PASCAL Journal of Computer Science and Informatics
https://jurnal.devitara.or.id/index.php/komputer

E-ISSN: 3047-5074 Volume 1 Nomor 2 Tahun 2024

Analysis and Implementation of Similarity Measurement in Documents Using Semantic Methods

[1]Satria Yudha Prayogi, [2]Sony Bahagia Sinaga
[1]Universitas Islam Sumatera Utara,
[2]STMIK Mulia Darma

ARTICLE INFO

Keywords:
Similarity,
Document,
Text Mining,
Semantic

Email :
Satria.yp@ft.uisu.ac.id,
sonybahagia@gmail.com

ABSTRACT

The number of documents available in digital form is increasing. Meanwhile, one document and another document may be related to each other, but they must not be plagiarized without including the reference source. For this reason, a mechanism for detecting similarities is needed. This research only discusses similarity in documents. In this research, the technique used to solve the above problem is to use text mining techniques to categorize the documents searched according to keywords. Meanwhile, to search for documents according to keywords, the indexing process is used to display documents that are searched for according to keywords. Semantics is a technique used by search engines to match key words on one page with another page. This method has been used very often before, because it is very precise and easy. The weight values (W) of D1 and D2 are the same. If the document weight sorting results cannot be sorted quickly, because both W values are the same, then a calculation process using the vector-space model algorithm is needed. The idea of this method is to calculate the cosine value of the angle of two vectors, namely W from each document and W from keywords. From the research results, it can be seen that document 3 (D3) has the highest level of similarity to keywords, followed by D2 and D1.

INTRODUCTION

The development of information technology at this time the exchange of information is getting younger, this not only brings a positive impact on technological advancement, but also brings a negative impact that is almost inevitable, namely plagiarism. The problem, to overcome the practice of plagiarism is proven in a *persuasive* way here that means approaching the student concerned that the action is not good to do because it has been proven to be ineffective. Document Similarity Measurement is the right solution that should be done so that the fraudulent act can be minimized.

Therefore, it is necessary to design an application for measuring document similarity. The similarity of documents or the similarity of documents can be known by providing keyword variables contained in the document whether there is the word or not, the process of checking existing documents may feel easy if there are only a few documents but what if hundreds of documents, for that it is necessary to design a system that can carry out the inspection process easily and quickly.

Text mining, also known as *text* data *mining*, is a process of extracting patterns in the form of useful information and knowledge from a large number of text data sources, such as *word documents*, *pdfs*, and text excerpts. *Text mining* looks for patterns in text in natural, unstructured language such as books, *emails*, articles, web pages. Activities that are usually carried out by *text mining* are *text categorization*, *text clustering*, *conception / entity extraction*. In a full-text search, the search engine checks all the words in each saved document as if trying to match the search criteria (*the text* is determined by the user). *Full-text-searching* techniques became common in online bibliographic databases in the 1990s. Many sites and application programs (such as word processing software) provide full-text searchability. Some web search engines, such as *AltaVista*, use *full-text-search* techniques, while others *index* only a portion of a *web* page that is checked by their indexing system. Semantics is the process after going through the scanning and parsing process. At this stage, checks are carried out on the final structure that has been obtained and checked for conformity with existing program components. Globally, the function of a *semantic analyzer* is to determine the meaning of a set of instructions contained in a source program. *Syntax* defines a correct form of programming of a language. Semantics define the meaning of a syntax-correct program of the language. The semantics of a language require some kind of expression to transmit a value of truth.

This study aims to analyze and implement a semantic-based document similarity measurement method. The main focus is to evaluate how effective this method is compared to a keyword-based approach in various application scenarios.

METHOD

This study was carried out through a series of steps designed to test and compare the effectiveness of the TF-IDF method and the semantic method in measuring document similarity. The stages of the research include:

a. Data Collection: The data used in this study consisted of a collection of text documents obtained from various sources, including corpus scientific texts, news articles, and other documents. This data is selected to ensure sufficient content and context variation.
b. Data Preprocessing: At this stage, the documents that have been collected are processed through several preprocessing steps such as tokenization, deletion of stop words, and stemming. This process is done to clean the data and prepare it for further analysis.
c. Implementation of TF-IDF: TF-IDF is applied to the entire corpus of documents to produce vector representations of words. Each document is represented as a vector that reflects the frequency and weight of words in the context of the corpus. The similarity measurement between documents was then carried out using cosine similarity on the TF-IDF vector.

