

MAKALAH
PENERAPAN ALJABAR LINEAR DALAM MENENTUKAN
KEMIRIPAN DUA BUAH DOKUMEN



DOSEN PENGAMPU:
NATALIS RANSI, S.Si., M.Cs

DISUSUN OLEH:
LA ODE MUHAMMAD YUDHY PRAYITNO
E1E122064

JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HALU OLEO
KENDARI
2023

KATA PENGANTAR

Segala puji syukur saya panjatkan atas kehadiran Allah SWT. Tuhan Yang Maha Esa. karena berkat dan rahmat-Nya lah saya dapat menyelesaikan Makalah ini dengan tepat waktu. Laporan ini disusun guna memenuhi tugas Aljabar Linear Jurusan Teknik Informatika Fakultas Teknik Universitas Halu Oleo.

Dalam penyusunan makalah ini, tentu tak lepas dari pengarahan dan bimbingan juga dukungan berbagai pihak. Maka dengan ini saya ucapkan rasa hormat dan terima kasih kepada semua pihak yang telah membantu proses pembuatan laporan hingga selesai. Pihak-pihak yang saya maksud ialah;

1. Bapak Natalis Ransi, S.Si., M.Cs. selaku dosen Aljabar Linear yang telah memberikan tugas ini.
2. Orang tua dan teman-teman saya yang ikut mendukung proses pembuatan laporan ini hingga selesai

Saya menyadari bahwa makalah ini masih memiliki banyak kekurangan. Maka dari itu. semua bentuk kritik dan saran yang membangun sangat saya harapkan dan tentu saja akan saya terima dengan senang hati. Dengan begitu, akan menjadi suatu pelajaran yang sangat berharga untuk saya supaya bisa menulis makalah yang lebih baik lagi di kemudian hari. Saya juga berharap semoga makalah “PENERAPAN ALJABAR LINEAR DALAM MENENTUKAN KEMIRIPAN 2 BUAH DOKUMEN” ini bisa bermanfaat bagi semua pihak yang membacanya.

Kendari, 25 Mei 2023

Penulis

DAFTAR ISI

| | |
|---------------------------------------------------------------------|----|
| KATA PENGANTAR | 2 |
| DAFTAR ISI..... | 3 |
| DAFTAR GAMBAR | 4 |
| BAB I | 5 |
| PENDAHULUAN | 5 |
| 1.1 Latar Belakang..... | 5 |
| 1.2 Rumusan Masalah | 6 |
| 1.3 Batasan Masalah..... | 6 |
| 1.4 Tujuan Penulis | 6 |
| 1.5 Manfaat Penulisan | 7 |
| BAB II..... | 8 |
| LANDASAN TEORI..... | 8 |
| 2.1 Vektor..... | 8 |
| 2.2 Operasi Vektor..... | 8 |
| 5.1 | 10 |
| 2.3 Ruang Vektor..... | 10 |
| 2.4 SubRuang Vektor | 11 |
| 2.5 Plagiarisme | 11 |
| 2.6 Cosine Similarity | 12 |
| BAB III | 13 |
| PEMBAHASAN | 13 |
| BAB IV | 17 |
| 4.1 Penerapan aljabar linear dalam pemrosesan bahasa alami | 17 |
| 4.2 Contoh pengukuran kemiripan dokumen dengan aljabar linear | 18 |
| BAB V..... | 20 |
| PENUTUP..... | 20 |
| 5.1 Kesimpulan..... | 20 |
| 5.2 Saran..... | 21 |
| DAFTAR PUSTAKA | 22 |

DAFTAR GAMBAR

| | |
|-------------------------------------------|----|
| Gambar 2. 1 Penjumlahan Vektor | 9 |
| Gambar 2. 2 Perkalian silang Vektor | 9 |
| Gambar 2. 3 Representasi vektor | 10 |

