JURNAL ILMIAH EKONOMI DAN BISNIS, Vol.13, No.1, Juli 2020, pp. 13 - 16
p-ISSN : 1979-0155 (print)
e-ISSN : 2614-8870 (online)
http://journal.stekom.ac.id/index.php/e-bisnis

The thesis topic similarity test with TF-IDF method

Sulartopo
Sekolah Tinggi Elektronika dan Komputer (STEKOM)
Jl. Majapahit No. 605 Semarang, 024-6723456, e-mail: sulartopo@stekom.ac.id

ARTICLE INFO

Article history:

Received 30 Mei 2020
Received in revised form 12 Juni 2020
Accepted 30 Juni 2020
Available online 1 Juli 2020

ABSTRACT

This research is to clarify how to test the thesis topic similarity, make it easy to check the topic thesis, whether it has been made by a student before. In this regard, an important issue that can be raised is how to make a thesis topic similarity test the manual way to be automated. The purpose of this study is a research method similarity of thesis topics using the TF-IDF method. In this research the system has two stages of process, the first mining the text that is categorizing the thesis that has been categorized using the TF-IDF algorithm, which is to read the appearance of each word in the contents of the document. The second stage results from the TF-IDF algorithm are reprocessed with the VSM algorithm. The end result of this program will get the names of documents that have a degree of similarity with keywords.

Keywords: Term Frequency Inverse Document Frequency, Vector Space Model, extracting text, text similarity test.

1. Introduction

Thesis is a scientific paper based on the results of field research and or literature studies compiled by students in accordance with the study program as the final project in formal studies at tertiary institutions. Where the research itself is the whole activity both in the mind and in real activities carried out by students to solve a problem in the field of scientific science in the context of thesis preparation. With the aim and usefulness to present scientific research findings that are useful for the community and scientific development.

The College of Electronics and Computers, a private university in Semarang, states that the thesis is the final official scientific paper of a student in completing the Bachelor Program. Thesis is a proof of student academic ability in research related to the problem discussed. Thesis prepared by students must have the following characteristics: is the original work, not a plagiarism for some / as a whole; have relevance to science / study programs; has theoretical or practical benefits; in accordance with scientific principles; use standard Indonesian, good and right.

One of the procedures of a student in preparing a thesis is to submit the title / topic of the thesis to the Head of the Study Program, to get approval. During this time in giving approval, the Study Program first checks the title / topic of the thesis, whether it has been made by a student before. This is so that the results of the student thesis are in accordance with the characteristics of the thesis above. In this study, what will be done is to fulfill the first

Received Mei 30, 2020; Revised Juni 39, 2020; Accepted Juli 1, 2020

p-ISSN : 1979-0155 e-ISSN : 2614-8870

characteristic, which is the original work of students, not plagiarized for some / as a whole by using text mining TF-IDF method.

2. Research Method

The research methods applied in this study are as follows:

The data used in this study is a student thesis database from the Computer Systems study program, the data amounted to 100 student thesis title data used as comparative / training data. Whereas the 5 proposed data for the students' thesis titles are used as data compared to testing.

Text mining is a variation of data mining that tries to find interesting patterns from a large collection of textual data. In this study utilizing a database of text mining results in previous studies, namely "Categorizing the thesis title using the NBC method" [1].

In the process of analyzing the similarity of documents, the author uses text mining techniques with the TF-IDF algorithm, to check the appearance of each word in the document content from the results of tokenizing, filtering, and word counting to calculate the TF-IDF formula that will produce the weight of the document [2].

$$TF-IDF(w,d) = \frac{TF-IDF(w,d)}{\sqrt{\sum_{k=1}^{t}\left(TF_{(w,k)}\right)^2 \times \left[log\left(\frac{N}{DF(w)}\right)+1\right]^2}}$$

To obtain good results, the results of the TF-IDF algorithm will be processed again with the VSM algorithm, using the calculation of similarity with Cosine's approach, which is stated in the formula [3]. The end result of this program will get the names of documents whose contents have a degree of similarity with keywords.

$$Similarity(d_j, q_k) = \frac{\sum_{i=1}^{n}\left(td_{ij} \times tq_{ik}\right)}{\sqrt{\sum_{i=1}^{n} td_{ij}^2 \times \sum_{i=1}^{n} tq_{ik}^2}}$$

3. Results and Analysis

3.1. Application Concept

The application to be built is the first document that will be tested by uploading files (pdf, doc, and txt), to get documents in text format. Next the text document will be processed with text mining techniques that will produce keywords that represent the contents of the document to determine the results of sorting documents. Furthermore, the keywords that have been obtained can be processed again with the TF-IDF algorithm to get the value of the weight of the document, then re-do the calculations with the VSM algorithm. After all processes are complete, the document weight value will appear from the largest to the smallest value, the document with the highest weight value is the document that has the highest level of similarity.

3.2. Document Similarity Process

In the process of analyzing the similarity of documents, the authors use text mining techniques with the TF-IDF and VSM algorithms. TF-IDF algorithm will check the appearance of each word in the document contents from the results of tokenizing, filtering, and word counting to calculate the TF-IDF formula that will produce the weight of the document. To obtain good results, the results of the TF-IDF algorithm will be processed again with the VSM algorithm.

