Soal 04 – Vector Space Model & Ranking

1. Tujuan

Tujuan dari eksperimen ini adalah untuk menerapkan **model ruang vektor (Vector Space Model / VSM)** dalam sistem temu kembali informasi.
Model ini digunakan untuk:

- Mewakili dokumen dan query dalam bentuk vektor numerik berdasarkan bobot TF-IDF.
- Mengukur kemiripan antara dokumen dan query menggunakan cosine similarity.
- Menampilkan hasil peringkat (ranking) dokumen yang paling relevan terhadap query pengguna.

2. Dasar Teori

a. TF-IDF (Term Frequency – Inverse Document Frequency)

Metode TF-IDF digunakan untuk menghitung bobot suatu kata terhadap sebuah dokumen.

- **TF (Term Frequency):** seberapa sering kata muncul dalam dokumen.
- **DF (Document Frequency):** jumlah dokumen yang mengandung kata tersebut.
- **IDF (Inverse Document Frequency):** logaritma kebalikan DF, untuk menurunkan bobot kata yang umum.

Rumus umum:

$$TFIDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right)$$

dengan N = jumlah total dokumen.

b. Vector Space Model (VSM)

VSM merepresentasikan dokumen dan query sebagai vektor dalam ruang multidimensi di mana setiap dimensi mewakili satu kata unik (term).
Kemiripan antara dua vektor dihitung menggunakan cosine similarity:

$$\text{cosine}(A,B) = \frac{A \cdot B}{||\ A\ ||\times||\ B\ ||}$$

Nilai cosine berkisar antara 0–1. Semakin mendekati 1 → semakin mirip dokumen dengan query.

c. Precision@k

Digunakan untuk mengukur ketepatan hasil pencarian.

$$Precision@k = \frac{\text{jumlah dokumen relevan pada top-k}}{k}$$

3. Implementasi

1. **Dataset:** 10 dokumen hasil preprocessing dari berbagai deskripsi buku (fantasi, romansa, sains, motivasi, horor, dll.) yang tersimpan di folder data/processed/.
2. **Bahasa Pemrograman:** Python
3. **Library yang digunakan:**
    - scikit-learn → TF-IDF dan cosine similarity
    - numpy, os, re untuk pemrosesan teks
4. **Langkah implementasi:**
    - Membaca seluruh dokumen hasil preprocessing.
    - Membentuk **TF-IDF matrix** menggunakan TfidfVectorizer().
    - Mengubah query pengguna menjadi vektor TF-IDF dengan vocabulary yang sama.
    - Menghitung **cosine similarity** antara query dan seluruh dokumen.
    - Mengurutkan hasil dan menampilkan **Top-3** dokumen teratas beserta snippet teks (120 karakter).
    - Mengukur **Precision@3** untuk setiap query berdasarkan *gold standard* dari tugas sebelumnya.

4. Hasil Uji

Query 1: pedang hutan

```
============================================================
Query: pedang hutan

Top-3 Hasil Ranking:
1. buku_fantasi.txt          | cosine=0.3487 | anak lakilaki bernama arka menemukan pedang ajaib tersembunyi hutan terlarang bantuan
penyihir tua menyelamatkan kerajaa
2. buku_sains.txt            | cosine=0.0000 | penuli fenomena alam mudah dipahami gerhana matahari misteri lubang hitam angkasa
3. buku_romansa.txt          | cosine=0.0000 | kisah cinta insan terhalang jarak berjuang mempertahankan hubungan kesibukan ambisi ma
singmas

Precision@3: 0.33
============================================================
```

Query 2: cinta motivasi

```
============================================================
Query: cinta motivasi

Top-3 Hasil Ranking:
1. buku_romansa.txt          | cosine=0.3015 | kisah cinta insan terhalang jarak berjuang mempertahankan hubungan kesibukan ambisi ma
singmas
2. buku_sains.txt            | cosine=0.0000 | penuli fenomena alam mudah dipahami gerhana matahari misteri lubang hitam angkasa
3. buku_petualangan.txt      | cosine=0.0000 | sahabat perjalanan menantang puncak gunung tertinggi menghadapi badai jurang ketakutan
 terbesar mencapai tujuan

Precision@3: 0.33
============================================================
```

Query 3: ilmu sains pengetahuan

```
Query: ilmu sains pengetahuan

Top-3 Hasil Ranking:
1. buku_sains.txt            | cosine=0.0000 | penuli fenomena alam mudah dipahami gerhana matahari misteri lubang hitam angkasa
2. buku_romansa.txt          | cosine=0.0000 | kisah cinta insan terhalang jarak berjuang mempertahankan hubungan kesibukan ambisi ma
singmas
3. buku_petualangan.txt      | cosine=0.0000 | sahabat perjalanan menantang puncak gunung tertinggi menghadapi badai jurang ketakutan
 terbesar mencapai tujuan

Precision@3: 0.33
PS D:\TUUUUGGGGGGAAAAASSSSSSS\stki-uts-A11.2023.15390-AtanasiusMarcello>
```