BAB I

PENDAHULUAN

1.1 Latar Belakang

Era globalisasi berkembang semakin pesat. Perkembangan pesat ini tentu juga terjadi pada teknologi informasi di Indonesia. Internet merupakan hal yang biasa lagi pada saat sekarang ini. Inovasi yang paling sering digunakan saat ini adalah mesin pencari (*search engine*). Situs-situs lokal maupun luar negeri banyak menyediakan fasilitas mesin pencari (*search engine*) ini seperti Google, Yahoo, Google Indonesia, dan sebagainya. Penerapan mesin pencari (*search engine*) ini sering digunakan untuk mencari (*surfing*) situs internet tertentu, mendownload data maupun gambar. Pengguna (*user*) cukup memasukkan kata kunci (*keyword*) yang diinginkan kemudian situs penyedia search engine akan bekerja dan kemudian menampilkan hasil (*result*) yang diinginkan. Dengan adanya mesin pencari seperti ini memudahkan orang untuk memperoleh informasi berupa data maupun dokumen dari internet. Perlu diketahui bahwa prinsip kerja mesin pencari (*search engine*) ini menggunakan konsep dasar aljabar linear, yaitu ruang vektor dan dekomposisi matriks ortogonal (Berry & Browne. 1992).

Prinsip kerja mesin pencari (*search engine*) tersebut ternyata dapat pula dimanfaatkan sebagai penduga terjadinya tindak plagiarisme (Rosliyanti, et al. 2006). Fakta bahwa banyaknya pengolah kata elektronik (*word processor*) semakin melengkapi fenomena tindak plagiarisme. Orang-orang dapat dengan mudah melakukan cut-paste pada suatu dokumen tertentu kemudian membuat dokumen baru dengan hanya melakukan sedikit revisi atau bahkan tidak sama sekali. Fakta inilah yang mendasari dikembangkannya suatu cara yang dapat menduga terjadinya tindak plagiarisme.

Tindakan plagiarisme merupakan tindakan merugikan baik di bidang akademik maupun bidang umum dan terapan lainnya. Jika dikaji lebih lanjut, plagiarisme adalah tindak kriminal meniru hak cipta orang lain. Hal ini tentu saja menimbulkan sejumlah kekhawatiran. Untuk membantu dan memudahkan pekerjaan tim pemeriksa untuk menduga terjadinya tindak plagiarisme maka karya tulis ini mencoba menggali lebih lanjut pemanfaatan konsep dasar aljabar linear

yang serupa dengan prinsip kerja mesin pencari (search engine) yaitu ruang vektor dan dekomposisi matriks ortogonal (Berry & Browne. 1992). Karya tulis ini juga mengembangkan hasil karya ilmiah sebelumnya yang juga mengkaji mengenai cara menduga terjadinya tindak plagiarisme dengan konsep dasar aljabar linear (Rosliyanti, et al. 2006). Pada karya ilmiah tersebut dokumen dimodelkan secara matematis sebagai vektor, pada karya tulis ini dokumen dimodelkan dalam bentuk matriks. Dengan kasus yang sama, karya tulis ini membatasi objek amatan pada dokumen berupa teks.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah di atas, maka penulis merumuskan masalahnya yaitu :

1. Bagaimana metode Vector Space Model dan matriks dalam mendeteksi adanya plagiarisme atau kemiripan 2 buah dokumen?

1.3 Batasan Masalah

Batasan Masalah Agar penelitian ini terarah dan permasalahan yang dihadapi tidak terlalu luas serta sesuai dengan tujuan penulis, maka ditetapkan Batasan batasan terhadap masalah yang sedang dikaji hal ini dimaksudkan agar langkah-langkah keseluruhan masalah tersebut tidak terjadi penyimpangan. Adapun batasan masalahnya adalah metode *vector space* hanya digunakan untuk melihat atau mengecek adanya kemiripan antara 2 buah dokumen.