RESULTS AND DISCUSSION

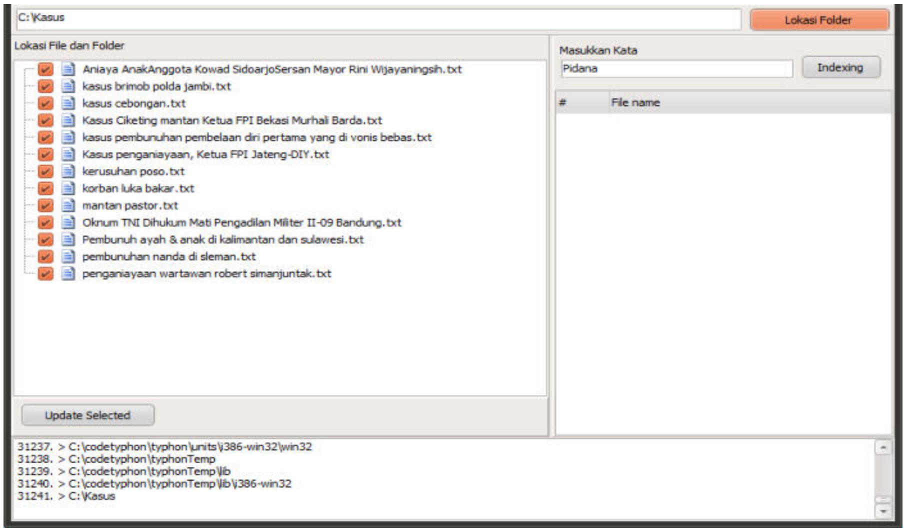
The searching process, as the name suggests, this process is the process to find data that has been stored. This process is often called the information retrieval process. Information retrieval is a term for studying the search system so that it gets the information it is looking for, starting from indexing, searching, and recalling data. This also applies to unstructured data searches. As a storage medium, all locations on the PC itself can be used, the search is done by looking for documents on the drive (where the file is stored). The working context of this search is to find documents stored on the computer. To analyze the level of similarity between a keyword in a document and another document, the stage that must be done is to select the document you want to compare and the document that is the comparison. The document that is selected for comparison has keywords, and these keywords will be analyzed for similarity with other documents. After getting the set of keywords in the document you want to compare, the program will repeat the number of keywords. In this loop process, each keyword will be compared with the entire comparison document, to get the value of the keyword weight (WK2), and the weight of the document to the keyword (WDK2). After the weight (W) of each document is known, a sorting process is carried out where the greater the value of W, the greater the level of similarity of the document to the keyword, and vice versa.

An example of a simple implementation of TF – IDF is as follows:

Keyword (kk)	= logistics knowledge
Document 1 (D1)	= logistics transaction management
Document 2 (D2)	= knowledge between individuals
Document3 (D3)	= in knowledge management there is a transfer of knowledge logistics

So the number of documents (D) = 3

After the tokenizing stage and the filtering process, the interjectives in document 2 and the inner words and in document 3 are deleted. From the final results, it can be seen that document 3 (D3) has the highest level of similarity to keywords, then followed by D2 and D1. This is a display of the implementation results of the software.



C:\Kasus

Lokasi File dan Folder

- Aniaya AnakAnggota Kowad SidoarjoSersan Mayor Rini Wijayaningsih.txt
- kasus brimob polda jambi.txt
- kasus cebongan.txt
- Kasus Ciketing mantan Ketua FPI Bekasi Murhali Barda.txt
- kasus pembunuhan pembelaan diri pertama yang di vonis bebas.txt
- Kasus penganiayaan, Ketua FPI Jateng-DIY.txt
- kerusuhan poso.txt
- korban luka bakar.txt
- mantan pastor.txt
- Oknum TNI Dihukum Mati Pengadilan Militer II-09 Bandung.txt
- Pembunuh ayah & anak di kalimantan dan sulawesi.txt
- pembunuhan nanda di sleman.txt
- penganiayaan wartawan robert simanjuntak.txt

