International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
471 – 473

Detecting Plagiarism In Academics Using Levenshtein Distance Algorithm And Semantic Similarity

Nikhil Ghode, Shubham Jadhav, Sampada Moon, Ashmina Khan, Shrutika Bhalkar
Student,Computer Engineering, Bapurao Deshmukh College of EngineerinWardha, India

Abstract— As the age of internet is growing the information is available to everyone at the click of their hand. So due to this copping of information and data have also increased at academic level. As it might be time consuming and complex work this can be made easy with some good algorithm, method and technique. In this paper we have use Levenshtein distance algorithm to detect the similarity or plagiarism between the student researches. We have also implemented other method and technique to find out the plagiarism. The main aim of the project to provide the high end but very simple software at the academic level. This can make student to do their own research and publish it rather than copping other work.

Keywords— *Plagiarism, Levenshtein distance, similarity,*

*****

I. INTRODUCTION

As the internet grows the problem of copping other work to fulfill our academic goals which include copying research paper, code, project etc which cut down all the work that is needed to be done. There are many studies done in recent year which prove that copying of other author work, project and there research has increased rapidly. Due to search increase in copying other work the originality of the individual have been compromised to very his level which also bring us to the point of academic integrity which play a vital role is the life of student and the academic goal. This small point of academic integrity shows the willingness of a student to achieve their goal.

The plagiarism can be defined by many definitions but a simple one or an understandable definition would be steeling other work and publishing that work under your name without the knowledge of original author to whom the work belongs. Plagiarism undermines the integrity of education and occurs at all level of scholarship. Now a day's plagiarism has become a problem at academic level.

A great way to avoid or minimize the plagiarism is to implement such easy but effective software at academic level like checking the research paper, code etc at academic level before publishing the work. The aim of this project to provide software that has different method applied in it to prevent academic plagiarism.

There are different types of plagiarism which make it more complex to detect the copied text. There are two main types of plagiarism which are intrinsic plagiarism which is also subdivided into four main parts which are near copies which can be defined as copying the text form file as it is without making any changes. The second one is translated which can be defined as translating the copied text in different language many time an than re-translating the same text in original language which will change the words and sentences as compared to original text and such translating and re-translating the text reduces the plagiarism but the work which they present remain copied.

The third one is idea this can simply define as stilling some one basic concept of their research and publishing it under your name. The idea can be copied for internet, article, or even form technical talk.

The fourth one which is some time very hard to detect is disguised. The document which is been used is recreated and redesigned by changing the sentence structure and using the synonyms for the word which can be difficult to catch and detect the plagiarism. Some time the document may contain all four types in it and it then become challenging to detect the plagiarism is the document.

There are different software present for free which can be used to change the word using synonyms of the same words. This software is design to catch such words which are converted using synonyms software. This software is capable to catch the document if all the four types of intrinsic plagiarism are present in it.

This software is also designed to find the relevant words line by line to detect the similarity in the words some time one word is used in many different styles throughout the document which express same meaning but different styles of expressing same words. Such words can reduce the plagiarism count in the document and a copied document will get accepted. Hence this project focuses on such words and sentences which can reduce plagiarism count and get expected. In this project we are using four to five methods and technique we are also using different algorithm which would help to catch different types of plagiarism in the document.

II. ANALZING DOCUMENT

In analyzing document the given document is analyze using four different approach or method. The four methods are as follow:-

- Tokenized Approach
- Detecting the Same Sentence

 ISSN: 2454-4248
471 – 473

- Cementing Based Analysis
- Levenshtein Distance

The first approach is to separate the document and words so that the same words can be eliminated from the document which will reduce the processing time of the document and will reduce the complexity of the document.

Such words which are used repeatedly in the document are called as stop words this words contain words like is, am, a, what, where, this, that, to, from, etc. This type of words provide body to the text and do not have an meaning individually so before doing any processing on the document the stop words must be removed which will reduce the complexity of the document and the software only go through the main words which define the contend of the document or of the research.

Some time the document may also contain delimiters which can be defined a set of different character which are included in document to specify the boundary. This different character are { } [ ] : ; + = & ^ # @ ! ( ) etc. sometime removal of such special character can reduce the complexity of the document. After removal of all stop words and special character from the document. The document contains main words which contain the idea of the research which is being published.

Now to process the document we have to Tokenized the words present in the document here we start using the first approach in analyzing the document which is Tokenized approach. In this approach the words are given numbers in the form of token which help to perform other method easily.

