222112058_Feza Raffa Arnanda_Penugasan Praktikum 2

August 31, 2023

0.0.1 Penugasan Praktikum Pertemuan 2 Information Retrieval

Feza Raffa Arnanda - 3SD2 - 222112058

1. Diketahui suatu dokumen berikut terdiri dari beberapa paragraf dan setiap paragraf terdiri dari beberapa kalimat. Paragraf yang berbeda dipisahkan dengan Enter, sedangkan kalimat dipisahkan dengan titik, tanda tanya, atau tanda seru. Buat kode fungsi python untuk memisahkan dokumen sehingga menghasilkan variabel list_paragraf (nama fungsi: paragraph_parsing), dan masing-masing paragraf menjadi variabel list_kalimat (nama fungsi: sentence_parsing)

Paragraph parsing function

[ ]:
```python
def paragraph_parser(text):
    paragraf = text.split('\n')

    parsed_paragraf = []

    for indeks, paragraf in enumerate(paragraf):
        label = f"p{indeks+1}"
        parsed_paragraf.append(f"{label}:{paragraf}") # untuk pemberian label␣
↪"p1" "p2" dll. f itu f-string di python untuk include expression di dalam␣
↪string.

    return "\n\n".join(parsed_paragraf)
```

Sentence parsing function. Menggunakan library re (regular expression) yang akan memudahkan dalam pemberian pola

[ ]:
```python
import re

def sentence_parser(paragraph):
    sentences = re.split(r'[.!?]', paragraph)  #re.split untuk split dari suatu␣
↪teks berdasarkan expression.

    parsed_output = []
    for index, sentence in enumerate(sentences):
        if sentence.strip():
            label = f"s{index + 1}"
            parsed_output.append(f"{label} : {sentence.strip()}")
```

```
    return parsed_output
```

Function test

[ ]:
```
teks = """
Mobilitas warga bakal diperketat melalui penerapan PPKM level 3 se-Indonesia di␣
 ↪masa libur Natal dan tahun baru (Nataru). Rencana kebijakan itu dikritik␣
 ↪oleh Epidemiolog dari Griffith University Dicky Budiman.
Dicky menyebut pembatasan mobilitas memang akan memiliki dampak dalam mencegah␣
 ↪penularan COVID-19. Tapi, kata dia, dampaknya signifikan atau tidak akan␣
 ↪bergantung pada konsistensi yang mendasar yakni testing, tracing, treatment,␣
 ↪(3T) hingga vaksinasi COVID-19.
"""

# Paragraf parser
paragraphs = paragraph_parser(teks.strip())
print("List paragraf : \n")
print(paragraphs, "\n")


# Kalimat parser
paragraphs = teks.strip().split('\n')  # paragraf parser

for index, paragraph in enumerate(paragraphs):
    sentence_list = sentence_parser(paragraph)
    print(f"List kalimat pada paragraf {index+ 1} :\n")
    print('\n'.join(sentence_list))
    print()
```

```
List paragraf :

p1:Mobilitas warga bakal diperketat melalui penerapan PPKM level 3 se-Indonesia
di masa libur Natal dan tahun baru (Nataru). Rencana kebijakan itu dikritik oleh
Epidemiolog dari Griffith University Dicky Budiman.

p2:Dicky menyebut pembatasan mobilitas memang akan memiliki dampak dalam
mencegah penularan COVID-19. Tapi, kata dia, dampaknya signifikan atau tidak
akan bergantung pada konsistensi yang mendasar yakni testing, tracing,
treatment, (3T) hingga vaksinasi COVID-19.

List kalimat pada paragraf 1 :

s1 : Mobilitas warga bakal diperketat melalui penerapan PPKM level 3 se-
Indonesia di masa libur Natal dan tahun baru (Nataru)
s2 : Rencana kebijakan itu dikritik oleh Epidemiolog dari Griffith University
Dicky Budiman
```

```
List kalimat pada paragraf 2 :

s1 : Dicky menyebut pembatasan mobilitas memang akan memiliki dampak dalam
mencegah penularan COVID-19
s2 : Tapi, kata dia, dampaknya signifikan atau tidak akan bergantung pada
konsistensi yang mendasar yakni testing, tracing, treatment, (3T) hingga
vaksinasi COVID-19
```

2. Lakukan case-folding (upper case dan lower case), tokenisasi, eliminasi stopword dan stemming pada dokumen di folder "berita" menggunakan library yang sudah tersedia (nltk, spacy, sastrawi, etc).

NLTK

Pada library NLTK, tersedia versi bahasa indonesia untuk melakukan text processing (case folding, tokenisasi, stemming, dll)

[ ]:
```python
import os
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer


# Path folder berita
path = "C:/Users/FEZA/My Drive/00. Drive PC/1.STIS/5. Semester 5/Information␣
 ↪Retrieval [IR] P/Pertemuan 1/berita"

# Stopwords pada library NLTK

stop_words = set(stopwords.words('indonesian'))

# Stemmer pada library NLTK
stemmer = PorterStemmer()

# Iterasi
for file in os.listdir(path):
    if os.path.isfile(os.path.join(path, file)):
        with open(os.path.join(path, file), 'r', encoding='utf-8') as f:
            content = f.read().lower()  # Casefolding pada NLTK

            # Tokenisasi
            words = word_tokenize(content)

            # menghilangkan stopwords sekaligus stemming
            # memeriksa apakah kata saat ini adalah alfanumerik (terdiri dari␣
 ↪huruf dan/atau angka) dan apakah
            # kata tersebut bukan termasuk dalam daftar stopwords.
```

```
            filtered_words = [stemmer.stem(word) for word in words if word.
 ↪isalnum() and word not in stop_words]


            # print hasil akhir
            print(filtered_words)
```

```
['terinfeksi', 'viru', 'corona', 'melonjak', 'negara', 'pemerintah', 'kera',
'mengatasi', 'penyebaran', 'viru']
['mencuci', 'tangan', 'rutin', 'mencegah', 'penularan', 'penyakit',
'penelitian', 'mencuci', 'tangan', 'mengurangi', 'risiko', 'infeksi']
['pandemi', 'corona', 'mengubah', 'aspek', 'kehidupan', 'mencari', 'solusi',
'mengatasi', 'negatifnya']
['hasil', 'survei', 'tingkat', 'kepuasan', 'masyarakat', 'layanan', 'kesehatan',
'menurun', 'pandemi', 'perbaikan', 'diambil']
['pemerintah', 'mengumumkan', 'kebijakan', 'terkait', 'pembatasan', 'sosial',
'mengendalikan', 'penyebaran', 'viru', 'warga', 'diharapkan', 'mematuhi',
'aturan']
['file']
```

