Deskripsi Tugas I Pemrosesan Bahasa Alami

Topik Terkait: *Language Modeling*

Judul Tugas: Prediksi Kemunculan Kata Menggunakan Model Bigram

Batas Pengumpulan: 10 Agustus 2018 (23:59)

Deskripsi:

Buatlah model Bigram berdasar kumpulan artikel Bahasa Indonesia (jumlah artikel yang disarankan kurang lebih 100 artikel, sumber bebas, topik juga bebas), dan gunakan model tersebut untuk memprediksi kata apa yang paling tepat muncul jika diberikan sebuah kata.

Misal kata yang diberikan sebagai masukan adalah “saya”, maka program akan memberikan keluaran berupa kata “akan”, karena kata “akan” mempunyai probabilitas muncul paling tinggi setelah kata “saya”.

Definisikan 10 kata yang digunakan pada pengujian, dan jelaskan alasan pemilihan kata-kata tersebut.

Program harus mengandung fungsi untuk:

1. Pembangunan model bigram dari kumpulan artikel
2. Pengujian prediksi kemunculan kata
3. Evaluasi model dengan perplexity. Silakan menyiapkan 5 kalimat pengujian untuk evaluasi dengan perplexity.

Deliverables:

1. Softcopy program beserta komentar yang mudah dimengerti
2. Softcopy artikel yang digunakan untuk membangun model
3. Petunjuk cara eksekusi program
4. Laporan singkat yang berisi:
    a. Keterangan sumber artikel, topik, dan alasan pemilihan.
    b. Analisis terhadap hasil pengujian prediksi kemunculan kata.
    c. Analisis terhadap hasil evaluasi perplexity.

Komponen penilaian:

1. Kebenaran, kelengkapan dan kejelasan program (nilai maksimum: 70)
2. Laporan (nilai maksimum: 30)
3. Bonus: penerapan smoothing atau add-one, atau metode lain untuk mengatasi persoalan frekuensi atau probabilitas yang melibatkan angka 0 (nilai maksimum: 20)