Gelombang 3

Panjang-pendeknya jawaban tidak akan mempengaruhi nilai.
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.

IR Models & Eval [33 poin]

---

1. **[Total 17 poin] Boolean model indexing**
   Boolean model memerlukan indeks. Salah satu metode pengindeksan yang dibahas dalam kelas adalah pengindeksan dengan posisi kata (*index with word positions*).

   Bob mengembangkan sistem pengindeksan baru, yaitu indeks dengan set kata setelahnya, yang mana ia mirip dengan indeks dengan posisi kata, namun informasi posisi kata *keyword* per dokumen digantikan dengan informasi kata setelah *keyword* tersebut. Contoh:

   Dokumen:
   - Dok1: “Saya makan nasi padang”
   - Dok2: “Saya membaca buku yang saya beli”
   - …

   Index:
   - “Saya” → {(Dok1, {“makan”}), (Dok2, {“membaca”, “beli”}), ...}
   - “buku” → {(Dok2, {“yang”}), ...}

   Jawab dua pertanyaan berikut:
   a. Saat melakukan *phrase query*, apakah sistem ini lebih baik atau lebih buruk dibandingkan dengan pengindeksan dengan posisi kata? Uraikan dari sisi *correctness* dan *running time*. **Hint:** coba pikirkan jika kita melakukan *phrase query* dengan 2 kata dan lebih dari 2 kata.
   b. Dari poin a, jika anda menemukan kelemahan dari sistem ini dari sisi *correctness* maupun *running time*, apa yang bisa dilakukan untuk memperbaiki kelemahan tersebut?

2. **[16 poin] TF-IDF dan BM25**
   Diberikan 4 dokumen:
   Dok1 = “saya makan makan makan nasi padang padang padang padang padang”
   Dok2 = “saya saya saya pesan pesan nasi nasi nasi nasi nasi padang”
   Dok3 = “pesan pesan pesan nasi nasi nasi rendang rendang rendang rendang padang padang padang enak enak”
   Dok4 = “makan makan makan makan makan nasi nasi pedas pedas pedas pedas pedas pedas enak enak enak enak enak”

Buatlah matriks:

- TF
- FancyTF (sebagaimana digunakan dalam BM25)
- IDF (yang original, bukan yang BM25)
- TF*IDF
- FancyTF*IDF original

Dengan menggunakan TF*IDF biasa, urutkan relevansi dokumen menggunakan cosine similarity untuk query:

- “nasi padang”
- “rendang enak”