1.4 Tujuan Penulis

Adapun tujuan dari pembuatan makalah “PENERAPAN ALJABAR LINEAR DALAM MENENTUKAN KEMIRIPAN 2 BUAH DOKUMEN” adalah sebagai berikut:

1. Memahami metode Vector Space Model didalam mendeteksi adanya tindakan plagiarisme.

1.5 Manfaat Penulisan

Manfaat dari pada pembuatan makalah ini baik bagi penulis maupun pembaca adalah sebagai berikut:

1. Dengan adanya sistem ini diharapkan bisa mengurangi terjadinya tindakan plagiarisme yang tentunya merugikan bagi penulis.
2. Para user seperti mahasiswa memudahkan dalam mencari sebuah dokumen dengan menggunakan query yang dimasukkan.

BAB II

LANDASAN TEORI

2.1 Vektor

Vektor adalah objek geometri yang memiliki besaran dan memiliki arah[4] . Setiap vektor dapat dinyatakan secara geometris sebagai segmen garis berarah pada bidang atau ruang. Vektor jika digambar dilambangkan dengan tanda panah (\rightarrow). Besar vektor proporsional dengan panjang panah dan arahnya bertepatan dengan arah panah. Vektor dapat melambangkan perpindahan dari titik A ke B. Vektor sering ditandai sebagai Sedangkan unsur vektor tersebut ditulis berurutan atau seperti matriks satu kolom atau memakai notasi vektor satuan \hat{i} , \hat{j} , \hat{k} .

Panjang sebuah vektor dalam ruang euklidian tiga dimensi dapat didefinisikan sebagai berikut.

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2}$$

yang merupakan konsekuensi dari Teorema Pythagoras karena vektor dasar e_1 , e_2 , e_3 merupakan vektor-vektor satuan orthogonal. Sebuah vektor yang memiliki panjang satu satuan disebut vektor satuan (unit vector). Biasanya vektor satuan digunakan untuk mendefinisikan arah. Untuk membentuk vektor satuan, bagilah vektor tersebut dengan panjang vektor tersebut.

$$\hat{a} = \frac{\vec{a}}{\|\vec{a}\|}$$

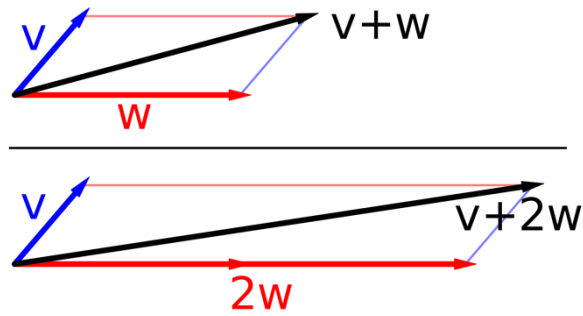
Sedangkan vektor nol (null vector) adalah suatu vektor yang panjangnya nol.

2.2 Operasi Vektor

Vektor pun dapat dikenakan operasi aljabar seperti penjumlahan, pengurangan, dan perkalian. Sebagai contoh vektor $a = a_1 \hat{i} + a_2 \hat{j} + a_3 \hat{k}$ dan $b = b_1 \hat{i} + b_2 \hat{j} + b_3 \hat{k}$. Hasil dari a ditambah b adalah:

$$a+b = (a_1+b_1)\hat{i} + (a_2+b_2)\hat{j} + (a_3+b_3)\hat{k}$$

Pengurangan vektor juga berlaku dengan cara mengganti tanda (+) menjadi tanda (-).



Gambar 2. 1 Penjumlahan Vektor

Perkalian vektor hanya dapat dilakukan jika kedua vektor berada pada ruang yang sama. Terdiri dari :

- Hasil kali titik (dot product)

Hasil kali titik akan menghasilkan besaran skalar. misal \mathbf{a} dan \mathbf{b} berada pada vector ruang yang sama, maka hasil kali titiknya didefinisikan

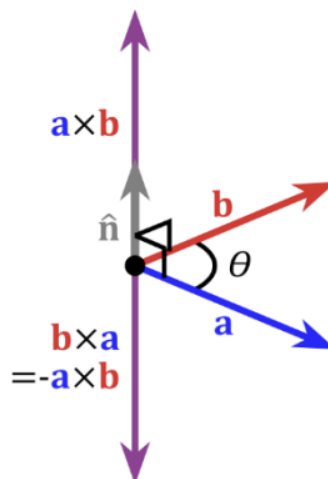
$$\bar{\mathbf{a}} \cdot \bar{\mathbf{b}} = |\bar{\mathbf{a}}| \cdot |\bar{\mathbf{b}}| \cos \alpha$$

Dimana $|\bar{\mathbf{a}}|$ dan $|\bar{\mathbf{b}}|$ masing-masing merupakan panjang vektor \mathbf{a} dan vektor \mathbf{b} Dan α adalah sudut yang dibentuk antara dua vektor tersebut.

