

**MAKALAH ALJABAR LINEAR
TERAPAN ALJABAR LINEAR UNTUK MENGUKUR
KEMIRIPAN DUA BUAH DOKUMEN**



**DISUSUN OLEH
LA ODE MUHAMMAD YUDHY PRAYITNO
(E1E122064)**

**JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HALU OLEO
KENDARI
2023**

KATA PENGANTAR

Tiada kalimat yang pantas saya ucapkan kecuali rasa syukur kepada Tuhan Yang Maha Esa atas selesainya makalah yang berjudul " Terapan Aljabar Linear Untuk Mengukur Kemiripan Dua Buah Dokumen ". Saya mengucapkan rasa terima kasih kepada semua pihak yang telah turut memberikan kontribusi dalam penyusunan makalah ini. Tentunya tidak akan bisa maksimal jika tidak mendapat dukungan dari berbagai pihak.

Tidak lupa pula kami mengucapkan rasa terima kasih kepada Bapak Natalis Ransi, S.Si.,M.Cs. selaku dosen pengampu matakuliah aljabar linear yang telah memberikan tugas ini.

Sebagai penyusun saya menyadari bahwa masih terdapat kekurangan, baik dari penyusunan maupun tata bahasa penyampaian dalam makalah ini. Oleh karena itu, saya dengan rendah hati menerima saran dan kritik dari pembaca agar kami dapat memperbaiki makalah ini. Saya berharap semoga makalah yang disusun ini memberikan manfaat dan juga inspirasi untuk pembaca.

DAFTAR ISI

KATA PENGANTAR	ii
DAFTAR ISI	iii
BAB I	1
1.1 Latar belakang	1
1.2 Rumusan masalah.....	2
1.3 Tujuan penelitian	2
1.4 Ruang lingkup makalah.....	2
BAB II.....	4
2.1 Definisi Matriks.....	4
2.2 Operasi Matriks (Penjumlahan, Perkalian, Invers).....	4
2.3 Transformasi Linier	6
BAB III	7
3.1 Konversi dokumen menjadi vector.....	7
3.2 Metode penghitungan bobot term dalam vektor dokumen.....	8
3.3 Penggunaan matriks term-dokumen	9
BAB IV	10
4.1 Metode cosine similarity	10
4.2 Perhitungan kemiripan dengan dot product.....	11
4.3 Penggunaan matriks cosine similarity	11
BAB V.....	13
5.1 Penerapan aljabar linear dalam pemrosesan bahasa alami	13
5.2 Contoh pengukuran kemiripan dokumen dengan aljabar linear.....	14
BAB VI	16
6.1 Kelebihan dan kekurangan pendekatan menggunakan aljabar linear.....	16
6.2 Perbandingan dengan metode lain yang digunakan dalam pengukuran kemiripan dokumen	17
BAB VII.....	18
7.1 Kesimpulan.....	18
7.2 Saran	19
DAFTAR PUSTAKA	20

BAB I

PENDAHULUAN

1.1 Latar belakang

Aljabar linear adalah cabang matematika yang mempelajari ruang vektor dan transformasi linear antara ruang vektor tersebut. Dalam konteks pengukuran kemiripan vektor, kita dapat menggunakan berbagai metode dan alat matematika yang disediakan oleh aljabar linear. Beberapa konsep yang relevan dalam aljabar linear yang dapat diterapkan dalam pengukuran kemiripan vektor antara lain:

1. Norma vektor: Norma vektor adalah ukuran atau panjang suatu vektor. Berbagai jenis norma, seperti norma Euclidean atau norma Manhattan, dapat digunakan untuk mengukur jarak atau kemiripan antara dua vektor.
2. Produk dalam: Produk dalam antara dua vektor memungkinkan kita untuk mengukur sudut antara vektor-vektor tersebut. Misalnya, produk dalam dot (titik) digunakan untuk menghitung sudut antara vektor-vektor dalam ruang Euclidean.
3. Matriks: Matriks adalah tabel berisi elemen-elemen yang terdiri dari baris dan kolom. Matriks dapat digunakan untuk merepresentasikan vektor-vektor dan melakukan operasi seperti perkalian matriks. Dalam konteks pengukuran kemiripan vektor, matriks dapat digunakan untuk mengubah atau mentransformasi vektor-vektor ke dalam ruang dimensi yang berbeda.
4. Ruang vektor: Konsep ruang vektor memungkinkan kita untuk memodelkan dan memanipulasi vektor-vektor secara matematis. Dalam pengukuran kemiripan vektor, kita dapat memperlakukan vektor-vektor sebagai titik-titik dalam ruang vektor, dan mengukur jarak atau kemiripan antara titik-titik tersebut menggunakan konsep-konsep dalam aljabar linear.