Sastrawi

Pada library Sastrawi tersedia juga versi bahasa Indonesia sehingga lebih mudah untuk text processing

[ ]:
```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from Sastrawi.StopWordRemover.StopWordRemoverFactory import␣
 ↪StopWordRemoverFactory
from nltk.tokenize import word_tokenize

# path ke folder berita
path = "C:/Users/FEZA/My Drive/00. Drive PC/1.STIS/5. Semester 5/Information␣
 ↪Retrieval [IR] P/Pertemuan 1/berita"

# methode sastrawi stemmer dan stopword remover
stemmer_factory = StemmerFactory()
stemmer = stemmer_factory.create_stemmer()

stopword_factory = StopWordRemoverFactory()
stopword_remover = stopword_factory.create_stop_word_remover()

# Iterai ke semua file di folder berita
for file in os.listdir(path):
    if os.path.isfile(os.path.join(path, file)):
        with open(os.path.join(path, file), 'r', encoding='utf-8') as f:
            content = f.read().lower()  # case folding

            # Tokenization
```

```
            words = word_tokenize(content)

            # mennghilangkan stopword dan dilakukan stemming sekaligus
            filtered_words = [stemmer.stem(stopword_remover.remove(word)) for␣
 ↪word in words]

            # Print hasil akhir
            print(filtered_words)
```

```
['kasus', 'baru', 'infeksi', 'virus', 'corona', 'lonjak', '', 'beberapa',
'negara', '', 'perintah', 'sedang', 'kerja', 'keras', '', 'atas', 'sebar',
'virus', '', '']
['penting', 'cuci', 'tangan', '', 'rutin', '', 'cegah', 'tular', 'sakit', '',
'teliti', 'tunjuk', '', 'cuci', 'tangan', '', 'kurang', 'risiko', 'infeksi', '']
['pandemi', 'corona', '', 'ubah', 'banyak', 'aspek', 'hidup', '', '', '',
'perlu', 'sama', 'cari', 'solusi', '', 'atas', 'dampak', 'negatif']
['hasil', 'survei', 'tunjuk', '', 'tingkat', 'puas', 'masyarakat', '', 'layan',
'sehat', '', 'turun', 'lama', 'pandemi', '', 'langkah', 'baik', 'perlu',
'segera', 'ambil']
['perintah', 'umum', 'bijak', 'baru', 'kait', 'batas', 'sosial', '', 'kendali',
'sebar', 'virus', '', 'semua', 'warga', 'harap', 'patuh', 'atur', 'sebut', '']
['', 'shellclassinfo', '', 'iconresource c', '', 'program', 'files google
drive', 'file', 'stream 79 0 2 0 googledrivefs exe 23']
```

SpaCy

Pada library SpaCy tidak tersedia versi bahasa Indonesia, sehingga disini saya menggunakan versi bahasa inggris dengan folder berita yang sudah berisi file berbahasa inggris juga, sehingga penggunaan library Spacy bisa kita gunakan secara maksimal dan terlihat bagaimana proses text processing yang baik

[ ]:
```
import spacy

# Load spaCy model untuk bahasa inggris.

nlp = spacy.load("en_core_web_sm")

# path ke berita folder berbahasa inggris
path = "C:/Users/FEZA/My Drive/00. Drive PC/1.STIS/5. Semester 5/Information␣
 ↪Retrieval [IR] P/Pertemuan 1/berita/english"

# Iterasi
for file in os.listdir(path):
    if os.path.isfile(os.path.join(path, file)):
        with open(os.path.join(path, file), 'r', encoding='utf-8') as f:
            content = f.read().lower()  # Casefolding

            # Proses NLP pada library Spacy
```

```
            doc = nlp(content)

            # Lemmatize (Stemming) dan penghilangan stopwords
            processed_words = [token.lemma_ for token in doc if token.is_alpha␣
↪and not token.is_stop]

            # Print output
            print(processed_words)
```

```
['iconresource', 'file']
['powerful', 'earthquake', 'strike', 'indonesia', 'thursday', 'kill', 'people',
'injure', 'hundred', 'earthquake', 'magnitude', 'strike', 'island', 'sumatra',
'local', 'time', 'epicenter', 'locate', 'mile', 'city', 'padang']
['nasa', 'thursday', 'unveil', 'new', 'space', 'telescope', 'james', 'webb',
'space', 'telescope', 'telescope', 'powerful', 'build', 'design', 'study',
'universe', 'unprecedented', 'detail']
['new', 'study', 'publish', 'thursday', 'find', 'climate', 'change', 'worsen',
'study', 'conduct', 'team', 'scientist', 'university', 'oxford', 'find',
'earth', 'climate', 'warm', 'alarming', 'rate']
```