- Hasil kali silang (cross product)

Hasil kali silang merupakan perkalian antara dua vektor yang akan menghasilkan suatu vektor baru.

$$\mathbf{a} \times \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \sin \theta$$



Gambar 2. 3 Perkalian silang Vektor

- Perkalian langsung

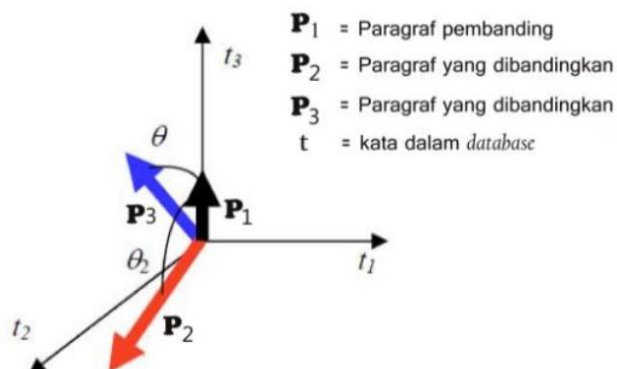
Perkalian ini tidak bersifat komutatif.

$$\begin{aligned}\vec{AB} &= (a_x \hat{i} + a_y \hat{j} + a_z \hat{k})(b_x \hat{i} + b_y \hat{j} + b_z \hat{k}) \\ &= \hat{i}(a_x b_x) \hat{i} + \hat{i}(a_x b_y) \hat{j} + \hat{i}(a_x b_z) \hat{k} \\ &\quad + \hat{j}(a_y b_x) \hat{i} + \hat{j}(a_y b_y) \hat{j} + \hat{j}(a_y b_z) \hat{k} \\ &\quad + \hat{k}(a_z b_x) \hat{i} + \hat{k}(a_z b_y) \hat{j} + \hat{k}(a_z b_z) \hat{k}\end{aligned}$$

5.1

2.3 Ruang Vektor

Ruang Vektor adalah struktur matematika yang dibentuk oleh sekumpulan vektor, yaitu objek yang dapat dijumlahkan dan dikalikan dengan suatu bilangan, yang dinamakan skalar. Contoh ruang vektor adalah Vektor Euclides yang sering digunakan untuk melambangkan besaran fisika seperti gaya. Vektor-vektor yang berada di ruang R_n dikenal sebagai vektor Eucides sedangkan ruang vektornya disebut ruang n -Euclides. Model ruang vektor merupakan teknik dasar dalam perolehan informasi yang dapat digunakan untuk penelitian relevansi dokumen terhadap kata kunci pencarian (query) pada mesin pencarian, klasifikasi dokumen, dan pengelompokan dokumen, sistem Temu-balik informasi (Information Retrieval System), dll.



Gambar 2. 4 Representasi vektor

2.4 SubRuang Vektor

Subhimpunan W dari sebuah ruang vektor V dinamakan subruang V jika W itu sendiri adalah ruang vektor yang tertutup terhadap operasi penambahan dan perkalian skalar yang didefinisikan pada V [2]. Dengan demikian, syarat agar W dikatakan sebagai sub ruang V adalah :

- $W \neq \{ \}$
- $W \subseteq V$
- Jika u dan v berada pada W , maka $u + v$ juga berada pada W
- Jika u berada di W maka ku juga berada di W , dimana k adalah suatu scalar Riil.

