Perhitungan Matriks Jarak pada Dokumen Teks

Luh Sukma Mulyani
2108541027

Universitas Udayana

Juni 11, 2025

Outline

Pendahuluan

Latar Belakang

Pentingnya Matriks Jarak dalam Analisis Teks

Matriks jarak merupakan fondasi utama dalam analisis data berbasis teks yang berfungsi untuk:

- Menggambarkan hubungan antar dokumen dalam bentuk numerik
- Memfasilitasi proses *klasifikasi*, *klusterisasi*, dan *pencarian informasi*
- Mengukur tingkat **kesamaan** atau **perbedaan** antar dokumen

Text Mining dan Document Similarity

Konsep Fundamental dalam Analisis Teks

Text Mining Merupakan cabang data mining yang mengekstraksi pengetahuan tersembunyi dari data teks tidak terstruktur untuk mendukung pengambilan keputusan. Tugas utama:

- Klasifikasi
- Klasterisasi
- Asosiasi

Kesamaan antar dokumen menjadi komponen fundamental dalam proses ini. **Document Similarity** Mengukur kesamaan atau perbedaan antar dokumen berdasarkan konten tekstual. Berperan penting dalam:

- Sistem rekomendasi
- Deteksi plagiarisme
- Information retrieval

Pengukuran similarity memerlukan *text preprocessing* untuk menghasilkan representasi numerik dalam perhitungan matriks jarak.

Text Preprocessing

Pipeline Preprocessing

6 Tahapan Utama:

1. **Lower Case** - Normalisasi huruf kapital
2. **Cleaning Data** - Hapus karakter khusus & punctuation
3. **Tokenisasi** - Pemecahan teks menjadi token
4. **Normalisasi** - Mengubah kata tidak baku menjadi baku
5. **Stopwords Removal** - Hapus kata-kata umum
6. **Stemming** - Kembalikan kata ke bentuk dasar

Contoh:

```
"Plz help me getting 1.000 Subscribers tonight/today. Thanks to
                    all who sub me ï»¿"
                              ↓
['pleas', 'help', 'get', 'numer', 'subscrib', 'tonight', 'today',
                    'thank', 'subscrib']
```

Representasi Dokumen

Bag of Words

Bag of Words (BoW)

- Representasi teks sebagai kumpulan kata tanpa mempertimbangkan urutan.
- Hanya menghitung **frekuensi kemunculan** kata dalam dokumen.
- **Contoh:** "belajar data mining" $\rightarrow$ {belajar: 1, data: 1, mining: 1}

Term Frequency–Inverse Document Frequency (TF-IDF)

Metode representasi teks yang mengukur kepentingan kata dalam sebuah dokumen relatif terhadap seluruh korpus.

TF-IDF Standar

- **TF:**

$$\mathrm{TF}_{ij} = \frac{f_{ij}}{\sum f_j} \tag{2.1}$$

  Frekuensi kata dalam dokumen tertentu

- **IDF:**

$$\mathrm{IDF}_i = \log\left(\frac{N}{df_i}\right) \tag{2.2}$$

  Ukuran jarangnya kata di seluruh dokumen

- **TF-IDF:**

$$\mathrm{w}_{ij} = \mathrm{TF}_{ij} \times \mathrm{IDF}_i \tag{2.3}$$

Implementasi Scikit-learn

- **Modul:** `TfidfVectorizer`, `TfidfTransformer`
- **IDF:**

$$\mathrm{idf}(t) = \log\left(\frac{1+nd}{1+df(t)}\right) + 1 \tag{2.4}$$

- **L2 Normalisasi:**

$$v_{\mathrm{norm}} = \frac{v}{\|v\|_2} \tag{2.6}$$

Document Term Matrix (DTM)

Definisi dan Konsep Dasar

Document Term Matrix adalah representasi numerik dari korpus yang terdiri dari N dokumen dan n kosakata unik, dalam bentuk matriks $A_{N \times n}$, dengan:

- N: Jumlah dokumen dalam korpus
- n: Jumlah kata unik dalam kamus
- A_{ij}: Frekuensi kemunculan kata ke-i pada dokumen ke-j

Contoh Implementasi (3 dokumen, 4 kata)

$$A = \begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 3 & 0 & 2 \\ 1 & 2 & 0 & 0 \end{bmatrix}$$

Metrik Jarak

Definisi Metrik Jarak

Metrik Jarak Metrik jarak adalah fungsi matematis yang mengukur tingkat kedekatan atau kemiripan antar objek dalam ruang berdimensi tinggi.