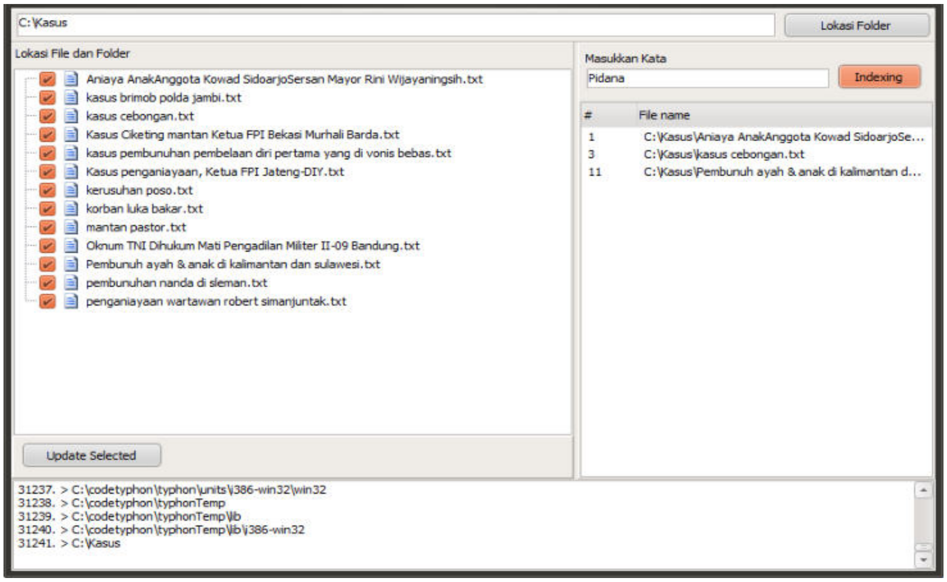
Masukkan Kata: Pidana

#	File name

```
31237. > C:\codetyphon\typhon\units\i386-win32\win32
31238. > C:\codetyphon\typhonTemp
31239. > C:\codetyphon\typhonTemp\lib
31240. > C:\codetyphon\typhonTemp\lib\i386-win32
31241. > C:\Kasus
```

Figure 1. Search Process

To determine the location of the file where the text *file* contains case information from the persecution, the location of the folder must be determined first.



C:\Kasus

Lokasi File dan Folder

- Aniaya AnakAnggota Kowad SidoarjoSersan Mayor Rini Wijayaningsih.txt
- kasus brimob polda jambi.txt
- kasus cebongan.txt
- Kasus Ciketing mantan Ketua FPI Bekasi Murhali Barda.txt
- kasus pembunuhan pembelaan diri pertama yang di vonis bebas.txt
- Kasus penganiayaan, Ketua FPI Jateng-DIY.txt
- kerusuhan poso.txt
- korban luka bakar.txt
- mantan pastor.txt
- Oknum TNI Dihukum Mati Pengadilan Militer II-09 Bandung.txt
- Pembunuh ayah & anak di kalimantan dan sulawesi.txt
- pembunuhan nanda di sleman.txt
- penganiayaan wartawan robert simanjuntak.txt