Terapan aljabar linear untuk mengukur kemiripan dua buah vektor, dapat menjelaskan konsep-konsep di atas dan mengilustrasikan penggunaannya dalam beberapa contoh kasus. Contohnya, bagaimana mengukur kemiripan antara dua

gambar menggunakan representasi vektor dari gambar-gambar tersebut, atau bagaimana mengukur kemiripan antara dua dokumen teks berdasarkan vektor representasi kata-kata yang ada di dalamnya.

1.2 Rumusan masalah

Adapun rumusan masalah dari makalah “Terapan Aljabar Linear Untuk Mengukur Kemiripan Dua Buah Dokumen” adalah sebagai berikut:

1. Bagaimana konsep dasar aljabar linear?
2. Bagaimana cara merepresentasi vektor dokumen?
3. Apa saja cara yang digunakan untuk mengukur kemiripan dokumen?
4. Bagaimana cara pengimplementasiannya dan contohnya?

1.3 Tujuan

Adapun tujuan dari pembuatan makalah “Terapan Aljabar Linear Untuk Mengukur Kemiripan Dua Buah Dokumen” adalah sebagai berikut:

1. Untuk mengetahui konsep dasar aljabar linear!
2. Untuk mengetahui cara mengimplementasikan vektor dokumen!
3. Untuk mengetahui cara yang digunakan untuk mengukur kemiripan dokumen!
4. Untuk mengetahui cara mengimplementasikan dan contohnya!

1.4 Ruang lingkup makalah

Ruang lingkup makalah “Terapan Aljabar Linear Untuk Mengukur Kemiripan Dua Buah Dokumen” dapat mencakup beberapa aspek berikut:

1. Representasi Vektor Dokumen: Fokus pada pengembangan metode representasi vektor dokumen menggunakan aljabar linear, di mana setiap dokumen direpresentasikan sebagai vektor dalam ruang vektor.
2. Metode Pengukuran Kemiripan: Melibatkan eksplorasi dan pengembangan metode pengukuran kemiripan antara dua vektor dokumen, seperti cosine similarity, dot product, atau metode lainnya yang berbasis aljabar linear.
3. Pemrosesan Vektor Dokumen: Membahas pemrosesan vektor dokumen menggunakan operasi aljabar linear, seperti penjumlahan, perkalian, dan pembagian vektor dokumen untuk mendapatkan hasil yang berguna dalam analisis kemiripan.

4. Reduksi Dimensi: Mempertimbangkan teknik reduksi dimensi, seperti Singular Value Decomposition (SVD) atau Principal Component Analysis (PCA), untuk mengurangi dimensi matriks term-dokumen dan memperbaiki kualitas representasi vektor dokumen.
5. Aplikasi Pemrosesan Bahasa Alami: Menjelajahi penggunaan metode aljabar linear dalam aplikasi pemrosesan bahasa alami, seperti temu kembali informasi, klasifikasi dokumen, analisis sentimen, atau penerjemahan mesin.
6. Evaluasi dan Pembandingan: Mengevaluasi kinerja metode aljabar linear dalam mengukur kemiripan dokumen dengan membandingkan hasilnya dengan metode lain yang ada, serta melakukan analisis dan diskusi terkait kelebihan dan kekurangan metode yang digunakan.

BAB II

KONSEP DASAR ALJABAR LINEAR

2.1 Definisi Matriks

Matriks adalah susunan bilangan-bilangan atau elemen-elemen dalam bentuk baris dan kolom. Matriks dinyatakan dalam bentuk tabel dengan bilangan-bilangan tersebut ditempatkan di dalamnya. Sebagai contoh, matriks 2×3 terdiri dari 2 baris dan 3 kolom:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

Dalam contoh di atas, a_{11} , a_{12} , a_{13} adalah elemen-elemen pada baris pertama, sedangkan a_{21} , a_{22} , a_{23} adalah elemen-elemen pada baris kedua. Matriks sering digunakan dalam berbagai bidang matematika, seperti aljabar linear, teori graf, dan pemrosesan sinyal. Setiap elemen dalam matriks dapat dinyatakan dengan indeks i dan j , di mana i menunjukkan nomor baris dan j menunjukkan nomor kolom. Misalnya, a_{ij} adalah elemen pada baris ke- i dan kolom ke- j .

Matriks juga memiliki operasi-operasi matematika yang dapat dilakukan, seperti penjumlahan, pengurangan, dan perkalian. Selain itu, terdapat juga konsep matriks khusus, seperti matriks identitas, matriks nol, dan matriks persegi. Matriks memiliki banyak aplikasi dalam berbagai bidang, seperti pemrograman komputer, analisis data, statistik, teori sistem, dan grafika komputer.

2.2 Operasi Matriks (Penjumlahan, Perkalian, Invers)

2.2.1 Penjumlahan Matriks

Operasi penjumlahan matriks dilakukan dengan menjumlahkan elemen-elemen yang berada pada posisi yang sama dalam dua matriks dengan ukuran yang sama.

