DoubleClick: Journal of Computer and Information Technology E-ISSN: 2579-5317
P-ISSN: 2685-2152
Vol. 3, No. 1, August 2019, Pages 1-5 http://e-journal.unipma.ac.id/index.php/doubleclick

Penerapan Metode *Eucliean Distance* Untuk Ekstraksi Ciri Dokumen dan Kemiripan Dokumen

Yessi Yunitasari

Universitas PGRI Madiun
e-mail: yunita.yessi@gmail.com

Abstrak

Ekstraksi ciri merupakan sebuah proses untuk mendapatkan fitur-fitur yang terkandung dalam dokumen untuk proses text minning. Fitur-fitur yang dimiliki berfungsi untuk membedakan satu pola dengan pola yang lain. Metode ekstraksi fitur yang digunakan pada penelitian ini adalah metode TF-IDF. Metode TF-IDF (Term Frequency Inverse Document Frequency) adalah metode yang umum digunakan dalam proses kategorisasi teks. TF-IDF memiliki dua buah komponen. Komponen pertama adalah term-frequency dan komponen kedua adalah inverse document frequency. Setelah proses ekstraksi fitur kemudian dilakukan perhitungan kemiripan dokumen yang didasari dari fitur-fitur yang telah diekstraksi dari sejumlah dokumen-dokumen yang akan diperiksa kemiripannya. Ada banyak metode yang dipakai untuk menghitung kemiripan dokumen seperti euclidean distance dan cosine-similarity. Metode yang dipilih dalam program adalah euclidean distance. Pada penelitian ini akan dilakukan penerapan metode Eucliean Distance untuk ekstraksi ciri dokumen dan kemiripan dokumen.

Kata kunci : Ekstraksi Ciri, TF-IDF, Eucliean Distance, Kemiripan Dokumen (Similiaritas)

PENDAHULUAN

Text processing merupakan salah satu fokus dalam pengenalan pola. Salah satu hal yang bisa diselesaikan dengan metode *text processing* adalah perhitungan seberapa mirip dokumen dengan dokumen lainnya. Perhitungan kemiripan dokumen ini didasari dari fitur-fitur yang telah diekstraksi dari sejumlah dokumen-dokumen yang akan diperiksa kemiripannya.

Tahap awal dalam ekstraksi fitur atau ciri dalam text processing adalah pre-processing. Dalam pre-processing ini dilakukan eliminasi komponenkomponen dari dokumen yang tidak berkontribusi terhadap kemiripan dokumen. Tahap selanjutnya merupakan ekstraksi fitur yang merupakan total term yang bersifat unik dari seluruh term di semua dokumen. Kemudian selanjutnya dilakukan perhitungan bobot tf-idf sebagai basis dari perhitungan kemiripan dokumen. Banyak metode yang dapat dipakai untuk menghitung kemiripan dokumen, seperti menggunakan euclidean distance. Rumus euclidean distance adalah akar dari kuadrat perbedaan antara 2 vektor. Euclidean Distance dipakai untuk menghitung kemiripan antara dua buah vektor (Sutoyo,2009).

KAJIAN TEORI

Text Minning

Text mining dan text analytics adalah cakupan dari data mining yang cukup luas, kemudian disatukan pada kebutuhan yang sama, yaitu untuk mengubah teks ke dalam angka sehingga teknik tersebut dapat diterapkan pada document database yang besar (Miner, 2012). Text mining mengembangkan teknik yang berasal dari bidang lain untuk menyelesaikan permasalahan. Teknik standar yang biasa digunakan dalam text mining diantaranya adalah text clustering, text classification, taxonomy creation, ontology, latent corpus analysis dan document summarization (Fienerer dkk., 2008).

Preprocessing

Preprocessing data merupakan proses dimana teks dibersihkan dan dipersiapkan terlebih dahulu sebelum teks dianalisis (Haddi dkk, 2013). Preprocessing terhadap data dilakukan untuk meminimalisir adanya data-data yang kurang sempurna, adanya inkonsistensi data, dan gangguan yang terdapat pada data.