Notasi Formal

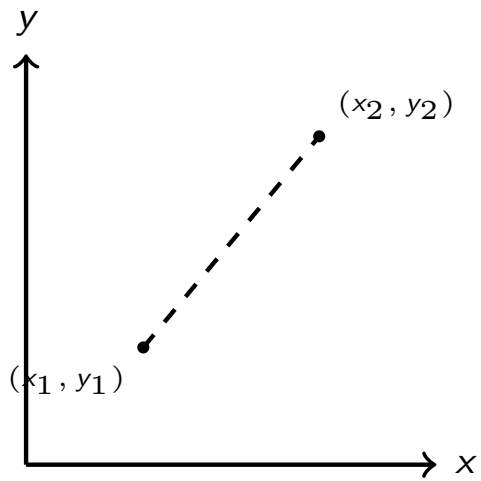
$$d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$$

Aksioma Metrik

- **Non-negativity**: $d(x, y) \geq 0$
- **Identity of Indiscernibles**: $d(x, y) = 0 \iff x = y$
- **Symmetry**: $d(x, y) = d(y, x)$
- **Triangle Inequality**: $d(x, y) \leq d(x, z) + d(z, y)$

Jarak Euclidean (L2-norm)

Definisi: Euclidean Distance (L2-norm) adalah ukuran jarak lurus terpendek antara dua titik dalam ruang berdimensi-n, berdasarkan teorema Pythagoras.



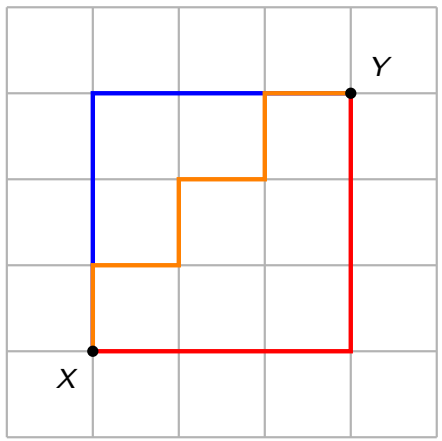
Ilustrasi Jarak Euclidean

Rumus

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.11}$$

Jarak Manhattan (L1-norm)

Definisi:
Manhattan Distance (L1-norm, City-block distance, atau Taxicab distance) mengukur jarak berdasarkan perpindahan sepanjang sumbu koordinat, seperti pola gerak taksi di kota.



Ilustrasi Jarak Manhattan

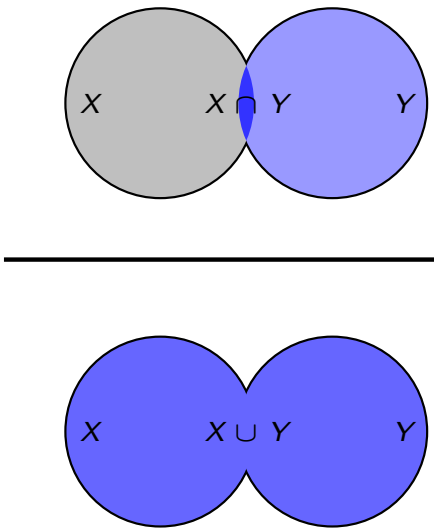
Rumus

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.12}$$

Jarak Jaccard (Jaccard Distance)

Definisi:

Jaccard Distance mengukur ketidaksamaan antara dua himpunan. Nilainya diperoleh dari komplemen Jaccard Similarity, yaitu perbandingan antara irisan dan gabungan dari dua himpunan.



Ilustrasi Jaccard: Irisan dan Gabungan

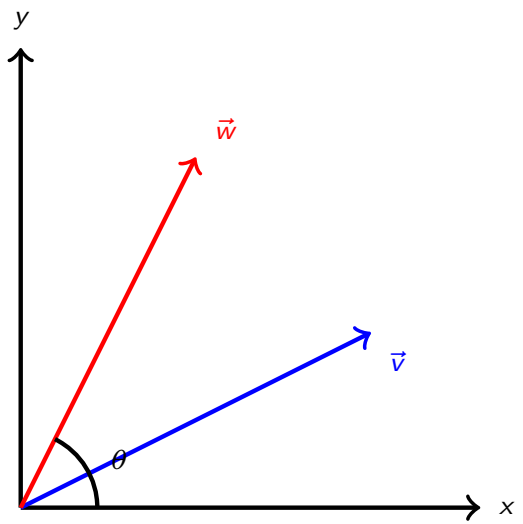
Rumus

$$\text{Jaccard}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{2.13}$$

$$d_{\text{Jaccard}}(X, Y) = 1 - \text{Jaccard}(X, Y) \tag{2.14}$$

Jarak Cosine (Cosine Distance)

Definisi:
Cosine Distance mengukur ketidaksamaan arah antara dua vektor dalam ruang berdimensi tinggi, mengabaikan besar (magnitudo) vektor.



