Tugas Pekan ke-2 Model Bahasa / *Language Model*

Batas pengumpulan: Jumat 25 September 2020, pukul 10.59 pagi, melalui LMS

Deskripsi

Buatlak sebuah model Bahasa **Bigram** dari 20 artikel berbahasa Indonesia. Setiap mahasiswa harus memilih topik yang berbeda, contoh: politik, ekonomi, kesehatan, dan lain-lain. Usahakan untuk memilih artikel-artikel dengan topik yang spesifik, misal: PSBB Jakarta Rem Darurat.

Setelah Anda membuat model Bahasa Bigram, lakukan evaluasi dengan menggunakan perplexity, dan analisis hasil yang diperoleh. Tuliskan hasil analisis dengan detail, dari rancangan kalimat uji yang mengandung:

a. Kalimat dengan topik yang serupa dengan data latih, paling sedikit tiga kalimat
b. Kalimat dengan topik yang berbeda dengan data latih, paling sedikit tiga kalimat

Persyaratan program model Bahasa/*language model*:

1. Dapat dibuat berdasarkan tutorial yang telah diberikan di kelas, atau dibangun dari awal sendiri.
2. Fungsi tamabahan yang harus diimplementasikan:
    a. Laplace smoothing
    b. Perhitungan nilai perplexity
3. Petunjuk bagaimana menjalankan program harus disertakan.
4. Program harus mengandung komentar cukup, sehingga mudah dipahami bagi pemeriksa.

Informasi yang harus dituliskan pada laporan:

1. Sepuluh unigram yang paling sering muncul.
2. Sepuluh bigram dengan pronability paling tinggi.
3. Keterangan pemilihan kalimat uji
4. Analisis perplexity, perbandingan antara hasil yang diperoleh dari kalimat dengan topik yang mirip dan yang tidak mirip.

File yang harus dikumpulkan:

1. Program dan kelengkapannya: 1 file kode program python (.py) + 20 file artikel (.txt) + petunjuk menjalankan program (.txt).
2. Laporan: 1 file pdf, maksimum Panjang laporan adalah 2 halaman.

Penilaian: 60% source code + 40% laporan, + maksimum 20 poin BONUS tambahan untuk implementasi model trigram.

Detail penilaian:

a. Program:

- kebenaran implementasi pembacaan file masukan dan pembangunan model Bahasa bigram (40 poin)

- kebenaran implementasi *Laplace smoothing* (10 poin)

- kebenaran implementasi perplexity (10 poin)

b. Laporan:

- kelengkapan (30 poin)

- analisis (10 poin)

c. Bonus:

kebenaran implementasi trigram dan analisis tambahan (20 poin)

Media diskusi

Semua mahasiswa yang mengambil MK Pemrosesan Bahasa Alami WAJIB bergabung pada forum diskusi di slack, link: https://join.slack.com/t/pemrosesanbah-c5k6846/shared_invite/zt-h6eo4p4n-AzRvTdtmV7pOStA~3xMTeA

Jika ada pertanyaan, silakan disampaikan melalui *channel* yang sesuai.