Nama : Raihan Rahmanda Junianto

NIM : 222112303

Kelas : 3SD2

**Responsi Praktikum 2 Information Retrieval**

Praktikum kali ini membahas mengenai term-vocabulary serta macam-macam teks processing, seperti tokenisasi, capitalization/case-folding, eliminasi stopword, normalisasi, dan stemming. Pada praktikum ini juga dibahas mengenai library-library yang digunakan untuk melakukan pemrosesan tersebut. Berikut disajikan jawaban dari masing-masing nomor penugasan yang terdapat pada modul 2 praktikum Information Retrieval.

1. **Buat kode fungsi python untuk memisahkan dokumen sehingga menghasilkan variabel list\_paragraf (nama fungsi: paragraph\_parsing), dan masing-masing paragraph menjadi variabel list\_kalimat (nama fungsi: sentence\_parsing).**

Langkah pertama yang dilakukan untuk menjawab soal di atas adalah mendefinisikan dokumen/teks yang dimaksud. Pada kasus kali ini, setiap paragraf dalam dokumen tersebut dipisahkan oleh enter sebanyak dua kali. Hal tersebut bertujuan untuk memudahkan penulisan regular expressionnya karena paragraf yang didefinisikan tidak berada dalam satu baris lurus.

A computer screen shot of a computer screen

Description automatically generated

Proses selanjutnya adalah mendefinisikan fungsi yang akan digunakan pada program, yaitu fungsi paragraph\_parsing dan sentence\_parsing.

A screen shot of a computer code

Description automatically generated

Pada fungsi paragraph\_parsing di atas, setiap paragraf dalam dokumen akan dipisahkan (split) berdasarkan enter sebanyak dua kali yang dituliskan dengan regular expression “\n\n”. Sedangkan, pada fungsi sentence\_parsing, setiap kalimat dalam paragraf akan dipisahkan berdasarkan tanda titik.

Selanjutnya, proses akan berlanjut ke dalam pemanggilan kedua fungsi tersebut dan mencetak output sesuai dengan yang tertera pada modul.

A screen shot of a computer code

Description automatically generated

Sintaks di atas akan menampilkan output berupa masing-masing paragraf yang terkandung dalam dokumen serta kalimat-kalimat pada masing-masing paragraf tersebut. Berikut disajikan tampilan program secara keseluruhan beserta outputnya.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Lakukan case-folding, tokenisasi, eliminasi stopword dan stemming pada dokumen di folder “berita” menggunakan library yang sudah tersedia (nltk, spacy, sastrawi, etc).**

* **Mendefinisikan library/modul yang digunakan**

**A screen shot of a computer

Description automatically generated**

Pada program kali ini, modul yang digunakan adalah modulo os, spacy, dan sastrawi. **Import os** berfungsi untuk mengimpor modul os yang digunakan untuk berinteraksi dengan sistem operasi, seperti mengakses file dan direktori. Sedangkan, **from spacy.lang.id import Indonesia** berfungsi untuk mengimpor modul Indonesian dari paket spacy.lang.id. Modul ini menyediakan dukungan untuk pemrosesan bahasa Indonesia menggunakan library spaCy. Selanjutnya, **import spacy** digunakan mengimpor modul spacy yang merupakan library utama untuk pemrosesan bahasa alami. Lalu, **from spacy.lang.id.stop\_words import STOP\_WORDS** digunakan untuk mengimpor variabel STOP\_WORDS dari modul stop\_words yang terdapat dalam paket spacy.lang.id. Variabel ini berisi kumpulan kata-kata yang umumnya dianggap sebagai kata penghubung atau kata yang tidak memiliki makna penting dalam bahasa Indonesia. Dan yang terakhir, **from Sastrawi.Stemmer.StemmerFactory import StemmerFactory** berfungsi untuk mengimpor kelas StemmerFactory dari modul StemmerFactory yang terdapat dalam paket Sastrawi. Kelas ini digunakan untuk membuat objek stemmer yang digunakan untuk melakukan stemming pada kata-kata dalam bahasa Indonesia.

* **Mendefiniskan fungsi dan lokasi folder berita**

A screenshot of a computer

Description automatically generated

* **Menampilkan seluruh list file yang terdapat pada folder berita dan menampilkan isi dari masing-masing file tersebut.**

**A computer screen shot of text

Description automatically generated**

* **Case Folding masing-masing isi dari folder berita**

**A screen shot of a computer program

Description automatically generated**

* **Tokenisasi masing-masing isi dari folder berita**

**A screen shot of a computer program

Description automatically generated**

* **Stopword elimination masing-masing isi dari folder berita**

**A computer screen shot of colorful text

Description automatically generated**

* **Stemming masing-masing isi dari folder berita**

**A screen shot of a computer code

Description automatically generated**

* **Sintaks program secara keseluruhan serta outputnya**

**A screen shot of a computer screen

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**A black screen with white text

Description automatically generated**

**A black screen with white text

Description automatically generated**

**A black screen with white text

Description automatically generated**

**A black screen with white text

Description automatically generated**

**A black screen with white text

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**A black and white image of text

Description automatically generated**

**A black and white screen with white text

Description automatically generated**

**A black screen with white text

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**A black screen with white text

Description automatically generated**