Ilustrasi Cosine Distance

Formula

$$\text{Cosine}(v, w) = \frac{v \cdot w}{\|v\| \cdot \|w\|} \tag{2.15}$$

$$d_{\text{Cosine}}(v, w) = 1 - \text{Cosine}(v, w) \tag{2.16}$$

Matriks Jarak

Matriks Jarak

Struktur Matriks Jarak:

- **Symmetric matrix:** $d(i, j) = d(j, i)$
- **Diagonal = 0 :** $d(i, i) = 0$
- **Ukuran:** $n \times n$ untuk n dokumen

Representasi Matematis:

$$D = \begin{bmatrix} d(\text{dok}_1, \text{dok}_1) & d(\text{dok}_1, \text{dok}_2) & \cdots & d(\text{dok}_1, \text{dok}_n) \\ d(\text{dok}_2, \text{dok}_1) & d(\text{dok}_2, \text{dok}_2) & \cdots & d(\text{dok}_2, \text{dok}_n) \\ \vdots & \vdots & \ddots & \vdots \\ d(\text{dok}_n, \text{dok}_1) & d(\text{dok}_n, \text{dok}_2) & \cdots & d(\text{dok}_n, \text{dok}_n) \end{bmatrix}$$

Interpretasi: Nilai kecil = dokumen similar, nilai besar = dokumen berbeda

Studi Kasus

Studi Kasus: Perhitungan Matriks Jarak

Overview Studi Kasus

- **Dataset:** 10 dokumen teks acak dari korpus tugas akhir
- **Vocabulary:** 20 kata unik setelah tahap preprocessing

Pipeline Pemrosesan Teks:

Hasil akhir: Matriks 10×10 *yang mencerminkan kedekatan antar dokumen berbasis representasi numerik teks.*

Tahapan Preprocessing Teks

- **Case Folding** –Mengubah seluruh teks menjadi huruf kecil
- **Cleaning** –Menghapus tanda baca dan karakter khusus
- **Tokenization** –Memecah teks menjadi unit kata (token)
- **Normalisasi** - Mengubah kata tidak baku menjadi baku (typo, singkatan)
- **Stopwords Removal** –Menghapus kata-kata umum yang tidak informatif
- **Stemming** –Mengubah kata menjadi bentuk dasar/akar katanya

Contoh Transformasi Teks:

Sebelum

```
"2 billion....Coming soon"
```

Setelah Preprocessing

```
[ 'numer', 'billion', 'come', 'soon' ]
```

Bag of Words (Binary)

Konsep: Representasi dokumen dalam bentuk vektor biner:

- 1 jika kata **muncul** di dokumen
- 0 jika kata **tidak muncul**

Contoh (10 Dokumen × 20 Kata)

- Doc1: [0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0]
- Doc6: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0]

TF-IDF Representation

- **Term Frequency (TF)** mengukur frekuensi kemunculan kata ke-*i* dalam dokumen ke-*j*. Contohnya:

$$tf(\text{'url'}, D8) = 1 \qquad tf(\text{'happy\_emoji'}, D10) = 3$$

- **Inverse Document Frequency IDF** Perhitungan IDF untuk fitur `'url'` :

$$n_d = 10, \qquad \text{df}_{\texttt{url}} = 4$$

$$\text{idf}_{\texttt{url}} = \log\left(\frac{11}{5}\right) + 1 = \log(2.2) + 1 \approx 1{,}788$$