Ekstraksi Fitur

Feature extraction atau biasa disebut dengan ekstraksi fitur merupakan proses ekstraksi untuk mengidentifikasi entitas-entitas yang dimaksud (Siqueirra dan Barros, 2010). Proses ekstraksi fitur mengunakan TFIDF (Term Frequency-Inverse Document Frequency).

Gagasan standar frekuensi dalam basis corpus (corpus-based) pengolahan bahasa alami biasa disebut Term frequency (Yamamoto dan Church, 2001). Term frequency (tf t,d) menghitung kemunculan suatu term dalam suatu corpus berdasarkan bobot suatu term t pada dokumen d.

Metric yang umum dipergunakan dalam proses kategorisasi teks adalah TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF terdiri dari dua buah nilai komponen. Komponen yg pertama adalah term-frequency dan komponen ke 2 adalah inverse document frequency. Skema pembobotan TF-IDF adalah pemberian bobot pada term t pada dokumen d seperti pada persamaan berikut (Manning, dkk. 2009).

$$tf.idf_{t,d} = tf_{t,d} \times idf_{t} \quad \text{.....(1)}$$

$$idf_t = log \frac{N}{df_t} \quad \text{.....(2)}$$

Keterangan:

$tf.idf_{t,d}$: Bobot tiap kata dari term *t* yang terdapat dari dokumen *d*

$tf_{t,d}$: Bobot term *t* yang terdapat pada dokumen *d*

idf_t : *Inverse document frequency* yang terdapat pada *term t*

N : Jumlah semua dokumen

dft : Banyak dokumen yang memuat term *t* didalamnya

t : Term atau kata

Similiaritas

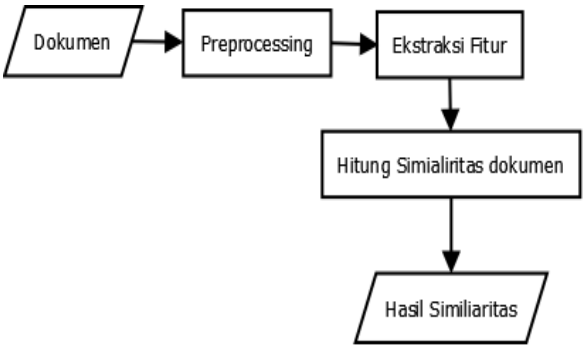
Metode Euclidean distance dapat dipakai untuk mengetahui kemiripan dokumen. Euclidean distance menghitung jarak paling pendek antara dua titik apabila digunakan didalam dua dimensi. Secara matematis formula tersebut dituliskan didalam persamaan dibawah ini:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} \quad \text{....(3)}$$

Keterangan:
d(i, j) = jarak 2 titik
Xin = nilai titik acuan
Xjn = nilai titik n target

METODE PENELITIAN

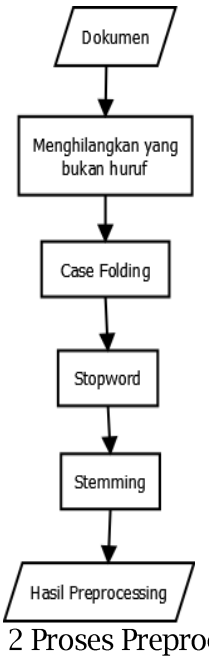
Penerapan metode *eucliean distance* untuk ekstraksi ciri dokumen dan kemiripan dokumen memakai metode penelitian seperti Gambar 1 .



Gambar 1 Metode Ekstraksi ciri dokumen dan kemiripan dokumen

Proses untuk ekstraksi ciri dokumen dan kemiripan dokumen ditampilkan pada Gambar 1. Hal pertama yang harus kita siapkan adalah dokumen yang akan kita ekstraksi fiturnya.

