

Tugas Pekan ke-9 Word Sense Disambiguation

Batas pengumpulan: Jumat 20 November 2020, pukul 10.59 pagi, melalui LMS

Deskripsi

Lakukan implementasi algoritma Lesk untuk menentukan *sense* yang tepat pada sebuah kata Bahasa Indonesia. Data definisi *sense* dan contoh penggunaan dalam sebuah konteks diambil dari Kamus Besar Bahasa Indonesia daring, <https://kbbi.kemdikbud.go.id>.

Perhatikan contoh berikut

Informasi kata “bisa” diambil dari Kamus Besar Bahasa Indonesia:

The screenshot shows the KBBI online interface for the word "bisa". It lists three senses:

- bi.sa¹ Sense 1**: *v* mampu (kuasa melakukan sesuatu); dapat: *ia -- membaca, tetapi tidak -- menulis*
- bi.sa² Sense 2**: *n* zat racun yang dapat menyebabkan luka, busuk, atau mati bagi sesuatu yang hidup (biasanya terdapat pada binatang)
- bi.sa³ Sense 3**: *n* upacara adat pingitan

Diberikan sebuah kalimat uji, misal : “Adik sudah bisa membaca”, program yang Anda buat akan mengeluarkan *sense* mana yang paling tepat, berdasarkan algoritma Lesk.

Program dan Analisis yang Harus Dibuat

1. Buatlah program untuk menghitung skor kemiripan/similarity antara 2 buah teks dengan algoritma Lesk, yaitu berdasarkan jumlah kata yang beririsan pada 2 buah teks tersebut.
2. Pilihlah **tiga** contoh kata ambigu, kemudian buat contoh kalimat yang mengandung kata tersebut. Jadi Anda akan memiliki 6 (3 x 2 kalimat), karena masing-masing *sense* kata diwakili oleh 1 kalimat. Tentukan *sense* mana yang paling tepat untuk kata pada tiap kalimat, berdasarkan algoritma Lesk. Perbandingan dengan algoritma Lesk dilakukan berdasar definisi dan contoh kalimat dari Kamus Besar Bahasa Indonesia daring, <https://kbbi.kemdikbud.go.id>. Pendefinisian *sense*, definisi, dan contoh konteks dapat dilakukan secara *hard-code* (dituliskan langsung pada kode program).
3. Lakukan analisis terhadap hasil identifikasi *sense* yang tepat pada keenam contoh kalimat yang disebutkan pada nomor 2.

File yang harus dikumpulkan:

1. Program, berisi pendefinisian fungsi Lesk dan fungsi untuk menentukan *sense* yang paling sesuai untuk sebuah kata target berdasar fungsi Lesk.
2. Laporan: 1 file pdf, maksimum panjang laporan adalah 2 halaman. Laporan berisi jawaban pertanyaan nomor 2 yang menjelaskan alasan pemilihan kata-kata ambigu, serta jawaban pertanyaan nomor 3 dan analisisnya. Apakah hasil yang Anda peroleh sesuai dengan *sense* seharusnya, jika tidak, apa dugaan penyebabnya.

Detail penilaian:

- a. Program:
 - Kebenaran implementasi fungsi Lesk [30 poin]
 - Kebenaran implementasi pemilihan *sense* yang paling tepat berdasarkan fungsi Lesk [30 poin]
- b. Laporan:
 - kelengkapan jawaban nomor 2 dan 3 [20 poin]
 - analisis jawaban nomor 3 [20 poin]

Jika ada pertanyaan, silakan disampaikan melalui *channel* pekan_9_tugas_wsd di slack.