Journal of Telematics and Informatics (JTI)
Vol.6, No.3, September 2018, pp. 224~229
ISSN: 2303-3703

Implementation of Correction to Answer Questions Essay Using Vector Space Model

Ipin Sugiyarto, Umi Faddillah, Arina Selawati
STMIK Nusa Mandiri, ASM BSI Jakarta, STMIK Nusa Mandiri
Jl. Damai No. 8, Margasatwa, Ragunan, Jakarta, Jl. Jatiwaringin Raya No. 18, Jatiwaringin, Pondok Gede, Jakarta, STMIK Nusa Mandiri, ASM BSI Jakarta, STMIK Nusa Mandiri
Jl. Damai No. 8, Margasatwa, Ragunan, Jakarta
e-mail: ipin.sugiyarto@gmail.com, umi.umf@bsi.ac.id, arinaselawati1513@gmail.com

Abstract

Correction answers to the essay still use manual way to judge based personal weight determined by each teacher. Correction value calculation result of the answers to the essay can use the quick way by computational methods. Used computational methods based text mining to test the system using the information retriveral vector space model. The result of the test method vector space model obtained by election the second-highest rank of valid answer content to answer valid student and teachers. Answer approach contained in rank-2 closest rank-1 with the result of the closest distance between the has a similarity value for student 0.45 and 0.10 teacher similarity value.

Keywords**: **correction, vector space model, rank, similarity

1. Introduction

Correction answers to the essay still using manual method weighted value to each question are has been determined by the teacher. Ratings way is still less effective as time because teacher must be read one by one student answers in detail and often obtained answers beyond text content which are already invalid of answers teachers. Method computation can be used to solve these problems.

Text mining is a multi-disciplinary science text which takes information involves analysis text, extraction information, clustering, categorization, visualization, database technology, machine learning and data mining [1]. Case who will discuss the invention back information based on keywords or query relevant as the similarity of answers essay between student and teachers.

In this case if the answer to the essay student according to the text content of the answers to the essay teachers and how big closeness between the two to be determine suitability and administration the correct decision or incorrect answers essay. The data used to use answers to the essay that has been determined by teachers in the form of language text Indonesia and text data from the essay question answers each student as much as three answers essay different matter.

Based on identification issues can underline how create application with effective algorithms for correct answers to the essay form text making it easier for teachers in determine the.

2. THEORETICAL

A. Text Mining

Text mining constitute result development of the data mining with the intention of to seek and find patterns draw from a set of text data with a large number [2]. Here are the steps to do with using text mining as follows:

B. Text Processing

Early stage text processing with way to lower case, change all characters capital letters to lowercase and do tokenizing namely the separation process description the initial form of

sentences into words and eliminate the word delimiter-delimiter such as a period (.), comma (,), space and character the figures in the word [3].

C. Feature Selection

Process feature selection constituten stage removal stopword and steaming against the word have uo [4] [5]. Stopword is vocabulary does not include characteristic (unique word) of a document [6]. For example , “di”, “oleh”, “pada”, “sebuah”, “karena” etc.

Steamming constitute process mapping and separation of various shapes (variants) of a word into a form of words basically (stem) [7]. The purpose of the process steaming is to remove affixes whether it be a prefix, suffix, or konfiks contained in each word.

D. Infromation Retriveral

Information retriveral constitute applied computer science that studies about taking a information based on the content and context of document. Process information restriveral could described as a process for search for relevant documents from a collection document by means of a search using keyword or query determined by user in finding the desired document. Salton explaint that the retriveral system information is a link between the user by source information is available at including a set of database search like, presented a set of ideas in a documents using batch concept, there some users who need ideas, but do not can identify and find well.

E. Vector Space Model

Vector space model a method used to find proximity or similarity term in a way give weight to term which has determined.

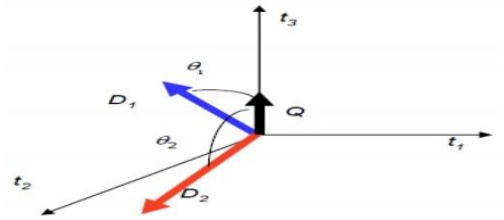

Figure 2.1 Illustration *Vector Space Model*

Information:
Ti = Word in database
Di = Documents
Q = Keywords

How to use the calculation equation on vector space models with calculate the value cosines the angle of the two vector, that is vector keywords and vector every document all-i

$$Cosine\,\theta_{D_i} = Sim(Q,D_i)$$

Equations …… 1st

Information:
Q = *Query*
Di = Document to (i)

$$Sim(Q,D_i) = \frac{\sum_j w_{i,j}\, w_{q,j}}{\sqrt{\sum_j w_{i,j}^2}\sqrt{\sum_i w_{q,j}^2}}$$