- **Pembobotan TF-IDF:**

$$\text{w}_{\texttt{url,D8}} = 1 \times 1{,}788 = 1{,}788$$

TF-IDF Representation

- **Normalisasi L2**

$$w_{D8} = [2.704, 2.704, 2.704, 1.788, 2.704]$$

Panjang vektor dihitung sebagai:

$$||D_8||_2 = \sqrt{2.704^2 + 2.704^2 + 2.704^2 + 1.788^2 + 2.704^2}$$

$$= \sqrt{32.44} \approx 5.6956$$

Vektor hasil normalisasi:

$$w_{norm}(D8) = \left[\frac{2.704}{5.6956}, \frac{2.704}{5.6956}, \frac{2.704}{5.6956}\frac{1.788}{5.6956}, \frac{2.704}{5.6956}\right]$$

$$w_{normalized}(D8) \approx [0.474,\ 0.474,\ \ 0.474,\ \ 0.313,\ \ 0.474]$$

Manhattan Distance : Dokumen 1 vs Dokumen 6

Rumus:

$$d_{\text{Manhattan}}(D_1, D_6) = \sum_{i=1}^{n} |d_{1,1} - d_{6,1}| + |d_{1,2} - d_{6,2}| + \ldots + |d_{1,20} - d_{6,20}|$$

Perhitungan Langkah-per-Langkah:

$$\sum_{i=1}^{20} |d_{1,i} - d_{6,i}| = |0.539 - 0| + |0.539 - 0| + |0 - 0.789| + |0.356 - 0.613| + |0.539 - 0|$$

$$= 0.539 + 0.539 + 0.789 + 0.257 + 0.539 = \mathbf{2.663}$$

Hasil Akhir:

$$d_{Manhattan}(D_1, D_6) = 2.663$$

Jaccard Similarity dan Distance: Dokumen 1 vs Dokumen 6

Hitung Matriks Biner:

$$M_{11} = 1, \quad M_{10} = 3, \quad M_{01} = 1$$

Jaccard Similarity:

$$J(A, B) = \frac{M_{11}}{M_{11} + M_{10} + M_{01}} = \frac{1}{1 + 3 + 1} = \frac{1}{5} = \mathbf{0.2}$$

Jaccard Distance:

$$D_J(A, B) = 1 - J(A, B) = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01}} = \frac{3 + 1}{5} = \mathbf{0.8}$$

Euclidean Distance: Dokumen 1 vs Dokumen 6 (TF-IDF)

Rumus:

$$d_{\text{Euc}}(D_1, D_6) = \sqrt{\sum_{i=1}^{n}(d_{1,1} - d_{6,1})^2 + (d_{1,2} - d_{6,2})^2 + \ldots + (d_{1,20} - d_{6,20})^2}$$

Perhitungan:

$$\sum_{i=1}^{20}(d_{1,i} - d_{6,i})^2 = \sqrt{(0.539)^2 + (0.539)^2 + (-0.789)^2 + (0.356 - 0.613)^2 + (0.539)^2}$$

$$= \sqrt{1.560133} \approx \textbf{1.24905284}$$

Hasil Akhir:

$$d(D_1, D_6) = 1.24905284$$

Cosine Similarity dan Cosine Distance: Dokumen 1 vs Dokumen 6 (TF-IDF)

Rumus Cosine Similarity:

$$S_C(D_1\, D_6) = \cos(\theta) = \frac{D_1 \cdot D_6}{\|D_1\| \cdot \|D_6\|} = \frac{\sum_{i=1}^{n} d_{1,i} d_{6,i}}{\sqrt{\sum_{i=1}^{n} d_{1,i}^2} \cdot \sqrt{\sum_{i=1}^{n} d_{6,i}^2}}$$

Langkah-langkah:

- **Dot Product:** $D_1 \cdot D_6 = 0.356 \times 0.613 = 0.218$
- **Panjang Vektor:**

$$\|D_1\| = \sqrt{0.539^2 + 0.539^2 + 0.356^2 + 0.539^2} = \sqrt{0.996} \approx 0.997$$

$$\|D_6\| = \sqrt{0.789^2 + 0.613^2} = \sqrt{0.997} \approx 0.998$$

Cosine Similarity:

$$\mathrm{SC}(D_1, D_6) = \frac{0.218}{0.997 \times 0.998} \approx 0.219$$

Cosine Distance: $\mathrm{DC}(\mathrm{D}_1, D_6) = 1 - 0.219 = 0.781$

Kesimpulan

Kesimpulan

Perhitungan matriks jarak memungkinkan pengukuran kemiripan antar dokumen teks secara numerik. Proses ini diawali dengan prapemrosesan, dilanjutkan dengan representasi vektor (BoW/TF-IDF), dan perhitungan jarak menggunakan metrik tertentu seperti Manhattan atau Jaccard. Pemahaman proses ini penting untuk analisis struktur dan pola dalam kumpulan dokumen teks.

Terima Kasih

Pertanyaan & Diskusi