2.5 Plagiarisme

Menurut KBBI, Plagiarisme atau sering disebut plagiat adalah penjiplakan atau pengambilan karangan, pendapat, dan sebagainya dari orang lain dan menjadikannya seolah karangan dan pendapat sendiri. Banyak definisi dan klasifikasi yang berbeda-beda tentang plagiarisme. Beberapa contoh yang dianggap sebagai tindakan plagiarisme:

- Copy paste artikel orang lain tanpa mencantumkan referensi.
- Mengganti nama pemilik karya tulis dengan nama sendiri.
- Mengambil ide orang lain tanpa mencantumkan sumbernya.
- Mengubah karya orang lain tanpa seizing pemiliknya

Menurut Sudigdo Sastroasmoro, (2007) dalam tulisannya menyatakan bahwa jenis-jenis plagiarisme yang dapat ditemukan adalah

1. Berdasarkan aspek yang dicuri
 - Plagiarisme ide
 - Plagiarisme isi (data peneletian)
 - Plagiarisme kata, kalimat, paragraph
 - Plagiarisme total
2. Berdasarkan proporsi konten
 - Plagiarisme ringan : $< 30\%$
 - Plagiarisme sedang : $30 - 70 \%$
 - Plagiarisme berat : $> 70\%$

2.6 Cosine Similarity

Cosine Similarity digunakan untuk mengukur kesamaan antara dua buah vektor. Cosine Similarity merupakan hasil cosinus dari sudut diantara kedua vektor. Dapat dirumuskan sebagai berikut.

$$sim(Q, D_1) = \cos \theta = \frac{Q \cdot D}{\|Q\| \|D\|} = \frac{\sum_{i=1}^n Q_i D_i}{\sqrt{\sum_{i=1}^n Q_i^2} \sqrt{\sum_{i=1}^n D_i^2}}$$

Keterangan:

Q = Query dokumen

D1 = Dokumen uji.

BAB III

PEMBAHASAN

Salah satu penerapan konsep aljabar linear dan vektor adalah model ruang vektor. Model ruang vektor ini awalnya diaplikasikan pada Sitem Temu-balik Informasi (Information Retrieval System). Dalam hal ini dokumen yang akan diperiksa dikonversi dahulu menjadi vektor berukuran $\mathbf{n} \times 1$, dimana n adalah banyaknya kata berbeda yang ada dalam dokumen sebagai kamus kata (vocabulary) atau indeks kata (term index). Kata-kata tersebut akan membentuk ruang vektor berdimensi \mathbf{n} . Setiap dokumen maupun query dinyatakan sebagai vektor $\mathbf{w} = (w_1, w_2, w_3, \dots, w_n)$ di dalam \mathbf{R}^n . Nilai w_i menyatakan jumlah kemunculan setiap kata i di dalam query atau dokumen (term frequency). Berikut adalah contoh representasi dokumen sebagai vector.

AKU CINTA BANDUNG DAN AKU ADALAH MAHASISWA INSTITUT
TEKNOLOGI BANDUNG YANG CINTA TEKNOLOGI.

| Kata | Frekuensi |
|-----------|-----------|
| Aku | 2 |
| Cinta | 2 |
| Bandung | 2 |
| Dan | 1 |
| Adalah | 1 |
| Mahasiswa | 1 |
| Institut | 1 |
| Teknologi | 2 |
| Yang | 1 |

Secara Matematis:

$$Q = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 1 \end{pmatrix}$$

Banyaknya kata berbeda ada 9, artinya, representasi kalimat judul sebagai vektor misal j dengan ukuran (9×1) merupakan subruang di R^9 .

