Tugas Pekan ke-4 POSTagging

Batas pengumpulan: Jumat 16 Oktober 2020, pukul 10.59 pagi, melalui LMS

Deskripsi

Buatlah POSTagger sederhana berbasis 3 metode seperti yang diberikan pada tutorial:

a. Metode baseline
b. Metode klasifikasi konvensional non-sekuensial
c. Metode HMM-Viterbi

Pembangunan model dilakukan berdasar **50 kalimat pertama** yang muncul pada file train01.tsv (terlampir) yang diambil dari https://github.com/kmkurn/id-pos-tagging . Kalimat uji yang digunakan adalah 10 kalimat berikutnya (**kalimat nomor 51-60**).

Modifikasi yang perlu Anda lakukan pada kode program (jika Anda menggunakan kode tutorial yang diberikan):

1. Pembacaan data latih dari file
2. Inisialisasi tagset pada POSTagger metode HMM-Viterbi
3. Pembentukan matriks emission dan transition probability pada POSTagger metode HMM-Viterbi
4. Pengukuran akurasi pada POSTagger metode baseline dan HMM-Viterbi

Catatan: TIDAK diperbolehkan menggunakan library/fungsi siap pakai untuk POStagger metode baseline dan metode HMM-Viterbi.

Informasi yang harus dituliskan pada laporan:

1. Hasil *tagging* setiap kalimat uji berdasar ketiga metode yang digunakan beserta analisisnya.
2. Perbandingan nilai akurasi pengujian dengan ketiga metode tersebut, metode mana yang paling tinggi akurasinya, beri analisis sederhana dugaan penyebabnya.

File yang harus dikumpulkan:

1. Program dan kelengkapannya: 3 file kode program python (.py) + 1 file data latih (.tsv) + 1 file data uji (.tsv) **dalam sebuah folder** + petunjuk menjalankan program (.txt).
2. Laporan: 1 file pdf, maksimum panjang laporan adalah 2 halaman.

Penilaian: 70% source code + 30% laporan

Detail penilaian:

a. Program:

- kebenaran implementasi pembacaan file data latih dan data uji dari sebuah direktori/folder (20 poin)
- kebenaran implementasi pengukuran akurasi pada POSTagger baseline (10 poin)
- kebenaran implementasi pengukuran akurasi pada POSTagger HMM-Viterbi (10 poin)

- kebenaran implementasi inisialisasi tagset pada POSTagger HMM-Viterbi (10 poin)

- kebenaran implementasi pembentukan matriks emission probability (10 poin)

- kebenaran implementasi pembentukan matriks transition probability (10 poin)

b. Laporan:

- kelengkapan (10 poin)

- analisis kesalahan tagging pada tiap kalimat uji (10 poin)

- analisis perbandingan akurasi POSTagging dengan 3 metode yang berbeda (10 poin)

Jika ada pertanyaan, silakan disampaikan melalui *channel* pekan_5_tugas_postagging di slack.