The end result of this program will get the names of documents whose contents have a degree of similarity with keywords. The following is a general description of the document similarity analysis program.

To analyze the level of similarity between a keyword in a document with another document, the step that must be done is to choose the document you want to compare and the

document you are comparing. The document selected for comparison has keywords, and those keywords will be analyzed for similarity with other documents.

After getting a collection of keywords in the document you want to compare, the program will repeat the number of keywords. In this iteration process, each keyword will be compared with all comparative documents, to get the value of the keyword weight (KW), and the weight of the document against the keyword (WDK).

The calculation process of the steps above is carried out for every one keyword with all comparative documents. So that the results are more optimal the results are combined with VSM calculations, with the formula Cosine Value = (WDK) / (KW / WD).

After calculating the cosine value in the VSM algorithm, the results of the calculation value in each document will be sorted from the highest cosine value. Documents that have the highest cosine value are documents that have the highest level of similarity with keywords.

3.3. Analysis of Thesis Topics Similarity Program Output

At the stage of applying the thesis topic similarity program to the data used, it is done by applying each 1 thesis topic document compared to 100 thesis topic documents that are already in the database. If a document that is to be compared also exists in a comparison document, the document will have a similarity value, with the similarity level as follows: VS > 0.85 : Very Similar; VS > 0.70 : Similar; VS >= 0.50 : Somewhat Similar; VS < 0.50 : Not Similar.

The results of implementing the program on the data being compared, produce various results. As shown in the following table:

Table 1: Document Similarity Level

Number	Topic Compared	Value of Similarity (VS) to Comparison Topics	Degree of Similarity
1	ProposedTopic01.pdf	Doc 1: Topic_046.pdf ; VS= 0,55185 Doc 2: Topic_071.pdf ; VS = 0,35252	Somewhat Similar Not Similar
2	ProposedTopic02.pdf	Doc 1: Topic_004.pdf ; VS = 0,82679 Doc 2: Topic_085.pdf ; VS = 0,21884	Very Similar Not Similar
3	ProposedTopic03.pdf	Doc 1: Topic_030.pdf ; VS = 0,74270 Doc 2: Topic_062.pdf ; VS = 0,26635	Similar Not Similar
4	ProposedTopic04.pdf	Doc 1: Topic_063.pdf ; VS = 0,62194 Doc 2: Topic_070.pdf ; VS = 0,20325	Somewhat Similar Not Similar
5	ProposedTopic05.pdf	Doc 1: Topic_062.pdf ; VS = 0,47839 Doc 2: Topic_089.pdf ; VS = 0,32451	Not Similar Not Similar
6	ProposedTopic06.pdf	Doc 1: Topic_013.pdf ; VS = 0,50872 Doc 2: Topic_083.pdf ; VS = 0,35648	Somewhat Similar Not Similar
7	ProposedTopic07.pdf	Doc 1: Topic_008.pdf ; VS = 0,37856 Doc 2: Topic_062.pdf ; VS = 0,12261	Not Similar Not Similar
8	ProposedTopic08.pdf	Doc 1: Topic_022.pdf ; VS = 0,36828 Doc 2: Topic_040.pdf ; VS = 0,28248	Not Similar Not Similar
9	ProposedTopic09.pdf	Doc 1: Topic_011.pdf ; VS = 0,76551 Doc 2: Topic_036.pdf ; VS = 0,39651	Similar Not Similar
10	ProposedTopic10.pdf	Doc 1: Topic_020.pdf ; VS = 0,55834 Doc 2: Topic_052.pdf ; VS = 0,09538	Somewhat Similar Not Similar

The results of the similarity value / document similarity, shows the level of similarity between the proposed thesis topic (which is compared) with the thesis topic database (comparison). Thus for the proposed thesis topic that results in a degree of similarity: Very Similar / Similar / Somewhat Similar, it is necessary to manually re-check by the Head of the Study Program on the recommended topic documents. Whereas the proposed thesis topic that results in a similarity level is Not Similar, there is no need to manually re-check.

p-ISSN : 1979-0155 e-ISSN : 2614-8870

4. Conclusion

Measurement of document similarity in this study was carried out on 100 documents of abstraction which resulted in 76 documents that had a corresponding similarity value and 24 documents that did not have an incompatible similarity value.

The appropriateness of the similarity level of this document is assessed if one document to be compared is also found in the compilation of documents, then the same document must have the highest similarity value than other documents, otherwise the results are said to be inappropriate. This result is not appropriate because the value of the document weighting of keywords compared to the value is smaller and the value of the document weight is large. So the highest document similarity value is determined by the weight of a document to the keywords and the small value of the weight of the document.

References

[1] Sulartopo. Pengkategorian Topik Skiripsi Dengan Metode NBC. *Jurnal Ilmiah Ekonomi dan Bisnis.* 2015; Vol.8(1).
[2] Feldman, R. & Sanger, J. The Text Mining Handbook. New York: Cambridge UniversityPress. 2007.
[3] Frakes, W. B. & Baeza, R. Information Retrieval Data Structure and Algorithms. NewJersey: Prentice-Hall. 1992.