CSCE604135 • Temu-Balik Informasi
Semester Genap 2024/2025
Fakultas Ilmu Komputer, Universitas Indonesia

Tugas Pemrograman 1: *Text Preprocessing*

Tenggat Waktu: Senin, 3 Maret 2025, 23.55 WIB

Ketentuan:

1. Anda diberikan beberapa *file* notebook yang berisi *template* kode.
2. Kumpulkan **semua program** (*.ipynb*) dengan penamaan **TP1_NPM.zip** melalui submisi SCeLe.

   Contoh penamaan file: TP1_2006524290.zip
3. Kumpulkan *file zip* tersebut pada submisi yang telah disediakan di SCeLe sebelum **Senin, 3 Maret 2025, 23.55 WIB**. Keterlambatan pengumpulan akan dikenakan penalti.
4. Tugas ini dirancang sebagai tugas mandiri. **Plagiarisme tidak diperkenankan dalam bentuk apapun**. Adapun kolaborasi berupa diskusi (tanpa menyalin maupun mengambil jawaban orang lain) dan memanfaatkan informasi dari literatur manapun masih diperbolehkan. **Pastikan** untuk mencantumkan nama kolaborator dan referensi literatur.
5. Anda boleh berkonsultasi terkait tugas ini asisten dosen berikut. Asisten dosen diperbolehkan membantu Anda dengan memberikan petunjuk.
   a. Jaycent Gunawan Ongris
      Email: jaycent.gunawan@ui.ac.id
      Discord: jaycentg
   b. Muhammad Ilham Ghozali
      Email: walangsigit@gmail.com
      Discord: myticalcat

Petunjuk Pengerjaan Tugas

Tugas ini akan melatih Anda dalam melakukan text preprocessing pada Bahasa Indonesia dan Inggris, serta mengimplementasikan data-driven tokenizer seperti Byte Pair Encoding (BPE) dan WordPiece.

Conventional Text Processing

Kita akan memulai dengan mempelajari cara melakukan text preprocessing pada teks Bahasa Inggris menggunakan dataset MSMARCO. Proses ini meliputi tokenization, stemming atau lemmatization, dan stopwords removal. Anda akan diberikan contoh kode untuk setiap tahapan, dan Anda diminta untuk menjawab pertanyaan terkait pipeline preprocessing yang diberikan, serta memperbaikinya berdasarkan insight yang Anda dapatkan.

Kemudian Anda akan membangun stopwords set Anda sendiri berdasarkan corpus yang diberikan. Stopwords dalam kasus ini adalah 200 token dengan frekuensi kemunculan tertinggi. Anda akan menggunakan hasil preprocessing dari dataset Bahasa Inggris yang telah dilakukan sebelumnya.

Selanjutnya, Anda akan mendefinisikan pipeline baru untuk melakukan text preprocessing pada teks Bahasa Indonesia menggunakan dataset `jakartaresearch/indonews`. Anda dapat menggunakan library seperti PySastrawi untuk melakukan preprocessing ini.

Kemudian kita akan menjelajahi berbagai alat pemrosesan teks untuk berbagai bahasa, yaitu Bahasa Inggris (NLTK), Bahasa Indonesia (Sastrawi), dan Bahasa Mandarin (Jieba). Anda akan membandingkan hasil stemming dan lemmatization di antara bahasa-bahasa ini.

Data-driven Tokenizer

Disini anda akan belajar mengimplementasikan beberapa data-driven tokenizer untuk bahasa indonesia.

Byte Pair Encoding (BPE)

Anda akan belajar mengimplementasikan salah satu data-driven tokenizer yang terkenal, yaitu BPE.

WordPiece Tokenizer

Terakhir, Anda akan mengimplementasikan WordPiece Tokenizer. Feel free untuk melakukan eksperimen dengan tokenizer ini karena sampai sekarang belum di-opensource oleh google.

Template Kode

- Template kode: notebook TP1

Deliverables

File jupyter notebook yang sudah dikerjakan

Catatan Revisi:

Rubrik Penilaian

Komponen	Proporsi
Implementasi dan jawaban soal bagian Pipeline	10%
Implementasi dan jawaban soal bagian Stopwords Set	5%
Implementasi dan jawaban soal bagian Indonesian Preprocessing Pipeline	10%
Implementasi dan jawaban soal bagian Stemming and Lemmatization Across Languages	5%
Implementasi Byte Pair Encoding (BPE) Tokenizer dan analisis	35%
Implementasi WordPiece Tokenizer dan analisis	35%

Selamat mengerjakan!