
  print('{} {} id '.format('ini isi parID ', texts[language_code] ))
print('lg2sentences', lg2senteces_tmp)
```

Cell 3

```
folder_path_output="/content/drive/MyDrive/Semester
7/ProyekEvaluasi/Output/"
folder_output_l2 = os.path.join(folder_path_output, "FlatL2") + "/"
folder_output_ivf = os.path.join(folder_path_output, "IVF") + "/"
```

Cell 4

```
language_to_embeddings = {}
for language_code, csv_file, language_name in corpus_metadata:
  print('\nComputing {} embeddings'.format(language_name))
  with tqdm(total=len(language_to_sentences[language_code])) as pbar:
    language_to_embeddings[language_code] = embed(texts[language_code])
    pbar.update(len(language_to_sentences[language_code]))
print('this is the language2embeddings: ', language_to_embeddings )
```

Cell 5

```
# Indexing with Faiss using Flat L2 Index

embedding_dimensions = len(list(language_to_embeddings.values())[0][0])
print('embeddings dimension : ', embedding_dimensions)

import faiss
for language_code, news_file, language_name in corpus_metadata:
  print('\nAdding {} embeddings to index'.format(language_name))
  language_to_embeddings[language_code] = np.array([embedding for
embedding in language_to_embeddings[language_code]]).astype("float32")
  index = faiss.IndexFlatL2(embedding_dimensions)
  index = faiss.IndexIDMap(index)
  index_i = faiss.IndexFlatL2(embedding_dimensions)
  index_i = faiss.IndexIDMap(index_i)
  index.train(language_to_embeddings[language_code])
  index.add_with_ids(language_to_embeddings['en'], parID['en'])
  index.add_with_ids(language_to_embeddings['id'], parID['id'])
print(index)
```

Cell 6

```
import panel as pn
pn.extension()

text = pn.widgets.TextInput(placeholder="query")
button = pn.widgets.Button(name='Press to submit',
button_type='success')
#button.on_click(lambda x: print("Button"))

@pn.depends(text, watch=True)
def query_processing(t):
    print("Query:", t)
    button.clicks += 1

pn.Column(text, button).servable()
```

Cell 7

```
query_id = '10042'
```

Cell 8

```
my_query = text.value_input
num_results = 10  #@param {type:"slider", min:0, max:100, step:1}
query_embedding = embed(my_query)
D, I = index.search(query_embedding, num_results)
print(f'L2 distance: {D.flatten().tolist()}\n\n paper ID:
{I.flatten().tolist()}')
```

Cell 9

```
results = I.flatten().tolist()
print('this is index results', results)
def id2text(lg2senteces_tmp, I):
  out_text=[]
  for kee, val in lg2senteces_tmp.items():
    for j, idx in enumerate(results):
      for i in range(len(val['DocParID'])):
        if idx == val['DocParID'][i]:
          annotated_output = '[{}] {}'.format(str(idx),
val['paragraphs'][i])
          out_text.insert(j, annotated_output)
  return out_text

sentences = id2text(lg2senteces_tmp, I)
list_output=[]
for sent in sentences:
  print(sent)
  list_output.append(sent)
```

```
#output = '\n'.join(list_output)
with open(os.path.join(os.path.dirname(folder_output_l2),
f'{query_id}.txt'), 'w') as fileout:
  fileout.write('\n'.join(list_output))
```

Cell 10

```
# indexing with FAISS's Inverted file with exact post-verification

embedding_dimensions = len(list(language_to_embeddings.values())[0][0])
print('embeddings dimension : ', embedding_dimensions)

import faiss
ncells = 5 #@param {type:"slider", min:0, max:30, step:1}
for language_code, news_file, language_name in corpus_metadata:
  language_to_embeddings[language_code] = np.array([embedding for
embedding in language_to_embeddings[language_code]]).astype("float32")
  quantizer = faiss.IndexFlatIP(embedding_dimensions)  # how the
vectors will be stored/compared
  index = faiss.IndexIVFFlat(quantizer, embedding_dimensions, ncells)
  index.train(language_to_embeddings[language_code])  # we must train
the index to cluster into cells
  index.add_with_ids(language_to_embeddings['en'], parID['en'])
  index.add_with_ids(language_to_embeddings['id'], parID['id'])
```

Cell 11