Contoh:

Misalkan kita memiliki dua matriks berikut:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

Maka, penjumlahan matriks A dan B adalah:

$$A + B = [[1+5, 2+6], \\ [3+7, 4+8]] = [[6, 8], \\ [10, 12]]$$

2.2.1 Perkalian Matriks

Terdapat dua jenis perkalian matriks: perkalian matriks skalar dan perkalian matriks antar-elemen (element-wise multiplication).

Perkalian Matriks Skalar:

1. Perkalian matriks skalar dilakukan dengan mengalikan setiap elemen matriks dengan skalar (bilangan tunggal).

Contoh:

Misalkan kita memiliki matriks A berikut:

$$A = [[1, 2], \\ [3, 4]]$$

Dan kita ingin mengalikan matriks A dengan skalar 2:

$$2A = [[2*1, 2*2], \\ [2*3, 2*4]] = [[2, 4], \\ [6, 8]]$$

2. Perkalian matriks antar-elemen dilakukan dengan mengalikan elemen pada posisi yang sama dalam dua matriks.

Contoh:

Misalkan kita memiliki dua matriks berikut:

$$A = [[1, 2], [3, 4]]$$

$$B = [[5, 6], [7, 8]]$$

Maka, perkalian antar-elemen matriks A dan B adalah:

$$A * B = [[1*5, 2*6], \\ [3*7, 4*8]] = [[5, 12],$$

$$[21, 32]]$$

3. Pembagian Matriks dilakukan dengan membagi setiap elemen dalam matriks pertama dengan elemen pada posisi yang sama dalam matriks kedua.

Contoh:

Misalkan kita memiliki dua matriks berikut:

$$A = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}$$

$$B = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

Maka, pembagian matriks A dengan B adalah:

$$A / B = \begin{bmatrix} 6/2 & 8/2 \\ 10/2 & 12/2 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix}$$

2.3 Transformasi Linier

Transformasi linear adalah konsep penting dalam aljabar linear yang menggambarkan pemetaan linear antara dua ruang vektor. Secara sederhana, transformasi linear adalah fungsi atau pemetaan yang menjaga sifat-sifat penting dalam struktur vektor, seperti penjumlahan dan perkalian skalar. Secara formal, transformasi linear adalah pemetaan antara dua ruang vektor, katakanlah V dan W , yang memenuhi dua sifat utama:

1. **Preservasi Penjumlahan**, Untuk setiap dua vektor v_1 dan v_2 dalam ruang V , transformasi linear T menjaga penjumlahan, yaitu $T(v_1 + v_2) = T(v_1) + T(v_2)$.
2. **Preservasi Perkalian Skalar**, Untuk setiap vektor v dalam ruang V dan skalar k , transformasi linear T menjaga perkalian skalar, yaitu $T(k * v) = k * T(v)$.

Dengan kata lain, transformasi linear mempertahankan operasi penjumlahan vektor dan perkalian skalar dalam pemetaan antara ruang V dan W . Contoh transformasi linear termasuk rotasi, dilatasi, refleksi, dan proyeksi dalam geometri, serta operasi matriks seperti perkalian matriks dan matriks rotasi. Transformasi linear juga dapat direpresentasikan dengan menggunakan matriks. Dalam representasi matriks, matriks transformasi linear menggambarkan bagaimana vektor dalam ruang asal (V) ditransformasikan menjadi vektor dalam ruang target (W).

BAB III

REPRESENTASI VEKTOR DOKUMEN

3.1 Konversi dokumen menjadi vector

Untuk mengkonversi dokumen menjadi vektor, kita dapat menggunakan teknik-teknik pemodelan ruang vektor (vector space modeling) yang umum digunakan dalam pemrosesan bahasa alami (natural language processing). Salah satu teknik yang populer adalah representasi Bag-of-Words (BoW). Berikut adalah langkah-langkah umum untuk mengkonversi dokumen menjadi vektor menggunakan representasi Bag-of-Words:

1. Pra-pemrosesan teks
 - Menghilangkan karakter khusus, tanda baca, dan karakter yang tidak relevan.
 - Mengubah teks menjadi huruf kecil (lowercase) untuk menghindari perbedaan kapitalisasi kata yang serupa.
 - Menghapus kata-kata umum (stop words) seperti "dan," "atau," "di," yang tidak memberikan informasi penting.
2. Membentuk vokabular
 - Membentuk kumpulan kata unik (vokabular) dari seluruh dokumen yang akan direpresentasikan.
3. Representasi dokumen:
 - Menghitung frekuensi kemunculan setiap kata dalam setiap dokumen.
 - Membentuk vektor untuk setiap dokumen dengan memperhatikan frekuensi kata-kata dalam vokabular.
 - Misalnya, jika terdapat N kata dalam vokabular, vektor akan memiliki dimensi N. Setiap elemen vektor akan menyatakan frekuensi kemunculan kata dalam dokumen tersebut.
4. Normalisasi vektor (opsional)
 - Jika diperlukan, vektor dapat dinormalisasi dengan menggunakan teknik seperti Normalisasi L2 untuk memastikan bahwa panjang vektor sama untuk mempermudah perhitungan kesamaan kosinus (cosine similarity).

