JuTech
Jurnal Ilmiah Teknologi & Informasi

Website: http://ojs.itb-ad.ac.id/index.php/JUTECH
E-ISSN: 2797 - 4111

MENGIPLEMENTASIKAN VECTOR SPACE MODEL SIMILARITY EUCLIDEAN DISTANCE MENGGUNAKAN TFIDF PADA KLASIFIKASI TEKS BAHASA INDONESIA

Reza Fitriansyah[1], Ellya Sestri[2], Vany Terisia[3] (*)

[1]*ITB Ahmad Dahlan, Jakarta*
[2]*ITB Ahmad Dahlan, Jakarta*
[3]*ITB Ahmad Dahlan, Jakarta*

Abstract

Weighting based on the term with stemming techniques to get the basic word form term in question. This will the application of the Indonesian language text classification machine using the K-Nearest Neighbor algorithm and the Vector Space Model method on the TF-IDF frequency weighting of the number of words and the Euclidean Distance function. comparison between the test documents and the test sample collection Using news documents as learning documents, a total of 10 (10) documents with 3 (three) categories, produces an Precision and Recall 90.00% for k = 5 using frequency weighting in words with the Euclidean Distance function.

Kata Kunci: K-Nearest Neighbor algorithm, Vector Space Model, Euclidean Distance, Precision and Recall

Juli - Desember 2022, Vol 3 (2) : hlm 158-163
©2022 Institut Teknologi dan Bisnis Ahmad Dahlan.
All rights reserved.

(*) *Korespondensi: Rezafsyh@gmail.com (Reza Fitriansyah), ellyasestri.24@gmail.com (Ellya Sestri), vterisia@gmail.com (Vany Terisia)*

PENDAHULUAN

Dalam penelitian kali ini bertujuan untuk mengenal karakteristik notasi pembobotan *TFIDF* serta mengimplementasikan model ruang vektor menggunakan notasi pada metode pembobotan *TFIDF* pada sistem *Information Retrieval (IR)*. Fungsi yang digunakan untuk pembobotan *TFIDF* pada penelitian ini adalah fungsi Euclidean Distance.

METODE

Metode yang digunakan untuk menguji klasifikasian dokumen *classifier* dengan Euclidean Distance. Instrumentasi pada penelitian yakni peneliti sendiri, dengan melakukan sebagai berikut: Identifikasi sebuh sistem yang telah digunakan saat ini, dalam segi analisa relevansi untuk pencarian dokumen maka belum ada yang menganalisis tingkat ketepatan dan kelengkapan antara *vector Space model TFIDF* dengan fungsi Euclidean Distance. Penulis mengambil sample dokmen *PDF* untuk di analisa nilai pembobotannya sehingga dapat di lihat tingkat *prcision* dan *recall* dokumennya.

Pengujian Data

Pengujian Ketepatan

Precision adalah sebuah proses hasil yang tepat dari sebuah dokumen relevan pada sebuah data yang dicocokan. Cara menghitung *precision* adalah :

$$Precision = \frac{Jumlah\ dokumen\ relevan\ yang\ terambil}{Jumlah\ dokumen\ yang\ terambil\ dalam\ pencarian}$$

Pengujian Kelengkapan

Recall adalah sebuah proses menggali sebuah data terbaca. Cara mencari *recall* adalah

$$Recall = \frac{Jumlah\ dokumen\ relevan\ yang\ terambil}{Jumlah\ dokumen\ relevan\ yang\ ada\ dalam\ database}$$

HASIL DAN PEMBAHASAN

Analisis Koleksi Dokumen

Menguji data yang sudah di siapkan berbentuk format file *pdf* data set/eksperiment, Di uji dari 10 dipisah menjadi dua data,dari sebuah data terdiri dari 5 dokumen, yaitu politik, teknologi, Dokumen bersumber dari koran online, lalu data yang sudah disapkan berdasarkan kategori.