Pada penelitian ini ada 5 dokumen yang digunakan. Kemudian 5 dokumen tersebut dilakukan proses preprocessing, dilanjutkan ekstraksi fitur, dan dilanjutkan proses perhitungan similiaritas dokumen dengan menggunakan eucliean distance. Langkah-langkah preprocessing dapat dilihat pada Gambar 2.



Gambar 2 Proses Preprocessing

Macam-macam proses preprocessing yang digunakan adalah :

1. Menghilangkan yang bukan huruf
   Masukan data yang merupakan karakter khusus atau karakter selain huruf akan dihilangkan karena tidak dibutuhkan dalam ekstraksi ciri dokumen. Adapun yang dimaksud dengan karakter khusus adalah karakter yang bukan huruf yang akan dibersihkan, dapat dilihat pada Tabel 1.
   Tabel 1. Karakter Khusus

Karakter Khusus				
.	(	;	3	7
,	)	!	4	8
-	?	1	5	9
%	:	2	6	0

2. Case folding
   Proses selanjutnya yaitu case folding. Proses untuk merubah term menjadi bentuk huruf kecil disebut case folding.
3. Stopword
   Stopword merupakan proses eliminasi terhadap kata-kata yang sering muncul atau tidak terlalu berperan terhadap ekstraksi dokumen.
4. Stemming
   Proses mengubah suatu kata berimbuhan menjadi bentuk kata dasarnya. Proses stemming bertujuan untuk mengurangi jumlah fitur. Proses stemming yang dilakukan disini menggunakan library "Sastrawi" karena bahasa yang digunakan adalah bahasa Indonesia.

Setelah preprocessing dilakukan proses ekstraksi fitur. Ekstraksi fitur menggunakan TFIDF menggunakan formula 1. Fitur-fitur tersebut kemudian dihitung similiaritasnya dengan menggunakan Eucliean Distance seperti formula 3.

HASIL DAN PEMBAHASAN

Tahapan pertama dalam penelitian ini yang harus dilewati adalah tahapan preprocessing dokumen. Pada penelitian ini akan dilakukan perhitungan terhadap 5 dokumen. Gambar 3 akan ditampilkan proses menghilangkan yang bukan huruf, case folding dan stopword pada Dokumen 1.

```
DOKUMEN 1
==========================================================================
ketahanan pangan salah prasyarat dasar dimiliki dalam
rangka mewujudkan kesejahteraan masyarakat kenyataannya meskipun
kabupaten cirebon salah pensuplai beras wilayah jawa barat ada
beberapa desa mengalami rawan pangan minimnya indikator oleh
bkpk kabupaten cirebon menentukan status rawan pangan tahan pangan masih
menjadi kendala penganalisaan penyebab rawan pangan penelitian mencoba
mengembangkan sistem membantu bkpk kabupaten cirebon untuk
penentuan cluster rawan pangan tahan pangan rekomendasi bantuannya
melalui parameter indikator ketahanan kerawanan pangan ditentukan sistem
ini dibangun metode fuzzy cmeans mengelompokkan daerah
rawan pangan tahan pangan metode takagi sugeno kang rulebase dalam
pemberian rekomendasi bantuannya pengujian  uji aspek yang
paling berpengaruh penentuan desa rawan pangan aspek ketersediaan pangan
aspek akses pangan penghidupan aspek kesehatan gizi jumlah
penduduk dibawah garis kemiskinan desa memiliki akses penghubung yang
memadai rt akses listrik areal tanam terkena puso jumlah
buruh buruh tani swasta  indikator memiliki pengaruh penting
dalam penentuan daerah rawan pangan
```

Gambar 3 Proses menghilangkan yang bukan huruf, case folding dan stopword pada Dokumen 1

Proses stopword adalah proses eliminasi terhadap kata-kata yang sering muncul atau tidak terlalu berperan terhadap ekstraksi dokumen. Daftar stopword yang digunakan sebanyak 758 kata. Beberapa contoh stopword yang digunakan untuk penelitian ini ditampilkan pada Tabel 2.

