PEMBUATAN APLIKASI PENENTUAN TINGKAT KEMIRIPAN ANTAR DOKUMEN TEKS MENGGUNAKAN METODE VECTOR SPACE MODEL

Alfine Candra Cuaca [1] **Lely Hiryanto** [2] **Tri Sutrisno**[3]

[1] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Grogol Petamburan, Jakarta Barat 11440 Indonesia
email : alfinecandracuaca@gmail.com
[2] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Grogol Petamburan, Jakarta Barat 11440 Indonesia
email : lelyh@fti.untar.ac.id
[2] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Grogol Petamburan, Jakarta Barat 11440 Indonesia
email : tris@fti.untar.ac.id

ABSTRACT

The application of similarity detection between text documents created using Vector Space Model (VSM) method gives result of similarity degree values and percentage of similarity between comparison document and the test documents. The process of this method, first calculates the word weight with pre-processing that is using Term Frequency – Inverse Document Frequency (TF-IDF), then calculates the dot product between comparison documents with each test document and dot product document with the document itself, then to calculate the angle using cosine similarity to get the similarity degree values between the comparison documents with each test document.

The test is done by using data from the students of Faculty of Information Technology, Tarumanagara University amounted to 21 people which divided into 4 sections, namely chapter I, chapter III, chapter IV, and chapter V. The test results showed that the application can operate the VSM method well and in the testing process, VSM method is not affected by the length of the documents and also not affected by the word order.

Key words

Cosine Similarity, Dot Product, Term Frequency – Inverse Document Frequency, Vector Space Model

1 Pendahuluan

Plagiarisme atau plagiat adalah penjiplakan atau pengambilan karangan, pendapat, dan sebagainya dari orang lain dan menjadikannya seolah menjadi karangan dan pendapat sendiri. [1] Tindakan ini dapat dikategorikan "mencuri", maka dari itu praktik plagiarisme tentu harus dihindarkan.

Untuk mengurangi tindak plagiarisme ada dua metode yang dapat diterapkan, yaitu mencegah plagiarisme dan mendeteksi plagiarisme. Mencegah plagiarisme berarti melakukan tindakan pencegahan agar tindakan plagiarisme tidak terjadi, misalnya dengan menetapkan sebuah kebijakan tentang tindakan plagiarisme atau sistek hukuman bila terbukti melakukan tindak plagiarisme. Sedangkan dalam mendeteksi plagiarisme berarti menemukan adanya tindakan plagiarisme yang telah dilakukan.

Mendeteksi plagiarisme hanya dapat mengurangi tindakan plagiarisme yang telah terjadi, tetapi mencegah plagiarisme tentu dapat mengurangi angka dari tindakan plagiarisme. Tetapi kenyataannya, tindakan plagiarisme sudah melekat pada moral masyarakat secara luas yang tidak dapat diselesaikan hanya dengan upaya Universitas atau Departemen.[2] Oleh karena itu, upaya yang mungkin untuk dilakukan adalah mendeteksi dari plagiarisme tersebut.

Dalam melakukan pendeteksi dari plagiarisme tersebut dapat secara manual atau otomatis dari bantuan komputer. Saat ini pendeteksian secara manual merupakan cara yang paling akurat dalam mendeteksi plagiat. Kelemahannya adalah sangat menghabiskan tenaga, waktu, serta tidak konsisten karena dipengaruhi oleh faktor emosional dari manusia. Oleh karena itu, sampai saat ini para akademisi berusaha mengembangkan sebuah sistem komputer untuk mendeteksi plagiat dengan tindak akurasi yang mendekati sistem manual. [3]

Tujuan yang ingin dicapai dari penelitian ini adalah untuk mendeteksi tingkat kesamaan teks dalam dokumen atau *file* sebagai data uji dengan dokumen pembanding dengan menggunakan metode *Vector Space Model* (VSM). Metode ini menggunakan pemanfaatan konsep aljabar linear yang serupa dengan prinsip kerja mesin pencarian yaitu ruang vektor.

2 Dasar Teori

2.1 Vektor

Vektor adalah objek geometri yang memiliki besaran dan memiliki arah. Setiap vektor dapat dinyatakan secara geometris sebagai segmen garis berarah pada bidang atau ruang. Vektor jika digambar dilambangkan dengan tanda panah (→). Besar vektor proporsional dengan panjang panah dan arahnya bertepatan dengan arah panah. Vektor dapat melambangkan perpindahan dari titik *A* ke titik *B*.[4]

Vektor memiliki sifat-sifat sebagai berikut[5]:

1. Vektor dikatakan sama jika memiliki besar dan arah yang sama.
2. Vektor harus memiliki unit yang sama agar dapat dijumlahkan atau dikurangan.
3. Negatif dari suatu vektor memiliki besar yang sama namun berlawanan arah.
4. Pengurangan vektor dapat dilakukan dengan menjumlahkan dengan vektor negatif.
5. Perkalian atau pembagian vektor dengan skalar akan menghasilkan vektor.
6. Proyeksi dari suatu vektor di sepanjang sumbu koordinat disebut sebagai komponen vektor.
7. Menjumlahkan vektor dilakukan dengan menjumlahkan komponen-komponen yang bersesuaian.