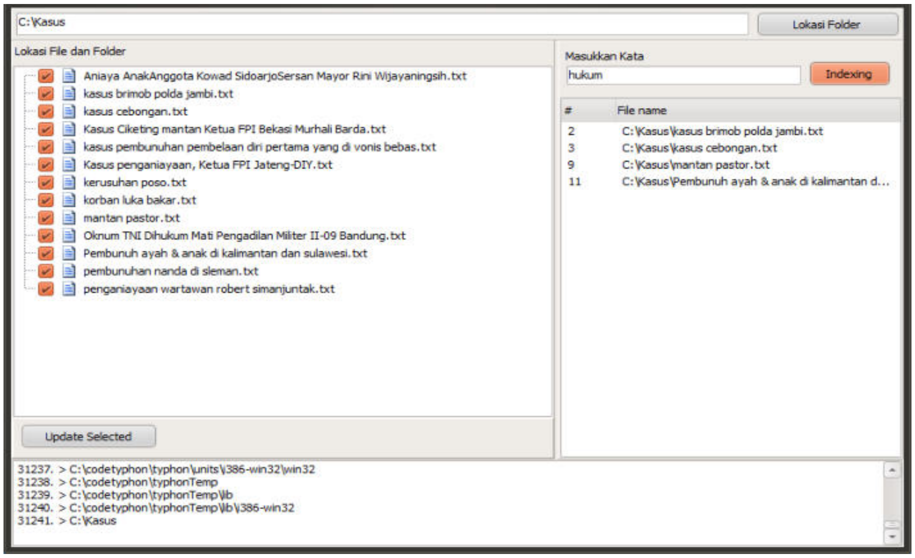
Masukkan Kata: Pidana

#	File name
1	C:\Kasus\Aniaya AnakAnggota Kowad SidoarjoSe...
3	C:\Kasus\kasus cebongan.txt
11	C:\Kasus\Pembunuh ayah & anak di kalimantan d...

```
31237. > C:\codetyphon\typhon\units\i386-win32\win32
31238. > C:\codetyphon\typhonTemp
31239. > C:\codetyphon\typhonTemp\lib
31240. > C:\codetyphon\typhonTemp\lib\i386-win32
31241. > C:\Kasus
```

Figure 2. Search Process Results

Pay attention to the image, when you select one *file* name in the *indexing* section, automatically the information in the *file* and *folder* sections is also selected and colored and at the bottom the information from *the selected* text is displayed.



C:\Kasus

☑ Aniaya AnakAnggota Kowad SidoarjoSersan Mayor Rini Wijayaningsih.txt
☑ kasus brimob polda jambi.txt
☑ kasus cebongan.txt
☑ Kasus Ciketing mantan Ketua FPI Bekasi Murhali Barda.txt
☑ kasus pembunuhan pembelaan diri pertama yang di vonis bebas.txt
☑ Kasus penganiayaan, Ketua FPI Jateng-DIY.txt
☑ kerusuhan poso.txt
☑ korban luka bakar.txt
☑ mantan pastor.txt
☑ Oknum TNI Dihukum Mati Pengadilan Militer II-09 Bandung.txt
☑ Pembunuh ayah & anak di kalimantan dan sulawesi.txt
☑ pembunuhan nanda di sleman.txt
☑ penganiayaan wartawan robert simanjuntak.txt

hukum

#	File name
2	C:\Kasus\kasus brimob polda jambi.txt
3	C:\Kasus\kasus cebongan.txt
9	C:\Kasus\mantan pastor.txt
11	C:\Kasus\Pembunuh ayah & anak di kalimantan d...

```
31237. > C:\codetyphon\typhon\units\i386-win32\win32
31238. > C:\codetyphon\typhonTemp
31239. > C:\codetyphon\typhonTemp\lib
31240. > C:\codetyphon\typhonTemp\lib\i386-win32
31241. > C:\Kasus
```

Figure 3. Legal Word Indexing

CONCLUSION

The information retrieval system application is designed using the *Embarcadero Delpih XE3* programming language. The use of tf / idf algorithm as term weighting as a measurement of text similarity in the process of searching for document information in the document similarity measurement application is able to obtain more accurate and effective search results. The performance of the application is fast enough to process or execute its functions.

REFERENCES

Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In Proceedings of the First International Conference on Machine Learning (pp. 133-142).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

H. Yan, N. Yang, Y. Peng, and Y. Ren, "Data mining in the construction industry Present status, opportunities, and future trends," *Automation in Construction*, vol. 119. Elsevier B.V., p. 103331, Nov. 01, 2020. doi 10.1016/j.autcon.2020.103331.

Z. Wang, Y. Li, D. Li, Z. Zhu, and W. Du, "Entropy and gravitation based dynamic radius nearest neighbor classification for imbalanced problem," *Knowl Based Syst*, vol. 193, no. xxxx, p. 105474, 2020, doi 10.1016/j.knosys.2020.105474.

J. Mahasiswa and U. Negeri, "View metadata, citation and similar papers at core.ac.uk".

M. Nurjannah and I. Fitri Astuti, "PENERAPAN ALGORITMA TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK TEXT MINING Mahasiswa S1 Program Studi Ilmu Komputer FMIPA Universitas Mulawarman Dosen Program Studi Ilmu Komputer FMIPA Universitas Mulawarman," *J. Inform. Mulawarman*, vol. 8, no. 3, pp. 110–113, 2013.

Gandhis Ulta Abriania, "Implementasi Metode Semantic Similarity Untuk Pengukuran Kemiripan Antar Kalimat", ILKOMNIKA, Vol. 1, No. 2

Davis Valentino, "Indexing dan Searching Document Menggunakan Metode Semantic Suffix Tree Clustering Berbasis Android", Jurnal Infra,