WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Similarity Detection System

[1]Murtaza Singapurwala, [2]Mayuri Balchandani, [3]Pratham Vishwakarma, [4]Mayur Chhalotre

B. Tech, Acropolis Institute of Technology and Research

Abstract: -

The purpose of similarity detection System is to automate the existing manual system by the help of computerized equipments and full-fledged computer software, fulfilling their requirements, so that their valuable data/information can be stored for a longer period with easy accessing and manipulation of the same. The required software and hardware are easily available and easy to work with.

Key-Words: *Similarity, plagiarism, copy, website.*

I. Introduction

A similarity detector detects similar or near duplicate occurrences of a document. The similarity detector determines similarity of documents by characterizing the documents as clusters each made up of a set of term entries, such as pairs of terms. A pair of terms, for example, indicates that the first term of the pair occurs before the second term of the pair in the underlying document. Another document that has a threshold level of term entries in common with a cluster is considered similar to the document characterized by the cluster.

II. Problem Formulation

The main objective of the Project on similarity detection system is to find the similarity percentage between the two text written in the text box or two documents that has been entered in the fields . The purpose of the project is to build an application program to reduce the manual work for the checkers to find if the text is copied from another source.

The Similarity Detection System to be implemented is divided into several modules. These modules just focus on Detecting Plagiarism rather than the other aspects of the System like server-client networking and security. Several words make a sentence, sentences makes a paragraph and paragraphs make a document. English language is very vast and rich in vocabulary, grammar and words. Using them wisely and appropriately can make a good thesis. But plagiarism can take place using vivid methods as mentioned above. Thus we have divided the modules with respect to the units in the English language such as words, sentences and paragraph. The processing of alphabets in this case is redundant. The system does include other smaller units as mentioned above but detecting plagiarism is the important task that we have to implement.

III. Literature Review

To identify plagiarism online tools like article checker, duplichecker, viper, turnitin, splat available. We focus on comparing the text document with article checker . To compare the source text with documents available online, (Automatic Plagiarism Detection Using Similarity Analysis 325 commercial tools do search for the exact words or sentences online. If those words are not available, they fail report plagiarism. We made a study pertaining to assignments collected by the students and asked them even to quote the exact reference from which they have copied. All these data's were stored in the warehouse and retrieved for processing. We could find by significantly removing the stop words and applying stemming, would identify plagiarism in a better way. Combining both the parameters leads us to even further benefits.

IV. Methodology

It is project about finding the similarity in percentage between the given text using the python cdifflib library and integrated in python flask prerequisite

1. Basic python
2. Python Flask [Web frame work]
3. Basic Html and Css to make website

To measure the content similarity we remove the stop words from the text document provided. This preprocessing is done in order to eliminate the relevancy among unwanted words. Stop words are ordinary or unusual words which occur in the document, which don't have significant meaning (e.g., connector words, conjunctions, single letter words). From a corpus of database, we eliminate such unwanted words [8]. We also eliminate special symbols that do not have significant part in text processing (e.g., ",", /,-, etc., in general symbols other than characters and numbers).

The system designed to detect similarity among text documents calculates content similarity among specified documents, after removal of stop words and stemming the terms. The similarity is estimated between the document samples (assignments) using various measures like cosine, dice, jaccard, hellinger and harmonic as the first vector (each term being normalized using the total number of words in the document) and t jh corresponds to the second vector. We performed analysis by normalized method, which useful in scenarios where we might not be interested in representing very high values. We perform the similarity analysis using the formulas listed above in equations 1-5. If two documents exactly plagiarized, then we have a value of 1, 0 otherwise, from the assignments in each group, we found that there is huge overlap among them. This is shown for all 5 measures in Table 2. From the values we infer that cosine is better to reflect the similarity of content among the peer groups. Columns marked 1-5 denote the respective values measuring the similarity adopting equations 1-5. We have three groups, for example A1- A2 denotes the relevance among group A1 and A2. Similarly the relevance among other groups is named as their group ID. From the values perused from Table 2, we find that there is strong inter relevance among the groups of similar assignments. Hence it is easier to finds the best assignment among the three using cosine metric. Then next task is to measure the similarity of plagiarised documents with original source referenced

V. Result Discussions

Similarity Detection system as described above, can lead to error free, secure, reliable and fast management system. It can assist the user to concentrate on their other activities rather to concentrate on the record keeping. Thus it will help organization in better utilization of resources.

VI. Conclusion

This complete project is has been built up using techniques from Design and Analysis of algorithms, and also from Theory of Computation to a certain extent. The best available techniques from these topics have to jotted down and used in an acceptable efficient way to detect plagiarism. Considering this layout as a standard any user or author can create his or her own standard for detecting plagiarism. Plagiarism Detection can cost huge amount of computation power and also there can be several ways in which the author can surpass the plagiarism test. Manual detection of plagiarism is always reliable. But certain amount of load can be reduced by the Plagiarism Detection System Standard given in this paper.

Acknowledgment

We express profound gratitude and heartfelt thanks to **Dr Kamal Kumar Sethi**, HOD IT, AITR Indore for his support, suggestion and inspiration for carrying out this project. I am very much thankful to other faculty and staff members of IT Dept, AITR Indore for providing me all support, help and advice during the project.

References

Reference Format for Journal Paper

[1] S.K. Gupta, L. Sharma, "*An Approach for Real Time Machine Learning Environment*", International Journal of Computer Science and Engineering, Vol.1**6**, Issue.**22**, pp.4-10, **2020**.

Reference Format for Book/Book Chapter

[2] G. Sharma, "A Proposed Novel Approach for Internet of Things Protocol Execution Environment", IJCSERT Publisher, India, pp. **142-145**, **2020**.

Reference Format for Conference Paper

[3] P.B. Verma, "A Proposed Approach for Real Time Cloud Environment using Machine Learning", In the Proceedings of the 2020 International Conference on Data Science and Engineering, India, pp.**542-545**, **2020**.