Setelah proses tersebut, setiap dokumen akan direpresentasikan sebagai vektor dalam ruang vektor yang dimensi-nya ditentukan oleh ukuran vokabular. Perlu dicatat bahwa teknik pemodelan ruang vektor seperti representasi Bag-of-Words sederhana dan tidak memperhatikan konteks atau urutan kata. Untuk representasi yang lebih canggih, teknik seperti Word Embedding (misalnya, menggunakan model Word2Vec atau GloVe) dapat digunakan.

3.2 Metode penghitungan bobot term dalam vektor dokumen

Dalam representasi vektor dokumen, terdapat beberapa metode yang digunakan untuk menghubungkan bobot term (kata) dalam vektor dokumen. Beberapa metode umum termasuk:

1. Binary Weighting, Metode ini memberikan bobot biner (1 atau 0) untuk setiap term dalam vektor dokumen. Jika term tersebut muncul dalam dokumen, bobotnya adalah 1; jika tidak, bobotnya adalah 0. Metode ini hanya memperhatikan kehadiran atau ketidakhadiran term dalam dokumen tanpa memperhitungkan frekuensi kemunculannya.
2. Term Frequency (TF), Metode ini memberikan bobot berdasarkan frekuensi kemunculan term dalam dokumen. Bobot term frequency (TF) dapat dihitung dengan rumus: $TF(\text{term}) = (\text{jumlah kemunculan term dalam dokumen}) / (\text{total jumlah term dalam dokumen})$. Bobot TF meningkat sesuai dengan jumlah kemunculan term dalam dokumen tersebut.
3. Inverse Document Frequency (IDF), Metode ini memberikan bobot berdasarkan kepentingan relatif term di seluruh koleksi dokumen. Bobot inverse document frequency (IDF) dapat dihitung dengan rumus: $IDF(\text{term}) = \log[(\text{total jumlah dokumen}) / (\text{jumlah dokumen yang mengandung term})]$. Bobot IDF memberikan penurunan bobot pada term yang muncul di banyak dokumen dalam koleksi, sehingga memberikan lebih banyak penekanan pada term yang lebih spesifik atau jarang muncul.
4. TF-IDF (Term Frequency-Inverse Document Frequency), Metode ini menggabungkan bobot TF dan IDF. Bobot TF-IDF untuk suatu term dalam dokumen dapat dihitung dengan rumus: $TF\text{-}IDF(\text{term}) = TF(\text{term}) * IDF(\text{term})$. Bobot TF-IDF memberikan penekanan pada term yang sering

muncul dalam dokumen tertentu (berdasarkan TF) dan jarang muncul dalam koleksi secara keseluruhan (berdasarkan IDF).

3.3 Penggunaan matriks term-dokumen

Penggunaan matriks term-dokumen adalah salah satu pendekatan umum dalam pemrosesan bahasa alami dan analisis teks. Matriks ini menggambarkan hubungan antara term (kata) dan dokumen dalam korpus teks. Setiap baris dalam matriks mewakili term tertentu, sedangkan setiap kolom mewakili dokumen tertentu. Nilai di dalam matriks menunjukkan bobot atau frekuensi kemunculan term tersebut dalam dokumen yang bersangkutan. Berikut adalah beberapa contoh penggunaan matriks term-dokumen:

1. **Representasi Vektor Dokumen:** Matriks term-dokumen dapat digunakan untuk mewakili setiap dokumen dalam korpus sebagai vektor berdasarkan bobot term yang terkandung di dalamnya. Setiap komponen vektor merupakan bobot atau frekuensi kemunculan term dalam dokumen. Representasi vektor ini memungkinkan perbandingan dan pengukuran kemiripan antara dokumen-dokumen menggunakan metode seperti cosine similarity atau Euclidean distance.
2. **Pencarian Informasi:** Dalam sistem temu kembali informasi, matriks term-dokumen dapat digunakan untuk mencari dokumen yang relevan dengan query pengguna. Query juga direpresentasikan sebagai vektor term yang kemudian dibandingkan dengan vektor dokumen menggunakan metode kemiripan seperti cosine similarity.
3. **Klasifikasi Dokumen:** Matriks term-dokumen juga digunakan dalam tugas klasifikasi dokumen. Setiap dokumen direpresentasikan sebagai vektor term dan dilatih menggunakan metode pembelajaran mesin seperti Naive Bayes, SVM, atau Decision Tree. Bobot term dalam matriks membantu dalam menemukan pola atau fitur-fitur yang membedakan kelas dokumen yang berbeda.