Equations …… 2nd

Information:
Di = Documents to-i
Q = *Query*

J = The word in the entire document

$$Cosine\ \theta_{D_i} = \frac{Q \cdot D_i}{|Q| * |D_i|}$$

Equations …… 3rd

Information:
Di = Document to-i
Q = *Query*
|Q| = *Vector* Q
|Di| = *Vector* Di

3. RESULT

Here are the results of the text mining on text content ot the answers to the essay. Use text answers to the essay of 18 word subject, predicate and object contained in each answers to the from respondens are students and teachers as well as one keywords or queries in the form of the word “norma”.

Table I Data Training

Query	D1 (Pengajar)	D2 (Murid_1)	D3 (Murid_2)	D4 (Murid_3)
Norma	Seperangkat aturan atau kaidah yang digunakan untuk mengatur tingkah laku masyarakat yang berisi perintah dan larangan	Adab yang mengatur perilaku seseorang	Norma itu aturan	adab

Proses Tokenisasi

Term	Doc ID
Seperangkat	1
aturan	1
kaidah	1
digunakan	1
mengatur	1
tingkah	1
laku	1
masyarakat	1
berisi	1
perintah	1
larangan	1
adab	2
mengatur	2
perilaku	2
seseorang	2
norma	3
aturan	3
adab	4

Proses Indexer

Term	Doc ID
adab	2
adab	4
aturan	1
aturan	3
berisi	1
digunakan	1
kaidah	1
laku	1
larangan	1
masyarakat	1
mengatur	1
mengatur	2
norma	3
perilaku	2
perintah	1
Seperangkat	1
seseorang	2
tingkah	1

Proses Filtering & Steaming

Term	Doc ID Freq.	Plot
adab	2	2, 4
atur	2	1, 3
isi	1	1
guna	1	1
kaidah	1	1
laku	1	1
larang	1	1
masyarakat	1	1
norma	1	3
laku	1	2
perintah	1	1
perangkat	1	1
seorang	1	2
tingkah	1	1

Figure I Tokenize Result

Tokenize stage classifying words as well as address identification documents. Process indexer sorted alphabetically, then process filtering and steaming that is aliminating the prefix

and said additional suffix and calssify it in one word unique and giving the number of times appearance of frequency document and mapping the location of the plot of the origin of the document.

The next stage is to calculate the weight document such as the following table:

	Token	tf					df	
No		Q	D1	D2	D3	D4		D/df
1	norma	1	0	0	1	0	1	4
2	perangkat	0	1	0	0	0	1	4
3	atur	0	1	1	1	0	3	1
4	tingkah	0	1	0	0	0	1	4
5	perilaku	0	1	1	0	0	2	2
6	masyarakat	0	1	0	0	0	1	4
7	isi	0	1	0	0	0	1	4
8	perintah	0	1	0	0	0	1	4
9	larang	0	1	0	0	0	1	4
10	adab	0	0	1	0	1	2	2
11	seorang	0	0	1	0	0	1	4
12	laku	0	1	0	0	0	1	4
13	guna	0	1	0	0	0	1	4
14	kaidah	0	1	0	0	0	1	4

Figure II Calculation of Weight Documents

Weight calculation documents using equation formula:

$$D/df = \frac{\text{The total number of documents}}{\text{The number of documents sought}}$$

After the calculation of the weights document obtained, the next step looking for value tf-idf weighting ie weight calculation results multiplication of documents with the log results from D/df with the following results.

IDF	W				
log (D/df)	Q	D1	D2	D3	D4
0.602	0.602	0	0	0.602	0
0.602	0	0.602	0	0	0
0.125	0	0.125	0.125	0.125	0
0.602	0	0.602	0	0	0
0.301	0	0.301	0.301	0	0
0.602	0	0.602	0	0	0
0.602	0	0.602	0	0	0
0.602	0	0.602	0	0	0
0.602	0	0.602	0	0	0
0.301	0	0	0.301	0	0.301
0.602	0	0	0.602	0	0
0.602	0	0.602	0	0	0
0.602	0	0.602	0	0	0
0.602	0	0.602	0	0	0

Figure III Calculation of tf-idf *weighting*

The nect stage after the tf-idf obtained then documents and calculate distance query of results the square root of each query (Q) and data (Di), the following calculation results:

Token	W				
	Q^2	D1^2	D2^2	D3^2	D4^2
norma	0.362	0	0	0.362	0
perangkat	0	0.362	0	0	0
atur	0	0.016	0.016	0.016	0
tingkah	0	0.362	0	0	0
perlaku	0	0.091	0.091	0	0
masyarakat	0	0.362	0	0	0
isi	0	0.362	0	0	0
perintah	0	0.362	0	0	0
larang	0	0.362	0	0	0
adab	0	0	0	0	0.091
seorang	0	0	0.362	0	0
laku	0	0.362	0	0	0
guna	0	0.362	0	0	0
kaidah	0	0.362	0	0	0
	SQRT Q	SQRT Di			
	0.602	1.835	0.748	0.615	0.301

Figure IV count within the document and query

Then the next step counting dot value results from multiplying the value of the square document all i with the square of the query. As the result of the following:

Menghitung dot			
Q*D1	Q*D2	Q*D3	Q*D4
0	0	0.131	0
0.131	0	0	0
0.006	0.006	0.006	0
0.131	0	0	0
0.033	0.033	0	0
0.131	0	0	0
0.131	0	0	0
0.131	0	0	0
0.131	0	0	0
0	0.033	0	0.033
0	0.131	0	0
0.131	0	0	0
0.131	0	0	0
0.131	0	0	0
Sum (Q*Di)			
1.220	0.203	0.137	0.033

Figure V Calculation Dot.

The next stage is to calculate similarity between document by the formula equation as follows::

$$Sim = \frac{\sum (Q*Di)}{\sqrt{Q^2} * \sqrt{D^2}}$$

The similarity calculation produces a value which can be used to determine ranking among documents as a determinant of outcome closeness documents that have level the highest similarity.

	Cosine Ø D1	Cosine Ø D2	Cosine Ø D3	Cosine Ø D4
	1.10	0.45	0.37	0.18
Rank	1	2	3	4

Figure VI Calculation of Similarity

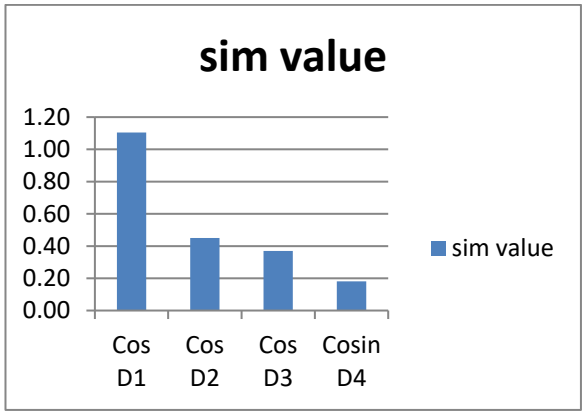

Figure VII Chart of Similarity Between Document

4. CONCLUSION

Result implementation of correction answers to the essay use space vector method model generate a value of similarity between document D1 to the value of 1.10, the document D2 with a value of 0.45, document D3 with value 0.35 and document D4 with the value of 0,18. The similarity of the calculation results it can be concluded that the correction of errors essay most appropriate answers in between D2, D3, D4 with document 1 (D1) the norma is a document to the query-2 (D2) with the highest level of similarity close to the value similarity document to-1.

5. References

[1] Feldman, R. & Dagan, L. (1995) Knowledge discovery in textual database (KDT). Inproceedings of the first International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, Agustus 20-21, AAAI Press, 122-117.

[2] Klasifikasi Konten Berita Dengan Metode Text Mining. Jurnal Dunia Teknologi Informasi Vol. 1, No.1, (2012), 14-19.

[3] Lin, S. 2008. A Document Classification and retriveral system for R&D in semiconductor industry-A hybrid approach. Expert system 18, 2:4753-4764.

[4] Berry, M. W. & Kogan, J. 2010. Text Mining application and theory. WILEY: United Kingdom.

[5] Draut, E. Fang, F. Sistla, P. Yu, S & Meng, W. 2009. Stop word and related problems in web interface integration. http://www.vldb.org/pvldb/2/vldb09384.pdf. Diakses pada tanggal 22 Jan 2018.

[6] Tala, Fadilla, Z. 2003. A Study of Stemming Effect on Information Retriveral in Bahasa Indonesia. Institute for Logic, language and computation Universitate van Amsterdam the netherland. http://www.illc.uva.nl/Research/Reports/MoL-2003- 02.text.pdf. Diakses pada tanggal 23 Januari 2018.

[7] Weiss, S. M. Indurkhya, N. Zang, T. Dhamerau, F. J. 2005. Text Mining: Predictif method of Analyzing Unstructured Information. Springer: New York.

[8] Salton, G. 1989. Automatic Text Processing, The Transformation, Analysis and Retriveral of Information by Computer, Addison – Westly Publishing Company, Inc. All right reserved