Gelombang 1

Open-world assumption: pada soal cerita, jika ada detail yang tidak disebutkan, maka hal tersebut bisa berlaku dan bisa juga tidak. Berikan jawaban anda sesuai dengan info yang ada. Tidak akan ada klarifikasi soal, kecuali jika ditemukan kontradiksi di dalam soal ujian sehingga tidak bisa dijawab.
Hint: Jawaban untuk masing-masing soal selain nomor terakhir pendek-pendek (1-2 kalimat). Jangan menghabiskan waktu terlalu lama untuk mereka. Jawab dengan singkat, benar, dan jelas.

A. Stopwords [16 poin]

---

1. **[3 poin]** Bob mengembangkan sistem perolehan informasi untuk korpus dengan domain tertentu dalam Bahasa Inggris. Setelah menganalisis data distribusi kemunculan kata, Bob menentukan bahwa 1000 kata dengan frekuensi tertinggi tidak deskriptif, sehingga ia masukkan sebagai stopwords. Setelah dievaluasi, Bob menemukan stopwords yang ia gunakan baik untuk sistemnya.

   Alice sedang mengembangkan sistem serupa dengan domain yang sama, tapi untuk korpus berbahasa Indonesia. Alice malas melakukan analisis ulang distribusi kata, sehingga dia hanya menghitung frekuensi kata dan, mengikuti Bob, langsung menentukan top-1000 kata sebagai stopwords.

   Apakah stopwords yang digunakan Alice bagus? Ya/tidak/tergantung, berikan alasan anda.

---

2. **[4 poin]** Alice diminta klien untuk membuat daftar *stopwords* untuk suatu sistem perolehan informasi yang menargetkan artikel domain kesehatan berbahasa Indonesia (detail penyakit/cedera dan cara penanganannya yang lengkap), namun dia tidak memiliki akses terhadap korpusnya, sehingga dia harus berpikir kira-kira kata-kata apa yang harus masuk ke stopwords.

   Bantu Alice menentukan apakah kata-kata di bawah ini (dipisahkan dengan koma) kira-kira **termasuk** stopwords atau **tidak**! Berikan alasan untuk masing-masing kata.
   (Kata-kata yang memiliki alasan yang sama bisa dikelompokkan untuk menghemat waktu)

   yang, dia, covid, gejala, kanker, pusing, budi, obat, rumput, hari, pijat

---

3. **[Total 9 poin]** Bob sedang mempertimbangkan mengembangkan stopwords domain kesehatan berbahasa Inggris dengan mempertimbangkan korpus lain yang dapat dia akses sehingga dia bisa melakukan analisis statistik untuk dijadikan landasan pembentukan stopwords.
   a. **[6 poin]** Untuk masing-masing opsi korpus di bawah ini, menurut anda apakah korpus tersebut baik untuk keperluan Bob. Jelaskan alasan anda.
      i. Wikipedia dump (https://dumps.wikimedia.org/enwiki/20201020/), yang kemudian didownload dan disimpan oleh Bob untuk analisis lebih lanjut

ii. Oscar corpus
iii. Brown corpus (https://en.wikipedia.org/wiki/Brown_Corpus)
iv. Korpus Journal of Medicine https://en.wikipedia.org/wiki/Journal_of_Medicine

b. **[3 poin]** Menurut anda, jika Bob hendak menggabungkan stopwords dari keempat opsi korpus di atas, kombinasi manakah yang akan menghasilkan hasil terbaik? Jelaskan!

B. Tokenization & Annotation [19 poin]

---

4. **[3 poin]** Alice mengembangkan sistem segmentasi kalimat dan tokenisasi untuk Bahasa Indonesia, ia beri nama “Tokenisakti”, yang ia klaim 100% akurat. Jika Bob ingin menggunakan Tokenisakti untuk Bahasa Inggris, kira-kira problem apa saja yang mungkin dia temukan? Sebutkan minimal 2 masalah.

5. **[Total 5 poin]** Bob sedang mengembangkan sistem tokenisasi dan anotasi named-entity recognition (NER) Bahasa Jepang yang ia beri nama “AnoTachiNeru”.
    a. [1.5 poin] Apa tantangan utama tokenisasi AnoTachiNeru yang tidak ditemukan pada Bahasa Inggris maupun Indonesia?
    b. [1.5 poin] Apa tantangan utama NER AnoTachiNeru yang tidak ditemukan pada Bahasa Inggris maupun maupun Indonesia?
    c. [2 poin] Misalkan Alice ingin memakai AnoTachiNeru untuk menganotasi *entities* pada korpus twitter berbahasa Indonesia *alay* (yang mungkin tercampur bahasa lain). Apakah AnoTachiNeru akan bisa melakukan anotasi sesuai keinginan Alice? Ya/tidak/tergantung, berikan alasan anda.

6. **[11 poin]** Bantu Alice mengembangkan program stemming Bahasa Indonesia. Untuk membantu anda, Alice sudah mempersiapkan 5 kamus untuk Anda gunakan.
    a. Kamus P: berisi *list* awalan (prefix) di Bahasa Indonesia (me, ber, etc.)
    b. Kamus S: berisi *list* akhiran (suffix) di Bahasa Indonesia (i, kan, nya, pun, etc.)
    c. Kamus D: berisi *list* kata dasar di Bahasa Indonesia (saya, kamu, etc.)
    d. Kamus U: Berisi *list* kata ulang semu (kupu-kupu, berang-berang, etc.)
    e. Kamus I: Berisi *map* dari kata bersisipan menuju kata dasarnya (seruling → suling)

    Anda dapat berasumsi bahwa kamus-kamus Alice ini benar dan lengkap. Tugas anda adalah untuk menuliskan *pseudocode* suatu fungsi/method yang menerima sebagai input suatu kata dan melakukan *stemming* sebagai outputnya. **Anda dapat berasumsi bahwa input yang diberikan akan selalu berupa kata yang valid dalam Bahasa Indonesia**. Anda tidak perlu memusingkan ambiguitas (e.g. beruang), asal pseudocode anda meng-output salah satu dari beberapa opsi yang benar. Alice akan menilai pseudocode anda dari *test cases* berikut: ikan → ikan, kupu-kupu → kupu-kupu, lari-lari → lari, berkuda → kuda, akhiri → akhir, memberlakukan → laku, peragawatinyapun → raga, berseruling → suling, berkejar-kejaran → kejar.
    Hint: anda dapat menggunakan fungsi bantu untuk mengakses dan memanipulasi string, seperti charAt, startsWith, endsWith, lowercase, uppercase, append, etc