Hitungan Manual: BoW, TF–IDF, dan Cosine Similarity

1. Dataset Mainan

Tiga dokumen pendek (lowercase, tanpa penghapusan stopword agar mudah dihitung):

- d_1: “barang cepat sampai”
- d_2: “pengiriman cepat dan rapi”
- d_3: “sangat lambat pengiriman”

Kosakata global (urut abjad) kita tetapkan sebagai

$$\mathcal{V} = [\text{barang}, \text{cepat}, \text{sampai}, \text{pengiriman}, \text{dan}, \text{rapi}, \text{sangat}, \text{lambat}],$$

dengan ukuran korpus $N = 3$ dokumen.

2. Bag-of-Words (BoW)

2.1 Definisi

Representasi BoW untuk sebuah dokumen d adalah vektor frekuensi:

$$\text{BoW}(d) = [f(w_1, d),\ f(w_2, d),\ \ldots,\ f(w_V, d)],$$

dengan $f(w_i, d)$ menyatakan jumlah kemunculan kata w_i dalam d, dan $V = |\mathcal{V}|$.

2.2 Perhitungan BoW

Panjang dokumen:

$$|d_1| = 3, \quad |d_2| = 4, \quad |d_3| = 3.$$

Susunan vektor mengikuti urutan $\mathcal{V}$.

$$\text{BoW}(d_1) = [1,\ 1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0]$$
$$\text{BoW}(d_2) = [0,\ 1,\ 0,\ 1,\ 1,\ 1,\ 0,\ 0]$$
$$\text{BoW}(d_3) = [0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1,\ 1]$$

3. TF, DF, dan IDF

3.1 Term Frequency (TF)

Gunakan TF ternormalisasi (proporsional terhadap panjang dokumen):

$$\text{tf}(t, d) = \frac{f_{t,d}}{|d|}.$$

Contoh:

- Di d_1: setiap kata yang muncul memiliki tf $= \frac{1}{3}$.
- Di d_2: setiap kata yang muncul memiliki tf $= \frac{1}{4}$.
- Di d_3: setiap kata yang muncul memiliki tf $= \frac{1}{3}$.

3.2 Document Frequency (DF)

$\mathrm{df}(t)$ adalah jumlah dokumen yang *mengandung* kata t:

$$\mathrm{df}(\text{barang}) = 1\ (d_1), \quad \mathrm{df}(\text{cepat}) = 2\ (d_1, d_2), \quad \mathrm{df}(\text{sampai}) = 1\ (d_1),$$
$$\mathrm{df}(\text{pengiriman}) = 2\ (d_2, d_3), \quad \mathrm{df}(\text{dan}) = 1\ (d_2), \quad \mathrm{df}(\text{rapi}) = 1\ (d_2),$$
$$\mathrm{df}(\text{sangat}) = 1\ (d_3), \quad \mathrm{df}(\text{lambat}) = 1\ (d_3).$$

3.3 Inverse Document Frequency (IDF)

Pakai varian IDF yang stabil terhadap pembagi nol:

$$\mathrm{idf}(t) = \ln\Big(\frac{N}{1 + \mathrm{df}(t)}\Big), \quad N = 3.$$

Maka:

$$\mathrm{df}(t) = 1 \ \Rightarrow\ \mathrm{idf}(t) = \ln\Big(\tfrac{3}{2}\Big) \approx 0.4055, \qquad \mathrm{df}(t) = 2 \ \Rightarrow\ \mathrm{idf}(t) = \ln\Big(\tfrac{3}{3}\Big) = 0.$$

Konsekuensinya:

$$\mathrm{idf}(\text{barang}) = \mathrm{idf}(\text{sampai}) = \mathrm{idf}(\text{dan}) = \mathrm{idf}(\text{rapi}) = \mathrm{idf}(\text{sangat}) = \mathrm{idf}(\text{lambat}) \approx 0.4055,$$
$$\mathrm{idf}(\text{cepat}) = \mathrm{idf}(\text{pengiriman}) = 0.$$

4. Vektor TF–IDF

Definisi dasar:

$$\mathrm{tfidf}(t, d) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t).$$

Kita tampilkan hanya komponen yang tidak nol (pembulatan 4 desimal).