Tabel I Dokumen Uji

No.	Judul	Kategori	Sumber
1	Di Balik Adu Surat Firli-Listyo	Politik	Tempo.com
2	Era Gelap Kebebasan Sipil	Politik	Tempo.com
3	Lalai Berulang Di Kilang Pertamina	Politik	Tempo.com
4	Tabrak Paksa Pengusutan Formula E	Politik	Tempo.com
5	Panas Dining KPK-MABe Polri	Politik	Tempo.com
6	Diakui mereka kesulitan dalam pelaksanaan ibadah di luar angkasa	Teknologi	CNN.com
7	Survei Telkomsel: 41 Persen Warga Pilih Pulang Cepat saat Ramadhan	Teknologi	CNN.com
8	Wajah Empat Astronaut NASA untuk Misi Artemis II ke Bulan	Teknologi	CNN.com
9	Rekor Temuan Ikan di Lautan Terdalam Pecah, Cek Caranya Bertahan Hidup	Teknologi	CNN.com
10	NASA Umumkan 4 Astronaut untuk Misi Artemis II, Salah Satunya Wanita	Teknologi	CNN.com

Di lihat dari table diatas, maka kategori dari dokumen uji coba yang terdiri dari 10 dokumen yang dapat dari dua koran online, diantaranya : tempo.com, CNN.com. maka memperoleh data, yang terbagi dari koran online dapat di lihat pada tabel berikut.

Tabel II Dokumen Eksperimen

No	Sumber	Jumlah Dokumen
1	Tempo.com	5
2	CNN.com	5

Prediksi Rangking K-Nearest Neighbour

Prediksi Rangking Q13 Untuk K10

Tabel III Hasil rangking Q13 Untuk K=10

Katergori Query	Dokumen Query	Data Set	Hasil Similarity	KNN (K=5)	Kategori	Keterangan
Keuangan	Q13	D13	1	1	Politik	Relevan
		D14	0.657754692965	2	Teknologi	Relevan
		D8	0.220925404771	3	Teknologi	Tidak Relevan
		D4	0.189847148284	4	Politik	Tidak Relevan
		D5	0.0949078829983	5	Politik	Tidak Relevan

Penjelasan Tabel III

Saat menguji, Rangking :

1. Menginput sebuah data lalu diproses oleh algoritma
2. Memasukan 10 dokumen Document Learning yang sudah disiapkan.
3. Menguji dokumen dengan klasifikasi 10 Document Classification, yang digunakan metode TFIDF serta similarity Euclidean Distance.
4. Ketika proses dari no.1 s/d no.3 dijalankan, lalukukan diakumulasi hasil
5. rangking (KNN).

Hasil *Precision* dan *Recall*

K-Nearest Neighbour dengan k=10 mengasilkan nilai *precision dan recall* sebagai berikut,

Tabel IV. Hasil *Precision* dan *Recall* untuk Untuk K=10

No	Kategori	Query	Precision	Recall
1	Politik	Q1	90%	90%
2	Politik	Q2	90%	90%
3	Politik	Q3	70%	70%
4	Politik	Q4	80%	80%
5	Politik	Q5	100%	100%
6	Teknologi	Q6	100%	100%

7	Teknologi	Q7	90%	90%
8	Teknologi	Q8	80%	80%
9	Teknologi	Q9	60%	60%
10	Teknologi	Q10	60%	60%

Penjelasan Tabel IV

Pada uji coba maka akan mehasilkan nilai *precision* dan *recall* yang relevan, berikut definisi sebuah hasil table *precision* dan *recall* diatas :

1. Q5(D5) dengan kategori Politik akan di hasilkan nilai yang valid precision 100% dan recall 100%.
2. Q6(D6) dengan kategori Politik akan di hasilkan nilai yang valid precision 100% dan recall 100%.
3. Q8(D8) dengan kategori Politik akan di hasilkan nilai yang valid precision 80% dan recall 80%.

KESIMPULAN

Nilai *precision* dan *recall* menggunakan pembobotan frekuensi kata dalam dokumen hasil rata-rata *precision* dan *recall***.**

Pengujian - Hasil

1. *Precision* - 90,00%
2. *Recall* - 90,00%

DAFTAR PUSTAKA

Adriani 2007 Adriani, M., NAzief, B., Asian J., & Williaws, H. E. Stemming Indonesia A Confixs Stripping Approach. ACMTransactions on Asian Language Information.

Asian, 2005 Asian, J., & Williams, H. E. Stemming Indonesia. Australia Computer Science Conference.

Barakbah, Ali Ridho 2010, Instance Base Classifier (Nearest Neightbour).

Ellis D, Fummer-Hines J, Willett P 1993, Measuring the degree of similarity between objects in text retrieval systems, Perspectives in Information Manajement, 3(2), 128-149.

Ignatow, G, dan Mihalcea. R, 2017, Text Mining A Guidebook for the Social Sciences, SAGE Publication, Inc, London, UK.

Salton, G. 1983, Introduction to Modern Information Retrieval. McGraw Hill.