BAB IV

PENGUKURAN KEMIRIPAN DOKUMEN

4.1 Metode cosine similarity

Metode cosine similarity adalah teknik yang digunakan untuk mengukur kemiripan antara dua vektor dalam ruang vektor berdasarkan sudut kosinus antara vektor-vektor tersebut. Ini sering digunakan dalam pemrosesan bahasa alami dan sistem pengambilan informasi untuk membandingkan dokumen, teks, atau representasi vektor lainnya. Berikut adalah contoh penggunaan metode cosine similarity:

Misalkan kita memiliki dua dokumen, A dan B, yang direpresentasikan sebagai vektor dalam ruang vektor dengan bobot term menggunakan metode TF-IDF. Misalkan vektor A adalah [0.2, 0.4, 0.1, 0.3] dan vektor B adalah [0.1, 0.3, 0.5, 0.2]. Langkah-langkah untuk menghitung cosine similarity antara vektor A dan vektor B adalah sebagai berikut:

1. Normalisasi vector, Kedua vektor sudah dalam bentuk yang dinormalisasi, sehingga tidak ada langkah normalisasi yang diperlukan.
2. Perhitungan cosine similarity,

- Hitung dot product ($A \cdot B$): $\text{dot_product} = (0.2 * 0.1) + (0.4 * 0.3) + (0.1 * 0.5) + (0.3 * 0.2) = 0.02 + 0.12 + 0.05 + 0.06 = 0.25$

- Hitung panjang atau norma vektor A dan vektor B:

$$\|A\| = \sqrt{((0.2^2) + (0.4^2) + (0.1^2) + (0.3^2))} = \sqrt{(0.04 + 0.16 + 0.01 + 0.09)} = \sqrt{0.3} = 0.5477$$

$$\|B\| = \sqrt{((0.1^2) + (0.3^2) + (0.5^2) + (0.2^2))} = \sqrt{(0.01 + 0.09 + 0.25 + 0.04)} = \sqrt{0.39} = 0.6245$$

- Hitung cosine similarity, $\text{cosine_similarity} = (A \cdot B) / (\|A\| * \|B\|) = 0.25 / (0.5477 * 0.6245) = 0.25 / 0.3416 = 0.7314$

Hasil cosine similarity antara vektor A dan vektor B adalah 0.7314, yang menunjukkan tingkat kemiripan antara kedua dokumen. Dalam pemrosesan bahasa alami, metode cosine similarity sering digunakan dalam aplikasi seperti pemadanan dokumen, klasifikasi teks, dan sistem pengambilan informasi untuk mengukur sejauh mana dua dokumen atau teks mirip berdasarkan representasi vektor mereka.

4.2 Perhitungan kemiripan dengan dot product

Perhitungan kemiripan antara dua vektor menggunakan dot product (produk titik) dapat dilakukan dengan mengalikan elemen-elemen vektor tersebut dan menjumlahkannya. Dot product mengukur sejauh mana dua vektor memiliki arah yang sama. Misalkan kita memiliki dua vektor $A = [a_1, a_2, \dots, a_n]$ dan $B = [b_1, b_2, \dots, b_n]$ dengan dimensi yang sama (n elemen). Langkah-langkah untuk menghitung kemiripan antara vektor A dan B menggunakan dot product adalah sebagai berikut:

1. Hitung dot product ($A \cdot B$): $\text{dot_product} = (a_1 * b_1) + (a_2 * b_2) + \dots + (a_n * b_n)$
2. Hitung panjang atau norma vektor A dan vektor B:
 - Norma vektor A ($\|A\|$) dihitung dengan menghitung akar kuadrat dari jumlah kuadrat semua elemen vektor A: $\|A\| = \sqrt{(a_1^2) + (a_2^2) + \dots + (a_n^2)}$
 - Norma vektor B ($\|B\|$) dihitung dengan menghitung akar kuadrat dari jumlah kuadrat semua elemen vektor B: $\|B\| = \sqrt{(b_1^2) + (b_2^2) + \dots + (b_n^2)}$
3. Hitung kemiripan menggunakan dot product: Kemiripan antara vektor A dan B dapat dihitung dengan rumus: $\text{similarity} = (A \cdot B) / (\|A\| * \|B\|)$

Hasil kemiripan adalah angka antara -1 hingga 1. Nilai 1 menunjukkan bahwa vektor A dan B sepenuhnya sejajar, nilai -1 menunjukkan bahwa vektor A dan B sepenuhnya berlawanan arah, dan nilai 0 menunjukkan bahwa vektor A dan B tegak lurus atau tidak berhubungan secara linear.

4.3 Penggunaan matriks cosine similarity

Penggunaan matriks cosine similarity adalah untuk mengukur kemiripan antara setiap pasangan data dalam suatu himpunan data. Matriks cosine similarity memiliki bentuk matriks persegi, di mana setiap elemen matriks menunjukkan nilai cosine similarity antara dua data. Berikut adalah beberapa contoh penggunaan matriks cosine similarity:

1. Rekomendasi Produk: Dalam sistem rekomendasi, matriks cosine similarity dapat digunakan untuk menghitung kemiripan antara profil pengguna dan item yang tersedia. Dengan memperoleh matriks cosine similarity, kita dapat

mengidentifikasi item yang paling mirip dengan preferensi pengguna untuk memberikan rekomendasi yang lebih relevan.