The second method is Detecting Same Sentence which is used to detect near copies in the document. When someone copy the same text from another document and paste as it is so that time Detecting Same Sentence method is very use full. In this method different line or sentence are compared to each other to detect the same sentence. This method is also use to detect same words which have different style but used for expressing same thought.

After every method or technique the plagiarism count will increase if copying of the text is detected. If the percentage of copied text is more than the threshold the process will not stop but also check on other parameter son we get a perfect result.

The third method is Cementing Based Analysis this method do one of the important work to finding the different words with same meaning this method is use to find the synonyms of the tokenized words. This method plays an important role in detecting plagiarism as now a day's people are using different API to get synonyms of same word. So to cover come this problem we are using WORD API which will help us to get list of synonyms for a given word. Using such API gives access to large set of words. Whenever a new word is detect which is not present is the database the WORD API will be called by the software and that word will get saved in the database of the software so if the same word is detect again the software will not have to call the WORD API ever time which will save lots of time and money as if we cross limit of free words than we have to pay for every word.

The Cementing Based Analysis also helps in detecting the words which are translated and re-translated many time in different language so they cannot be detected. This is one of the important processes in the Cementing Based Analysis and it is very helpful as people are creating different ways and method to not get caught.

The fourth approach is using Levenshtein Distance Algorithm. This algorithm can be defined as a measure of the similarity between two strings which can be referred as the source string (s) and the target string (t). The distance is the number of deletions, insertion or substitution required to transform s into t.

The Levenshtein distance algorithm is named after the Russian scientist Vladimir Levenshtein. The Levenshtein distance algorithm work as follow.

- If s is "test" and t is "test", then LD(s, t) =0, because no transformation are needed. The strings are already identical.
- If s is "test" and t is "tent" then LD(s, t) =1, because one substation (change "s" to " n") is sufficient to transform s into t

  The greater the Levenshtein Distance the more different the string are. The three main conditions in Levenshtein distance algorithm are as follow
- The cell immediately above plus 1: d[i-1, j]+1
- The cell immediately to the left plus 1: d[i, j-1]+1
- The cell diagonally above and to the left plus the cost: d [i-1, j-1] +cost.

III. CONCLUSION

In this project we have use different method, techniques and algorithm to detect the plagiarism of different kind using only one software. This project implements combined approach to detect the plagiarism using technique like cementing based analysis, Levenshtein distance algorithm and other approach. The academic plagiarism is growing rapidly now a day's so implementing such software will reduce the rapid growth of academic plagiarism

We kept our focus on the intrinsic plagiarism which is being use more this day's. Which contain copying of text as it is, translating and then re-translating the text so the software cannot detect the plagiarism, using different words which have same meaning etc. This all different way can be stopped using this software.

IV. REFERENCES

[1] E. Gharavi, K. Bijari, H. Veisi, K. Zahirnia" *A DeepLearning Approach to Persian Plagiarism*

 ISSN: 2454-4248
471 – 473

Detection" FIRE 2016 International Workshop,Kolkata,India.

[2] S. Vuković, B. Kopić, "Plagiranje - Svcučilište u Zarebu još nije uvelo sustav provjere radova" in Global, Zagreb:, vol. 21, pp. 24, Nov 2016.

[3] J. O. Shea, Z. Bandar, K. Crockett, D. Mclean, "A Comparative Study of Two Short Text Semantic Similarity Measures", *Artif. In tell.*, vol. 4953, pp. 172-181, 2015.

[4] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, J. Montmain, "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain", *J. Biomed. Inform.*, vol. 48, pp. 38-53, Apr. 2014.

[5] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", *J. Artif. In tell. Res.*, vol. 11, pp. 95-130, May 2011.

[6] R. Mihalcea, C. Corley, C. Strapparava, *Corpus-based and knowledge-based measures of text semantic similarity*, vol. 6, pp. 775-780, 2006.

[7] A. H. Osman, N. Salim, M. S. Binwahlan, "Plagiarism detection using graph-based representation", *J. Comput.*, vol. 2, no. 4, 2010.

[8] M. Chong, L. Speciali, R. Mitkov, *Using natural language processing for automatic detection of plagiarism*, 2010.

[9] A Maedche, S. Staab, "Comparing ontologies-similarity measures and a comparison study", *Proc. of EKAW-2002*, no. 408, 2002.

[10] P. Turney, *Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL*, 2001.