Tabel 2 Contoh Stopword

Daftar Stopword		
ada	amatlah	atas
adalah	anda	atau
adanya	andalah	ataukah
adapun	antar	ataupun
agak	antara	awal
agaknya	antaranya	awalnya
agar	apa	bagai
akan	apaan	bagaikan
akankah	apabila	bagaimana
akhir	apakah	bagaimanakah
akhiri	apalagi	bagaimanapun
akhirnya	apatah	bagi
aku	artinya	bagian
akulah	asal	bahkan
amat	asalkan	bahwa
bahwa	belakangan	kan
baik	belum	kapan
bakal	belumlah	kapankah
bakalan	benar	kapanpun

balik	benarkah	karena
banyak	benarlah	karenanya
bapak	berada	kasus
baru	berakhir	kata
bawah	berakhirlah	katakan
beberapa	berakhirnya	katakanlah
begini	berapa	katanya
beginian	berapakah	ke
beginikah	berapalah	keadaan
beginilah	berapapun	kebetulan
begitu	berarti	kecil
begitukah	berawal	kedua
begitulah	berbagai	keduanya
begitupun	berdatangan	keinginan
bekerja	beri	kelamaan
belakang	berikan	kelihatan

Gambar 4 akan ditampilkan hasil dokumen 1 yang telah dilakukan proses stemming.

```
DOKUMEN 1==========================================================
tahan pangan salah prasyarat dasar milik dalam rangka wujud
sejahtera masyarakat nyata meski kabupaten cirebon salah pensuplai
beras wilayah jawa barat ada beberapa desa alami rawan pangan minim
indikator oleh bkpk kabupaten cirebon tentu status rawan pangan
tahan pangan masih jadi kendala penganalisaan sebab rawan pangan
teliti coba kembang sistem bantu bkpk kabupaten cirebon untuk tentu
cluster rawan pangan tahan pangan rekomendasi bantu lalu parameter
indikator tahan rawan pangan tentu sistem ini bangun metode fuzzy
cmeans kelompok daerah rawan pangan tahan pangan metode takagi
sugeno kang rulebase dalam beri rekomendasi bantu uji uji aspek
yang paling pengaruh tentu desa rawan pangan aspek sedia pangan
aspek akses pangan hidup aspek sehat gizi jumlah duduk bawah garis
miskin desa milik akses hubung yang pada rt akses listrik areal
tanam kena puso jumlah buruh buruh tani swasta indikator milik
pengaruh penting dalam tentu daerah rawan pangan
```

Gambar 4 Proses Stemming pada Dokumen 1

Proses selanjutnya adalah perhitungan *Term–Frequency*, *Document Frequency* dan kemudian dihitung TF-IDF. Sample hasil *Term–Frequency* ditampilkan pada Tabel 2. Sedangkan sample hasil *Document Frequency* ditampilkan pada Tabel 3.

Tabel 3 Sample hasil *Term–Frequency*

TF	milik	guna	kota	kasus	akses	aksi
DOK 0	3	0	0	0	3	0
DOK 1	0	0	1	0	0	3
DOK 2	0	0	0	1	0	0
DOK 3	1	2	1	1	0	0
DOK 4	0	1	0	0	0	0
TF	akurasi	alami	garis	ambil	analisis	metode
DOK 0	0	1	1	0	0	2
DOK 1	1	0	0	0	0	4
DOK 2	0	0	0	0	0	0
DOK 3	0	0	0	1	0	2
DOK 4	0	0	0	0	1	0

Nilai term frequency (tf) frekuensi dihitung dari kemunculan term pada dokumen. Pada Tabel 2 dapat dilihat misalnya term “milik” pada dokumen ke 1 berjumlah 3 dan pada dokumen 4 berjumlah 1.