Lalu kita memerlukan dokumen uji disini. Kita akan menghitung sudut antara dua dokumen. Ukuran vektor haruslah sama. Pengukuran sudut antar dua buah dokumen diidentikan dengan pengukuran sudut antara dua buah vektor yaitu dengan rumus.

$$\cos \theta = \frac{Q \cdot D}{\|Q\| \|D\|}$$

Dimana

- θ adalah sudut yang diapit oleh kedua vektor
- $Q \cdot D$ adalah hasil perkalian titik dari vektor Q dan vektor D yang didefinisikan sebagai $Q \cdot D = Q_1D_1 + Q_2D_2 + \dots + Q_nD_n$.
- $\|Q\| \|D\|$ masing-masing adalah panjang *Euclidean (norm)* dari vektor Q dan D . Panjang *Euclidean (norm)* diperoleh dari akar penjumlahan kuadrat elemen-elemen vektor tersebut.

$$\|D\| = \sqrt{a^2 + b^2}$$

Lalu kita misalkan ada dokumen dua buah dokumen uji untuk dibandingkan kesamaannya dengan dokumen utama. Untuk saat ini dibatasi kedua dokumen uji harus merepresentasikan vektor yang ukurannya sama.

$$Q = \begin{pmatrix} 1 \\ 2 \\ 5 \\ 3 \\ 1 \\ 2 \\ 3 \\ 2 \\ 1 \end{pmatrix} \quad Q = \begin{pmatrix} 0 \\ 5 \\ 3 \\ 2 \\ 0 \\ 0 \\ 1 \\ 1 \\ 5 \end{pmatrix}$$

D1 = (1,2,5,3,1,2,3,2,1) artinya dalam dokumen pertama terdapat kata “Aku” sebanyak 1 buah, 2 buah kata “Cinta”, 5 buah kata “Bandung”, dan seterusnya. Begitu juga dengan dokumen kedua tidak ada kata “Aku”, ada 5 buah kata “cinta” dan seterusnya. Kemudian dugaan plagiarisme dilihat dari besar sudut θ yaitu sudut antara dokumen yang ingin di cek dengan dokumen uji. Jika $\cos \theta = 1$ berarti $\theta = 0$, vektor Q dan D berhimpit, yang berarti dokumen D sama dengan dokumen Q, atau dengan kata lain, menjiplak. Jika sudut yang dibentuk semakin kecil, maka mengindikasikan telah terjadi plagiarisme. Walaupun begitu, tetap dibutuhkan pemeriksaan lebih lanjut setelah pendugaan ini.

Pada contoh diatas,

$$\begin{aligned} Q \cdot D1 &= (2)(1) + (2)(2) + (2)(5) + (1)(3) + (1)(1) + (1)(2) \\ &\quad + (1)(3) + (2)(2) + (1)(1) \\ &= 2 + 4 + 10 + 3 + 1 + 2 + 3 + 4 + 1 \\ &= 30 \end{aligned}$$

$$\begin{aligned} Q \cdot D2 &= (2)(0) + (2)(5) + (2)(3) + (1)(2) + (1)(0) + (1)(0) \\ &\quad + (1)(1) + (2)(1) + (1)(5) \\ &= 0 + 10 + 6 + 2 + 0 + 0 + 1 + 2 + 5 \\ &= 26 \end{aligned}$$

$$\begin{aligned} \|Q\| &= \sqrt{2^2 + 2^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2 + 2^2 + 1^2} \\ &= \sqrt{21} \end{aligned}$$

$$\begin{aligned} \|D1\| &= \sqrt{1^2 + 2^2 + 5^2 + 3^2 + 1^2 + 2^2 + 3^2 + 2^2 + 1^2} \\ &= \sqrt{58} \end{aligned}$$

$$\begin{aligned} \|D2\| &= \sqrt{0^2 + 5^2 + 3^2 + 2^2 + 0^2 + 0^2 + 1^2 + 1^2 + 5^2} \\ &= \sqrt{65} \end{aligned}$$

$$\begin{aligned}
 \text{sim}(Q, D_1) &= \cos \theta_1 = \frac{Q \cdot D_1}{\|Q\| \|D_1\|} = \frac{30}{\sqrt{21} \sqrt{58}} \\
 &= \frac{30}{\sqrt{1218}} = 0.8596 \\
 \theta_1 &= 30.7283^\circ
 \end{aligned}$$

$$\begin{aligned}
 \text{sim}(Q, D_2) &= \cos \theta_2 = \frac{Q \cdot D_2}{\|Q\| \|D_2\|} = \frac{26}{\sqrt{21} \sqrt{65}} \\
 &= \frac{26}{\sqrt{1365}} = 0.703732 \\
 \theta_2 &= 45.2728^\circ
 \end{aligned}$$