```
import panel as pn
pn.extension()

text = pn.widgets.TextInput(placeholder="query")
button = pn.widgets.Button(name='Press to submit',
button_type='success')
#button.on_click(lambda x: print("Button"))

@pn.depends(text, watch=True)
def query_processing(t):
    print("Query:", t)
    button.clicks += 1

pn.Column(text, button).servable()
```

Cell 12

```
my_query = text.value_input
num_results = 10  #@param {type:"slider", min:0, max:100, step:1}
query_embedding = embed(my_query)
D, I = index.search(query_embedding, num_results)
```

```
print(f'L2 distance: {D.flatten().tolist()}\n\n paper ID:
{I.flatten().tolist()}')
```

Cell 13

```
results = I.flatten().tolist()
print('this is index results', results)
def id2text(lg2senteces_tmp, I):
  out_text=[]
  for kee, val in lg2senteces_tmp.items():
    for j, idx in enumerate(results):
      for i in range(len(val['DocParID'])):
        if idx == val['DocParID'][i]:
          annotated_output = '[{}] {}'.format(str(idx),
val['paragraphs'][i])
          out_text.insert(j, annotated_output)
  return out_text

sentences = id2text(lg2senteces_tmp, I)
list_output=[]
for sent in sentences:
  print(sent)
  list_output.append(sent)
#output = '\n'.join(list_output)
with open(os.path.join(os.path.dirname(folder_output_ivf),
f'{query_id}.txt'), 'w') as fileout:
  fileout.write('\n'.join(list_output))
```

EVALUASI

Cell 1

```
!pip install -U scikit-learn
```

Cell 2

```
from sklearn.metrics import confusion_matrix,
precision_recall_fscore_support
```

Cell 3

```
precisions = []
recalls = []
f1 = []
```

Cell 4

```
tid_path = os.path.join(folder_path, 'tid_pars.txt')
with open(tid_path, 'r', encoding="utf8") as fh:
```

```
  tid = fh.read()
tid_par = tid.split('\n')
tid_par =[par.split('\t') for par in tid_par]
df_tid_par = pd.DataFrame(tid_par, columns =['par', 'id', 'en'])
df_tid_par.head()
```

Mengambil dokumen relevan dari korpus

Cell 5

```
filtered_df = df_tid_par.loc[df_tid_par['par'].str.contains('penyakit
yang sedang berlangsung di seluruh dunia adalah pandemi covid-19.
sindrom pernapasan akut berat 2sars-cov-2 merupakan penyebab
penyakitnya. kasus positif covid-19 pertama di indonesia terdeteksi
pada  2 maret 2020, saat dua orang terkonfirmasi tertular dari seorang
warga negara jepang.')]
filtered_df
```

Cell 6

```
relevant = filtered_df.iloc[:, 1:]
relevant = relevant.to_numpy()
relevant
```

Mengambil id dokumen ter-*retrieve*

Cell 7

```
import re
out_path = folder_output_l2 + '10011.txt'
with open(out_path, 'r') as fh:
  txt = fh.read()
retrieved_id = re.findall('\[\d{5}\]', txt)
retrieved_id = [id.replace('[','').replace(']','') for id in
retrieved_id]
retrieved_id = np.array(retrieved_id)
retrieved_id
```

Cell 8

```
relevant_item_retrieved = np.intersect1d(retrieved_id, relevant)
relevant_item_retrieved
```

Hitung Precision dan Recall

Cell 9

```
precision = len(relevant_item_retrieved) / len(retrieved_id)
precision
```

Cell 10

```
recall = len(relevant_item_retrieved) / len(relevant)
recall
```