2.2 Operasi Vektor

Vektor pun dapat dikenakan operasi aljabar seperti penjumlahan, pengurangan, dan perkalian. Perkalian vektor hanya dapat dilakukan jika kedua vektor berada pada ruang yang sama, yang terdiri dari:

- Hasil kali titik (*dot product*)
  Hasil kali titik akan menghasilkan besaran skalar. Misalnya *a* dan *b* berada pada vektor ruang yang sama, maka hasil kali titiknya akan didefinisikan sebagai berikut[6]:

$$\bar{a} \cdot \bar{b} = \|\bar{a}\| \, \|\bar{b}\| \cos \propto$$

....(1)

Dimana $\|\bar{a}\|$ dan $\|\bar{b}\|$ masing – masing merupakan panjang vektor *a* dan *b*. Dan $\propto$ adalah sudut yang dibentuk antara dua vektor tersebut.

2.3 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[4]:

1. *Copy paste* artikel orang lain tanpa mencantumkan referensi.
2. Mengganti nama pemilik karya tulis dengan nama sendiri.
3. Mengambil ide orang lain tanpa mencantumkan sumbernya.
4. Mengubah karya orang lain tanpa seizin pemiliknya.

Menurut Sudigdo Sastroasmoro, (2007) dalam tulisannya menyatakan bahwa jenis-jenis plagiarisme yang dapat ditemukan adalah[7]:

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

2.4 *Cosine Similarity*

Cosine Similarity digunakan untuk melihat kemiripan antar dokumen teks. Kemiripan dalam VSM ini ditemukan oleh vektor dari dokumen pembanding dan vektor dari dokumen uji. [4] *Cosine Similarity* akan menghasilkan sebuah matriks yang saling berelasi antara dokumen-dokumen dengan melihat besar sudutnya. Cosinus sering digunakan untuk membandingkan dokumen – dokumen. Apabila nilai cos dari sudutnya adalah 1 maka kemiripan teks adalah 100%, sedangkan apabila nilai cos dari sudutnya adalah 0 maka kemiripan teks adalah 0% atau dapat dikatakan orisinil.[8] Dapat dirumuskan sebagai berikut[6]:

$$\cos\theta = \frac{D_1 \cdot D_2}{\|D_1\| \cdot \|D_2\|}$$

....(2)

Keterangan:
D_1 = dokumen pembanding
D_2 = dokumen uji

3. Hasil Percobaan

Anggap ada 3 contoh, 1 sebagai dokumen pembanding dan 2 sebagai dokumen uji.

- Dokumen A: “Saya makan nasi”
- Dokumen B: “Saya makan ikan dan sayur”
- Dokumen C: “Saya tidak suka nasi”

3.1 Tokenizing / Tokenisasi

Pertama-tama dokumen akan mengalami proses tokenisasi. Tokenisasi adalah proses membagi teks menjadi bagian-bagian yang lebih kecil yang disebut *token*. Proses tokenisasi umumnya memisahkan kata-kata dari spasi dan juga semua huruf akan dibuat menjadi huruf kecil semuanya. Ini terlihat efisien, karena kalimat umumnya terdiri dari spasi, sehingga dapat dipisahkan menjadi *token*.[9] Jadi, setelah melalui tahap tokenisasi,

- Dokumen A: "saya makan nasi"
- Dokumen B: "saya makan ikan dan sayur"
- Dokumen C: "saya tidak suka nasi"

3.2 *Stopword Removal*

Stopword adalah kata-kata fungsional khusus yang tidak membawa informasi apapun, yaitu contohnya kata ganti, kata depan, dan kata hubung. Daftar *stopword* dalam bahasa Indonesia berkisar lebih dari 500 kata. *Stopword* banyak digunakan di kalimat atau paragraf, sehingga sangat penting dalam melakukan pendeteksian tingkat kemiripan.[10] Dokumen-dokumen tersebut setelah melalui tahap *stopword removal* akan menjadi:

- Dokumen A: makan nasi
- Dokumen B: makan ikan sayur
- Dokumen C: suka nasi

3.3 *Pre-processing*

Tahap *pre-processing* terdiri dari 4 tahap yaitu *Term Frequency* (TF), *Document Frequency* (DF), *Inverse Document Frequency* (IDF), dan *Term Frequency – Inverse Document Frequency* (TF-IDF).

