ABSTRACT

IMPLEMENTION OF TF-IDF IN DOCUMENTS SIMILARITY MEASURE

***By*:**
Adi Ryansyah
3.10.007

Documents similarity measure is a time consuming problem. The large amount of documents and the large number of pages per document are causing the similarity measures to become a complicated and hard job to do manually.

In this research, a system that can automatically measuring similarity between documents is built by implementing TF-IDF. Measurements are carried by first creating a vector representation of documents being compared. This vector representation containing the weight of each term in the documents. After that, the similarity values are calculated using cosine similarity.

This research used waterfall model. System design is done using Unified Modeling Language (UML) and implemented in Java programming language with Netbeans as IDE. The documents used in this research are ten thesis reports from computer science major, eight from information system major, and ten reports from industrial engineering major of Sekolah Tinggi Teknik Musi.

The finished system can carry out comparison of documents in pdf or word format. Document comparison can be done using all the chapters in the report, or just a few selected chapters that are considered significant. Based on experiment, it can be concluded that tf-idf needs at least three documents to be available in the document collection being processed. The test of correlation shows that for document in pdf format, there is a significant correlation between the amount of characters in the document with the processing time.

Keywords: documents similarity measure, tf-idf, vector, cosine similarity