4.1 Dokumen d_1 ("barang cepat sampai")

$$\mathrm{tfidf}(\text{barang}, d_1) = \tfrac{1}{3} \cdot 0.4055 \approx 0.1352$$
$$\mathrm{tfidf}(\text{cepat}, d_1) = \tfrac{1}{3} \cdot 0 = 0$$
$$\mathrm{tfidf}(\text{sampai}, d_1) = \tfrac{1}{3} \cdot 0.4055 \approx 0.1352$$

Vektor TF–IDF (urutan $\mathcal{V}$):

$$[\,0.1352,\ 0,\ 0.1352,\ 0,\ 0,\ 0,\ 0,\ 0\,].$$

4.2 Dokumen d_2 ("pengiriman cepat dan rapi")

$$\mathrm{tfidf}(\text{pengiriman}, d_2) = \tfrac{1}{4} \cdot 0 = 0, \quad \mathrm{tfidf}(\text{cepat}, d_2) = \tfrac{1}{4} \cdot 0 = 0,$$
$$\mathrm{tfidf}(\text{dan}, d_2) = \tfrac{1}{4} \cdot 0.4055 \approx 0.1014, \quad \mathrm{tfidf}(\text{rapi}, d_2) = \tfrac{1}{4} \cdot 0.4055 \approx 0.1014.$$

Vektor TF–IDF:

$$[\,0,\ 0,\ 0,\ 0,\ 0.1014,\ 0.1014,\ 0,\ 0\,].$$

4.3 Dokumen d_3 ("sangat lambat pengiriman")

$$\mathrm{tfidf}(\text{sangat}, d_3) = \tfrac{1}{3} \cdot 0.4055 \approx 0.1352, \quad \mathrm{tfidf}(\text{lambat}, d_3) = \tfrac{1}{3} \cdot 0.4055 \approx 0.1352,$$
$$\mathrm{tfidf}(\text{pengiriman}, d_3) = \tfrac{1}{3} \cdot 0 = 0.$$

Vektor TF–IDF:

$$[\,0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.1352,\ 0.1352\,].$$

5. Cosine Similarity (Contoh)

Misalkan ada *query* q: “barang cepat”. Dengan kosakata yang sama,

$$|q| = 2, \quad \text{tf}(\text{barang}, q) = \tfrac{1}{2}, \quad \text{tf}(\text{cepat}, q) = \tfrac{1}{2}.$$

Dengan $\text{idf}(\text{barang}) \approx 0.4055$ dan $\text{idf}(\text{cepat}) = 0$:

$$\text{tfidf}(\text{barang}, q) = \tfrac{1}{2} \cdot 0.4055 \approx 0.2027, \qquad \text{tfidf}(\text{cepat}, q) = \tfrac{1}{2} \cdot 0 = 0.$$

Vektor TF–IDF q (urutan $\mathcal{V}$):

$$\mathbf{q} = [\,0.2027,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0\,].$$

Cosine similarity antara $\mathbf{q}$ dan $\mathbf{d}$ didefinisikan:

$$\cos(\theta) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\|\,\|\mathbf{d}\|}.$$

Dengan d_1.

$$\mathbf{d_1} = [\,0.1352,\ 0,\ 0.1352,\ 0,\ 0,\ 0,\ 0,\ 0\,].$$

$$\mathbf{q} \cdot \mathbf{d_1} = 0.2027 \times 0.1352 + 0 \times 0.1352 = 0.0274.$$

$$\|\mathbf{q}\| = \sqrt{0.2027^2} \approx 0.2027, \qquad \|\mathbf{d_1}\| = \sqrt{0.1352^2 + 0.1352^2} \approx \sqrt{2 \times 0.0183} \approx 0.1913.$$

$$\cos(\mathbf{q}, \mathbf{d_1}) \approx \frac{0.0274}{0.2027 \times 0.1913} \approx \frac{0.0274}{0.0388} \approx 0.706.$$

Dengan d_2 dan d_3. Karena $\mathbf{q}$ hanya memiliki komponen “barang” dan d_2, d_3 tidak memiliki “barang”,

$$\mathbf{q} \cdot \mathbf{d_2} = 0, \quad \mathbf{q} \cdot \mathbf{d_3} = 0 \ \Rightarrow\ \cos(\mathbf{q}, \mathbf{d_2}) = 0, \quad \cos(\mathbf{q}, \mathbf{d_3}) = 0.$$

Kesimpulan: q paling mirip ke d_1 (nilai cosine terbesar).

6. Catatan Singkat tentang N-gram

Jika menggunakan *bigram* untuk menangkap negasi:

$$\text{dokumen: “tidak bagus”} \quad \Rightarrow \quad \text{fitur bigram: “tidak\_bagus”}.$$

BoW bigram:

$$\text{BoW}_{\text{bigram}}(d) = [\,f(\text{“tidak\_bagus”}, d), \ldots\,].$$

Skor TF–IDF bigram dihitung sama: gunakan frekuensi bigram sebagai $f_{t,d}$ lalu kalikan dengan $\text{idf}(t)$.

7. Ringkasan

- BoW menghitung *frekuensi kata* per dokumen.
- TF–IDF menurunkan bobot kata yang *terlalu umum* dan meninggikan kata *informatif*.
- Cosine similarity mengukur kedekatan vektor TF–IDF antar teks.
- Vektor TF–IDF siap dipakai sebagai fitur untuk model klasik (LogReg/SVM/NB).