Term Frequency (TF) adalah jumlah frekuensi suatu kata pada dokumen. Jika ingin melambangkan jumlah frekuensi kata *t* pada dokumen *d* dengan $f_{t,d}$, maka TF suatu kata adalah[6]:

$$tf(t,d) = f_{t,d}$$

....(3)

Keterangan:

$f_{t,d}$: jumlah frekuensi kata pada dokumen

Nantinya dokumen dokumen tersebut akan menjadi pada **Tabel 1** seperti berikut:

Tabel 1 *Term Frequency* (TF)

Kata	Dokumen A	Dokumen B	Dokumen C
makan	1	1	
nasi	1		1
ikan		1	
sayur		1	
suka			1

Document Frequency (DF) adalah jumlah dokumen yang memuat *term* atau kata tersebubut. Hasil dari DF bisa dilihat pada **Tabel 2**. Dokumen pembanding tidak dimasukkan dalam hitungan DF. Sehingga kata makan yang ada di dokumen A dan dokumen B, akan hanya dianggap ada di dokumen B saja.

Tabel 2 *Document Frequency* (DF)

Kata	DF
makan	1
nasi	1
ikan	1
sayur	1
suka	1

Inverse Document Frequency (IDF) adalah pengukuran seberapa penting informasi yang terdapat pada suatu kata. IDF berfungsi untuk mengurangi bobot suatu *term* atau kata jika kemunculannya banyak tersebar di seluruh koleksi dokumen. Hampir sama dengan konsep DF, apabila ada tiga dokumen, maka jumlah *N* adalah 3 (termasuk dokumen pembanding). Jika jumlah seluruh dokumen adalah *N*, dan jumlah dokumen yang memuat kata adalah *t* adalah df_t, maka IDF dari suatu kata adalah[6]:

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right)$$

....(4)

Keterangan:

idf_t: *Inverse Document Frequency* suatu kata

N: jumlah seluruh dokumen

df_t: jumlah dokumen yang memuat suatu kata

Tabel 3 *Inverse Document Frequency* (IDF)

Kata	DF
makan	0.301
nasi	0.301
ikan	0.301
sayur	0.301
suka	0.301

Term Frequency – Inverse Document Frequency (TF-IDF) adalah sebuah perhitungan statistik yang bertujuan untuk menggambarkan seberapa penting sebuah kata terhadap sebuah koleksi dokumen. TF-IDF sering digunakan sebagai pembobotan dalam *Information Retrieval* (IR) dan *text mining*. Nilai TF-IDF meningkat seiring dengan banyaknya jumlah kata tersebut muncul di dalam dokumen, tetapi diimbangi dengan jumlah dokumen yang memuat kata tersebut. Semakin banyak dokumen yang memuat kata tersebut, nilai TF-IDF kata tersebut akan semakin kecil. Nilai TF-IDF suatu kata pada dokumen dihitung dengan cara mengalikan TF suatu kata pada dokumen dengan IDF kata tersebut, sehingga dapat ditulis sebagai berikut[6]:

$$tfidf(t,d) = tf(t,d) * idf_t$$

....(5)

Keterangan:
$tfidf(t,d)$: nilai bobot TF-IDF suatu kata pada dokumen
$tf(t,d)$: frekuensi kemunculan kata pada dokumen
idf_t: nilai IDF dari suatu kata

Hasil dari nilai pembobotan TF-IDF dapat dilihat pada **Tabel 4.**

Tabel 4 TF-IDF

Kata	Dokumen A	Dokumen B	Dokumen C
makan	0.301	0.301	0
nasi	0.301	0	0.301
ikan	0	0.301	0
sayur	0	0.301	0
suka	0	0	0.301

Lalu setelah mendapatkan nilai bobot TF-IDF masing-masing kata, maka selanjutnya akan dihitung *dot product* dan selanjutnya *cosine similarity* untuk melihat derajat kemiripan antar dokumen teks. Nilai dari derajat kemiripan dapat dilihat pada **Tabel 5.**

Tabel 5 Hasil perhitungan

Dok. Pembanding	Dok. Uji	Cosine Similarity	Derajat
Dokumen A	Dokumen B	0.408	65.905
Dokumen A	Dokumen C	0.5	60

Dengan menghitung derajat kemiripan, makan bisa dihitung juga persentase kemiripan dengan membagi total jumlah kata yang serupa dengan total keseluruhan kata pada dokumen. Hasil perhitungan dapat dilihat pada **Tabel 6.**