Nilai document frequency (df) dapat dihitung dari banyaknya dokumen dimana suatu term muncul.

Tabel 4 Hasil Perhitungan *Document Frequency*

DF	milik	guna	kota	kasus	akses	aksi
	2	2	2	2	1	1
DF	akurasi	alami	garis	ambil	analisis	metode
	1	1	1	1	1	3

Setelah diketahui tf dan df maka dapat dilakukan perhiungan TF-IDF seperti pada Tabel 5.

Tabel 5 Hasil TF-IDF

TF-IDF	milik	guna	kota	kasus	akses	aksi
DOK 1	2.0794415	0	0	0	4.8283137	0
DOK 2	0	0	0.6931472	0	0	4.8283137
DOK 3	0	0	0	0.6931472	0	0
DOK 4	0.6931472	1.3862944	0.6931472	0.6931472	0	0
DOK 5	0	0.6931472	0	0	0	0
TF-IDF	akurasi	alami	garis	ambil	analisis	metode
DOK 1	0	1.6094379	1.6094379	0	0	0
DOK 2	1.6094379	0	0	0	0	0
DOK 3	0	0	0	0	0	0
DOK 4	0	0	0	1.6094379	0	0
DOK 5	0	0	0	0	1.6094379	0

Metode Euclidean distance dapat dipakai untuk mengetahui kemiripan dokumen. Ketika digunakan didalam dua dimensi euclidean distance merupakan jarak paling pendek antara dua titik. Hasil perhitungan Euclidean distance di representasikan dalam bentuk matrix 5 x 5 yang diagonalnya semua nilainya 0. Hasil kemiripan dokumen dapat dilihat pada Tabel 6

Tabel 6 Hasil Euclidean Distance

	Dok 1	Dok 2	Dok 3	Dok 4	Dok 5
Dok 1	0	28.29	0	0	0
Dok 2	28.29	0	28.29	28.29	28.29
Dok 3	0	28.29	0	0	0
Dok 4	0	28.29	0	0	0
Dok 5	0	28.29	0	0	0

KESIMPULAN DAN SARAN

Metode jarak dapat dipergunakan untuk menentukan tingkat ketidaksamaan (disimilarity degree) atau tingkat kesamaan (similarrity degree) antara dua vektor fitur. Teknik pengenalan pola dengan menggunakan metode jarak dengan metode Euclidean Distance telah dilakukan dan dipaparkan pada penelitian ini. Namun, penelitian ini masih dapat dikembangkan dengan metode-metode yang lain misalkan cosinus similiarity, Haming Distance ,Chebyshev, Angular Sparation, ataupun Correlation Coefficient.

DAFTAR PUSTAKA

Feinerer, Ingo, Kurt Hornik, and David Meyer. 2008. "Text Mining Infrastructure in R." *Journal Of Statistical Software* 25(5): 1–54.

Haddi, Emma, Xiaohui Liu, and Yong Shi. 2013. "The Role of Text Pre-Processing in Sentiment Analysis." *Procedia Computer Science* 17: 26–32.

Manning, Raghavan, dan Schutze, H., 2009, *Introduction to Information Retrieval*, Cambridge University Press.

Miner, G., Delen, D., Elder, J., Fast, A., dan Nisbet, R., 2012, *Practical Text Mining and Statistical Analysis for Non–Structured Text Data Applications*, Elsevier Inc.

Siqueira, H., dan Barros, F., 2010, A Feature Extraction Process for Sentiment Analysis of Opinions on Services, *Proceedings of International Workshop on Web and Text Intelligence*.

Sutoyo, S.Si., M.Kom,dkk. 2009 *Pengolahan Citra Digital*, Yogyakarta, Andi Offset.

Wurdianarto,dkk. 2014 Perbandingan Euclidean Distance Dengan Canberra Distance Pada Face Recognition, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro Semarang.

Yamamoto, M., dan Church, 2001, Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in A Corpus, *Computational Linguistics*, 27(1), 1-30.