Dari dua hasil diatas, dokumen query memiliki beda sudut yang dibentuk dengan dokumen 1 sebesar 30.7283° , sedangkan dengan dokumen 2 memiliki beda sudut 45.2728° . Juka sudut yang dibentuk antar Subruang memiliki nilai yang kecil, berarti kedua dokumen memiliki kemiripan yang tinggi. Karena nilai sudut yang terbentuk antara dokumen Query dan dokumen pertama lebih kecil daripada sudut antara dokumen Query dengan dokumen kedua, maka dapat dikatakan dokumen query lebih mirip dengan dokumen pertama.

BAB IV

IMPLEMENTASI DAN CONTOH

4.1 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.

4.2 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 V

PENUTUP

5.1 Kesimpulan

Konsep aljabar linear dapat dijadikan dasar untuk pencocokan dua buah dokumen. Pencocokan ini diaplikasikan pada system temu-balik informasi (information retrieval system) dan dapat juga digunakan untuk melakukan pengecekan awal untuk plagiarisme. Pengecekan ditinjau dari jumlah kata yang direpresentasikan sebagai subruang di R^n . Vektor ini nantinya akan diukur sudutnya dengan dokumen uji. Pasangan dokumen yang sudutnya kurang dari sudut batas, diduga memiliki kecenderungan terjadi plagiarisme. Jika θ yang dibentuk $= 0$ maka kedua dokumen dikatakan sama atau identik. Jika $\theta = \pi/2$, maka dua dokumen dikatakan berbeda. Tetapi hal ini masih belum bisa dikatakan pasti. Masih banyak faktor lain yang tidak diperhitungkan selain yang diatas. Masih banyak hal yang seharusnya diperhitungkan atau bahkan harus dibuang tetapi masih diperhitungkan. Seperti contohnya kita harusnya menggunakan stopword yaitu kata-kata yang tidak deskriptif yang dapat dibuang seperti “di”, “dan”, “yang” dan lainnya, juga teknik stemming yaitu teknik yang dilakukan pada perolehan informasi untuk menghilangkan variasi morfologi atau dengan kata lain membuang imbuhan. Selain itu, untuk memastikan terjadinya tindak plagiarisme, kita juga harus melakukan uji empiris, berupa pemeriksaan manual. Diharapkan kedepannya, pemodelan ini akan bisa menghasilkan kinerja yang lebih baik, lebih efisien, lebih akurat dari yang sudah ada.

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.

5.2 Saran

Saran bagi pembaca terhadap makalah ini adalah untuk membaca dengan cermat dan menjaga konsentrasi. Aljabar linear dalam pengolahan citra digital merupakan topik yang kompleks, oleh karena itu, penting bagi pembaca untuk memahami setiap konsep dan aplikasi yang dijelaskan dalam makalah. Dalam membaca makalah ini, disarankan untuk membaca dengan teliti agar tidak melewatkan informasi penting. Jika pembaca mengalami kesulitan dalam memahami beberapa bagian, disarankan untuk membaca ulang atau mencari referensi tambahan untuk mendapatkan pemahaman yang lebih baik.

DAFTAR PUSTAKA

- Aggarwal, C. C., & Zhai, C. (2012). Mining Text Data. Springer Science & Business Media.
- Leman, Dedi, and Khusaeri Andesa. 2015. "IMPLEMENTASI VECTOR SPACE MODEL UNTUK MENINGKATKAN KUALITAS PADA SISTEM PENCARIAN BUKU PERPUSTAKAAN." *Seminar Nasional Informatika 2015* 8-15. Diakses 22 Mei 2023
- Sentosa, Johan. 2015. "Aplikasi Model Ruang Vektor dan Matriks untuk Mendeteksi Adanya Plagiarisme." *INFORMATIKA STEI ITB* 1-5. Diakses 22 Mei 2023
- Strang, G. (2016). Introduction to linear algebra (5th ed.). Wellesley-Cambridge Press.