2. **Klasterisasi Data:** Dalam analisis klasterisasi, matriks cosine similarity dapat digunakan untuk mengukur kemiripan antara vektor fitur atau profil objek dalam suatu dataset. Matriks cosine similarity digunakan sebagai matriks jarak untuk membangun klaster yang berdasarkan kemiripan fitur atau atribut.
3. **Pemadanan Teks:** Dalam pemrosesan bahasa alami, matriks cosine similarity sering digunakan untuk membandingkan dokumen teks. Dengan menerapkan metode vektorisasi teks seperti TF-IDF atau Word2Vec, kita dapat menghasilkan vektor representasi untuk setiap dokumen dan menghitung matriks cosine similarity untuk menemukan dokumen yang paling mirip.
4. **Identifikasi Topik:** Dalam analisis topik, matriks cosine similarity dapat digunakan untuk mengukur kemiripan antara vektor representasi topik dari dokumen. Dengan memperoleh matriks cosine similarity, kita dapat mengidentifikasi hubungan antara topik-topik tertentu dan membangun pemodelan topik yang lebih baik.
5. **Pencarian Informasi:** Dalam sistem pengambilan informasi, matriks cosine similarity dapat digunakan untuk membandingkan query pengguna dengan dokumen yang ada. Matriks cosine similarity digunakan untuk memeringkat dan memilih dokumen yang paling relevan dengan query.

BAB V

IMPLEMENTASI DAN CONTOH

5.1 Penerapan aljabar linear dalam pemrosesan bahasa alami

Aljabar linear memiliki penerapan yang luas dalam pemrosesan bahasa alami (Natural Language Processing, NLP). Beberapa penerapan utama aljabar linear dalam NLP meliputi:

1. **Representasi Vektor Kata:** Aljabar linear digunakan untuk mengonstruksi dan memanipulasi representasi vektor kata dalam pemrosesan teks. Salah satu metode yang umum digunakan adalah Word2Vec, di mana kata-kata diwakili sebagai vektor dalam ruang semantik. Aljabar linear digunakan untuk menghitung kesamaan atau perbedaan antara vektor kata dalam konteks seperti perhitungan cosine similarity, penjumlahan atau pengurangan vektor kata, dan transformasi vektor.
2. **Klasifikasi Teks:** Aljabar linear digunakan dalam pembelajaran mesin untuk klasifikasi teks. Misalnya, algoritma seperti Naive Bayes dan Support Vector Machines menggunakan representasi vektor kata dan operasi aljabar linear untuk mengklasifikasikan teks menjadi kategori yang relevan. Aljabar linear juga dapat digunakan untuk memperoleh batas keputusan yang memisahkan kelas-kelas yang berbeda dalam ruang vektor.
3. **Pemadanan Teks:** Aljabar linear digunakan dalam pemadanan atau pencocokan teks, seperti pencarian informasi atau sistem rekomendasi berbasis teks. Dalam kasus ini, aljabar linear digunakan untuk menghitung kesamaan atau jarak antara vektor representasi dokumen atau kalimat. Algoritma pemadanan seperti cosine similarity atau Jaccard similarity menggunakan operasi aljabar linear untuk mengukur kemiripan antara dokumen atau kalimat.
4. **Pengurangan Dimensi:** Aljabar linear digunakan untuk mengurangi dimensi ruang vektor teks. Teknik seperti Principal Component Analysis (PCA) atau Latent Semantic Analysis (LSA) menggunakan operasi aljabar linear seperti matriks singular value decomposition (SVD) untuk menghasilkan representasi vektor teks yang lebih ringkas dengan mempertahankan informasi penting.

5. Pemrosesan Bahasa Alami Lanjutan: Aljabar linear digunakan dalam berbagai tugas pemrosesan bahasa alami lanjutan seperti parsing, analisis sintaksis, dan pemodelan bahasa. Model seperti Recurrent Neural Networks (RNNs) dan Transformers menggunakan operasi aljabar linear dalam komputasi dan propagasi informasi melalui jaringan.