Tabel 6 Hasil Persentase Kemiripan

Dok. Pembanding	Dok. Uji	Derajat	Persentase
Dokumen A	Dokumen B	65.905	33.33%
Dokumen A	Dokumen C	60	50%

4. Kesimpulan

Kesimpulan yang dapat diperoleh berdasarkan pembuatan dan pengujian dari aplikasi ini adalah sebagai berikut:

1. Aplikasi pendeteksi tingkat kemiripan antar dokumen teks ini dapat berjalan dan mengoperasikan metode *Vector Space Model* (VSM) dengan baik, juga dapat memberikan hasil derajat kemiripan tiap dokumen pembanding dengan dokumen uji masing-masing.
2. Berdasarkan hasil pengujian, derajat kemiripan mungkin bisa sangat berbeda dengan persentase kemiripan. Hal ini dikarenakan bahwa dengan metode VSM yang menghitung *dot product* dari masing-masing dokumen uji dengan dokumen pembanding dan juga menghitung panjang vektor masing-masing dokumen.
3. Hasil dari derajat kemiripan lebih efektif dibandingkan persentase kemiripan, karena persentase kemiripan hanya menghitung kata-kata yang mirip yang dibagi dengan jumlah total kata di dokumen.
4. Berdasarkan hasil pengujian diketahui bahwa metode VSM akan terasa buruk jika digunakan pada dokumen yang sangat panjang karena akan menghasilkan vektor dengan dimensi yang besar.

Saran untuk yang ingin mengembangkan aplikasi pendeteksi tingkat kemiripan antar dokumen teks dengan metode VSM adalah sebagai berikut:

1. Aplikasi dapat dikembangkan agar dapat melihat kata-kata yang sama persis dengan dokumen pembanding.
2. Aplikasi dapat untuk tidak memproses kutipan, referensi, atau catatan kaki yang sudah dicantumkan.

REFERENSI

[1] Departemen Pendidikan Nasional, "Kamus Besar Bahasa Indonesia Daring", http://badanbahasa.kemdikbud.go.id, 11 September 2017.

[2] Romans, Lukashenko; Graudina V dan Grundspenskis J., 2017, "Computer-based Plagiarism Detection Methods and Tools: An Overview", Proceedings of the 2007 International Conference on Computer Systems and Technologies, New York, NY, USA.

[3] Mahathir, Fakhri, 2011, "Sistem Pendeteksi pada Dokumen Teks Berbahasa Indonesia Menggunakan Metode ROUGE-N, ROUGE-L, dan ROUGE-W", Bogor: Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor.

[4] Sentosa, Johan, 2015, "Aplikasi Model Ruang Vektor dan Matriks untuk Mendeteksi Adanya Plagiarisme", Bandung: Institut Teknologi Bandung

[5] Unwinnipeg, "Scalars and Vectors", http://theory.uwinnipeg.ca/physics/twodim/node2.html., 29 Desember 2017.

[6] Manning, Christopher D; Prabhakar Raghavan dan Hinrich Schutze, 2008, "Introduction to Information Retrieval", Cambridge: Cambridge University Press.

[7] Sudigdo, Sastroasmoro, 2007, "Beberapa Catatan Tentang Plagiarisme, Majalah Kedokteran Indonesia"

[8] Perone, Christian S., 2013, "Machine Learning::Cosine Similarity for Vector Space Model (Part III)", http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/, 26 Desember 2017.

[9] Turney, Peter D., Patrick Pantel, 2010, "From Frequency to Meaning: Vector Space Model of Semantics", Journal of Artificial Intelligence Research.

[10] Sharma, Dharmendra; Suresh Jain, 2015, "Evaluation of Stemming and Stopword Techniques on Text Classification Problem", Journal of Scientific Research in Computer Science and Engineering, Vol. 3, No. 2

[11] Mao, Yaobin., Chen, Guanrong., 2003, "Chaos-Based Image Encryption", Handbook of Computational

Geometry for Pattern Recognition, Computer Vision, Neural Computing and Robotics, Springer, Berlin.

Alfine Candra Cuaca, seorang mahasiswa pada program studi Fakultas Teknologi Informasi di Universitas Tarumagara

Lely Hiryanto, memperoleh gelar ST dari Universitas Tarumanagara, Indonesia tahun 2001. Kemudian tahun 2006 memperoleh M.Sc. dari Curtin University of Technology, Australia. Saat ini sebagai Staf Pengajar program studi Teknik Informatika, Universitas Tarumanagara

Tri Sutrisno, memperoleh gelar S.Si dari Universitas Dipenogoro, Indonesia tahun 2011. Kemudian tahun 2015 memperoleh M.Sc. dari Universitas Gadjah Mada, Indonesia. Saat ini sebagai Staf Pengajar program studi Teknik Informatika, Universitas Tarumanagara