5.2 Contoh pengukuran kemiripan dokumen dengan aljabar linear

Salah satu contoh pengukuran kemiripan dokumen menggunakan aljabar linear adalah dengan menggunakan metode cosine similarity dengan representasi vektor dokumen. Berikut adalah langkah-langkah umum dalam pengukuran kemiripan dokumen menggunakan aljabar linear:

1. Pra-pemrosesan Teks: Lakukan pra-pemrosesan teks pada dokumen yang akan diukur kemiripannya. Ini termasuk langkah-langkah seperti penghapusan stop words, stemming, normalisasi teks, dan tokenisasi.
2. Ekstraksi Fitur: Ubah setiap dokumen ke dalam representasi vektor. Salah satu metode umum adalah metode Term Frequency-Inverse Document Frequency (TF-IDF), di mana setiap kata dalam dokumen dihitung frekuensinya (TF) dan dibobotkan dengan kebalikannya terhadap frekuensi kemunculan kata tersebut di seluruh dokumen (IDF).
3. Matriks Dokumen: Bentuk matriks dokumen dengan menggabungkan vektor representasi setiap dokumen. Setiap baris mewakili dokumen, dan setiap kolom mewakili fitur (kata) dalam vektor.
4. Normalisasi Vektor: Normalisasi setiap vektor dokumen agar memiliki panjang unit. Ini dilakukan dengan membagi setiap vektor dokumen dengan norma Euclidean-nya.
5. Pengukuran Kemiripan: Hitung cosine similarity antara pasangan vektor dokumen. Cosine similarity mengukur sudut antara vektor dokumen dalam ruang vektor dan memberikan nilai kemiripan antara 0 (tidak mirip) hingga 1 (sangat mirip).

Berikut adalah contoh kode Python sederhana untuk mengukur kemiripan dokumen menggunakan cosine similarity dengan menggunakan pustaka scikit-learn:

```
from sklearn.feature_extraction.text import
TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
# Contoh dokumen
documents = [
    "Dokumen A: Saya suka makan pizza",
    "Dokumen B: Pizza adalah makanan favorit saya",
    "Dokumen C: Hamburger lebih enak daripada pizza" ]
# Ekstraksi fitur dengan TF-IDF
vectorizer = TfidfVectorizer()
vectors = vectorizer.fit_transform(documents)
# Menghitung matriks kemiripan
similarity_matrix = cosine_similarity(vectors)
# Menampilkan hasil kemiripan
for i in range(len(documents)):
    for j in range(i+1, len(documents)):
        print(f"Kemiripan antara Dokumen {i} dan
        Dokumen {j}: {similarity_matrix[i, j]}")
```

Hasil output akan menampilkan kemiripan antara pasangan dokumen dalam bentuk nilai cosine similarity.

BAB VI

ANALISIS DAN PEMBAHASAN

6.1 Kelebihan dan kekurangan pendekatan menggunakan aljabar linear

Kelebihan pendekatan menggunakan aljabar linear dalam pemrosesan data, termasuk pemrosesan bahasa alami, meliputi:

1. Representasi Vektor: Aljabar linear memungkinkan representasi data dalam bentuk vektor yang dapat dengan mudah dimanipulasi dan dihitung.
2. Pengolahan dan Manipulasi Data: Aljabar linear menyediakan operasi matematis yang kuat untuk memanipulasi dan mengolah data. Misalnya, operasi penjumlahan, perkalian, pembagian, dan transformasi linear dapat digunakan untuk menggabungkan, memperoleh, dan memproses data dalam pemrosesan bahasa alami.
3. Pemodelan Kompleksitas: Aljabar linear memungkinkan pemodelan kompleksitas yang tinggi dalam pemrosesan bahasa alami. Dengan menggunakan matriks, vektor, dan operasi aljabar linear lainnya, kita dapat menggambarkan dan memodelkan hubungan kompleks antara entitas bahasa, seperti kata-kata, kalimat, dan dokumen.
4. Skalabilitas: Pendekatan menggunakan aljabar linear cenderung memiliki skala yang baik dan dapat diterapkan pada volume data yang besar. Algoritma dan teknik aljabar linear sering dapat diterapkan secara efisien pada perhitungan matriks dan vektor yang kompleks, sehingga memungkinkan pemrosesan data yang skalabel.

Namun, pendekatan menggunakan aljabar linear juga memiliki beberapa kekurangan, antara lain:

1. Kurangnya Representasi Semantik yang Kaya: Representasi vektor dalam aljabar linear mungkin tidak secara eksplisit menangkap nuansa semantik dan makna yang lebih kompleks dalam pemrosesan bahasa alami. Beberapa konsep semantik yang kompleks, seperti ambiguitas, ironi, atau konotasi, mungkin sulit direpresentasikan dengan baik dalam ruang vektor.
2. Keterbatasan dalam Memodelkan Konteks: Pendekatan aljabar linear sering kali memperlakukan kata-kata sebagai entitas independen dan mengabaikan konteks yang lebih luas. Ini dapat menjadi keterbatasan dalam pemodelan

dan memahami hubungan antara kata-kata dalam konteks yang lebih kompleks, seperti analisis sintaksis atau pemahaman bahasa alami yang lebih mendalam.

3. Keterbatasan dalam Struktur Data: Aljabar linear mungkin tidak selalu efektif dalam memodelkan struktur data yang kompleks, seperti pohon sintaksis, grafik semantik, atau jaringan. Representasi vektor dalam aljabar linear umumnya linear dan datar, sehingga tidak selalu mampu menangkap hubungan hierarkis atau jaringan kompleks dalam pemrosesan bahasa alami.

6.2 Perbandingan dengan metode lain yang digunakan dalam pengukuran kemiripan dokumen

Terdapat beberapa metode lain yang umum digunakan dalam pengukuran kemiripan dokumen. Berikut ini adalah perbandingan dengan beberapa metode tersebut:

1. Term Frequency-Inverse Document Frequency (TF-IDF): Metode TF-IDF digunakan untuk menghitung bobot kata-kata dalam dokumen berdasarkan frekuensi kemunculan kata tersebut dalam dokumen dan invers dari frekuensi kemunculan kata tersebut dalam seluruh korpus dokumen. Metode ini memperhitungkan pentingnya kata-kata yang jarang muncul dalam korpus tetapi sering muncul dalam dokumen tertentu.
2. Word2Vec: Word2Vec adalah metode yang digunakan untuk mempelajari representasi vektor kata-kata dalam ruang vektor berdasarkan konteks kata tersebut. Metode ini menghasilkan representasi yang menggambarkan relasi semantik antara kata-kata dan dapat digunakan dalam pengukuran kemiripan dokumen berdasarkan kesamaan representasi vektor kata-kata yang terkandung dalam dokumen.
3. Latent Semantic Indexing (LSI): LSI adalah metode yang digunakan untuk mengurangi dimensi dari matriks term-dokumen dengan menerapkan Singular Value Decomposition (SVD). Metode ini memungkinkan pemodelan konsep latent dalam dokumen dan dapat digunakan dalam pengukuran kemiripan dokumen berdasarkan representasi vektor yang lebih abstrak.

BAB VII

PENUTUP

7.1 Kesimpulan

Aljabar linear dapat digunakan sebagai pendekatan yang efektif dalam mengukur kemiripan antara dua dokumen. Dengan mewakili dokumen sebagai vektor dalam ruang vektor, metode aljabar linear seperti cosine similarity atau dot product dapat digunakan untuk menghitung tingkat kemiripan antara dokumen-dokumen tersebut.

Representasi vektor dokumen, seperti matriks term-dokumen atau vektor berbasis bobot seperti TF-IDF, memberikan fondasi penting dalam pengukuran kemiripan dengan aljabar linear. Representasi ini memungkinkan perhitungan yang efisien dan mempertimbangkan kemunculan term dalam dokumen-dokumen.

Penggunaan metode reduksi dimensi, seperti SVD atau PCA, dapat membantu dalam meningkatkan efisiensi dan mengurangi dimensi vektor dokumen. Ini memungkinkan pengolahan yang lebih cepat dan mengatasi masalah sparsitas dalam representasi vektor dokumen.

Pendekatan aljabar linear untuk mengukur kemiripan dokumen memiliki kelebihan, seperti kesederhanaan konseptual dan kemampuan untuk menangani dokumen-dokumen dalam skala besar. Ini juga dapat diterapkan dalam berbagai tugas pemrosesan bahasa alami, termasuk temu kembali informasi, klasifikasi dokumen, dan analisis sentimen.

Namun, pendekatan menggunakan aljabar linear juga memiliki beberapa kekurangan. Misalnya, mereka cenderung mengabaikan konteks dan urutan kata dalam dokumen, yang dapat menjadi faktor penting dalam beberapa kasus. Selain itu, mereka mungkin tidak dapat menangani secara efektif teks yang sangat kompleks, seperti dokumen yang mengandung banyak entitas atau struktur yang kompleks.

Dalam kesimpulannya, pendekatan menggunakan aljabar linear untuk mengukur kemiripan dokumen menawarkan metode yang sederhana dan efisien untuk menganalisis teks. Namun, perlu juga mempertimbangkan konteks dan

karakteristik khusus dari dokumen yang sedang diproses untuk mendapatkan hasil yang lebih akurat dan relevan.

7.2 Saran

Saran bagi pembaca terhadap makalah ini adalah untuk memahami konsep dasar aljabar linear dan representasi vektor dokumen yang digunakan, serta melihat kelebihan dan kekurangan pendekatan aljabar linear dalam konteks pengukuran kemiripan dokumen. Selain itu, pemahaman tentang tujuan penelitian dalam makalah tersebut juga penting untuk mengaplikasikan dan mempertimbangkan hasil yang diperoleh secara relevan.

DAFTAR PUSTAKA

- Aggarwal, C. C., & Zhai, C. (2012). Mining Text Data. Springer Science & Business Media.
- Dhillon, I. S., Mallela, S., & Kumar, R. (2003). A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification. *Journal of Machine Learning Research*, 3, 1265-1287.
- Ganesan, K., & Li, Y. F. (2008). Mining text data. Springer Science & Business Media.
- Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Morgan & Claypool.
- Higham, N. J. (2008). Functions of matrices: Theory and computation. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).
- Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.). Pearson.
- Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
- Strang, G. (2016). Introduction to linear algebra (5th ed.). Wellesley-Cambridge Press.