INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 2 | Issue: 3 | Year: 2020

Cognizance of Cribbing using Data Excavation

S Phani Kumar

[1] Assistant Professor,
Department of Information Technology,
Sasi Institute of Technology and Engineering (A), Tadepalligudem.
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

G Likhitha [2], CH Mounika [3], M S N Rama Lakshmi [4]

[2, 3, 4] IV B.Tech. Students,
Department of Information Technology,
Sasi Institute of Technology and Engineering (A), Tadepalligudem.
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

Abstract – *In the present era, we know that intellectuality of an individual cannot be gloomed. But, in this contemporary world, with the help of advanced technology, people made it easy to access or to copy one's intellectual work. This results in plagiarism. Finding the plagiarized content manually is not that easy, but using automation we can identify the plagiarism. In the field of data mining, data can be mined thoroughly through existing relations among them.*

Index Terms- SWR, Tokenization, Data Mining

1. Introduction

Plagiarism can be visually perceived as the work that is used by other individuals without intimation to the original author. Not only in academics, it is done in almost all fields. The case where George Harrison used the same tune became an earnest issue in academia is an example of facsimileing in different fields.

Plagiarism can be found in many forms. Some of the forms are as follows:

a) Ecumenical plagiarism: it means the copying if entire work of others.

b) Paraphrasing plagiarism: it signifies manipulating the pristine text with its lexicon and restarting it.

c) Verbatim plagiarism (copy paste): it occurs when an author copies the text of another author without utilizing special characters and representing the work as their own.

d) Mosaic plagiarism (patchwork plagiarism): it means getting resources from two or more sources and present results as their own.

e) Citing incorrectly: it signifies claiming others' papers work and information as our own.

f) Plagiarizing your work (self-plagiarism): it is nothing but reuse of our own work without citing that it was antecedently utilized. Republished is the best example of self-plagiarism.

This plagiarism can avail to minimize the time for completion of work but it kills your originality. Being incipient is a designation of discipline towards achieving your goal.

Data Mining is the field which avails to solve this quandary and can withal find valuable information obnubilated in the astronomically immense volumes. By utilizing data mining techniques the patterns and their regularities eliminate the arbitrariness present in it. It avails in detecting plagiarism and additionally amends the efficiency of the process.

2. Literature Survey

Eisa et al. [1] they analysed the state-of-the-art techniques used to detect plagiarism in terms of their limitations, features, taxonomies and processes. Their findings revealed that existing plagiarism detection techniques require further enhancements. Robust techniques for detecting plagiarism are required, especially now that authors can easily modify texts, change ideas from one form to another or hide correct references to prevent detection. Using structural features and contextual information integrated with semantic similarity methods can help to detect these types of plagiarism.

Chowdhury et al. [2] they used methods for plagiarism detection based on machine learning techniques and analysis on the pros and cons of these methods and finally highlighted a list of issues and research challenges related to this evolving research problem.

Kasprzak et al. [3] they discussed the computational cost of each step of implementation, including the performance data from two different computers. They have also presented several improvements to the system for detecting plagiarism. The result was the best-performing system in the PAN 2010 plagiarism detection competition.

Shkodkina et al.[4] compares plagiarism detection systems according to the list of criteria, compiled from

the most challenging and important features for users in Ukrainian higher educational institutions. It provides an overview of the most common and serious forms of plagiarism around the world. It can also allow to counter- attack plagiarism.

Zaher et al. [5] they defined plagiarism is generally defined as "literary theft" and "academic dishonesty" in the literature, and it is really has to be well-informed on this topic to prevent the problem and stick to the ethical principles. With the evolution of the internet and the need for information the plagiarism continues to be a concern problem to universities, teachers, policy-makers and students. So authors conclude that the need of plagiarism detection systems become very important issues and the use of plagiarism detection systems in E-Learning improve academic integrity, and also instances of plagiarism can be greatly reduced, if not eliminated, with the use of plagiarism detection systems.

Dahl [6] has been an increasing interest in plagiarism detection systems, such as the web-based Turnitin system. This exploratory study examines the attitudes of students on a postgraduate module after using Turnitin as their standard way of submitting work and getting feedback. Overall, students reacted positively towards the system. However, the study also found evidence of a group of students who were less positive.

3. Existing System

There are several implements to detect plagiarism they may be free or commercial. In those tools, the user can check plagiarism efficaciously but it sanctions only one search per day if he/she was unregistered. Some websites store the data and utilize it for other purposes. According to our survey, the user has to constrain the times of checking plagiarism for the smooth running of web servers, and the system must be reliable. Different users can use this system and get their results instead of with the same user.

4. Proposed System

According to the survey discussed above, we propose a system for detecting plagiarism that checks the similarity between documents. The main purpose of this system is to amass the information from users and check whether they are replicated. Tokenization and duffle techniques are used for detecting plagiarism in this system. And an algorithm is used for comparing the documents. It involves various phases:

a. Text preprocessing: In this phase, the accumulated information will be converted to a common format.

b. This preprocessing includes various stages like converting into lowercase\uppercase, removal of stop words, removal of punctuations, and replacement of synonyms.

c. Comparison of text: In this, the input data will be compared with local database files. Here we use word tokenization to compare with other files.

d. A report analyzing: Determinately, the percent of plagiarism found will be exhibited along with matches in individual files.

Conclusively, we propose a system that aims to highlight the copied data as well as show the percent of unique text found.

5. Methodology

By utilizing the text mining techniques plagiarism can be detected facilely and efficiently. To do this the following steps are:

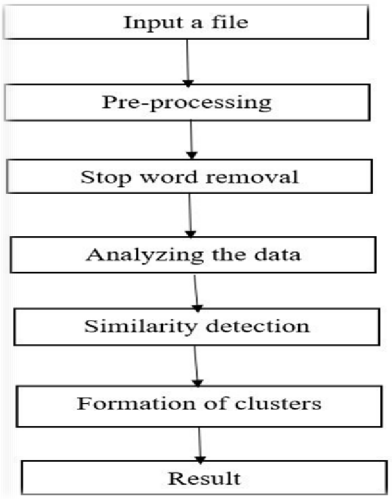

Figure 1: Architecture

a. Input a file: A file is taken as input in electronic format to detect the plagiarism automatically.

b. Pre-Processing: In this phase, all the files in the database including the input file are converted into a particular format i.e., it eliminates all the tables and images present.

c. Stop Word Abstraction: It abstracts all the pre-defined or common words that are present in the text i.e., the, is, in, and so on.

d. Analyzing the data: Here we analyze the data from all the documents in the database with the uploaded file.

e. Similarity detection: In this step, a comparison between the documents is performed with the avail of the duffle technique.

f. Formation of clusters: The plagiarised data and unplagiarized data are composed into separate clusters.

g. Result: Similarity report will be calculated in the form of a percentage, if the percentage is high it signifies the plagiarism content is high which admonishes to stay pristine.

An algorithm called Sequence Matcher is used for comparing every word in the file. For the abstraction of stop words, we import nltk.corpus package. In this, all the prevalent words will be compared with user-uploaded text. We can withal integrate user-defined text in the stop words directory. Utilizing set command we can fine-tune the words to a particular language.

SEQUENCEMATCHER

It is a class available in python which is used for checking similarity between two pairs of input data. It checks various files at a time and shows the information difference in various formats. The main of this is to find the longest contiguous matching subsequence which has no stop words.

This applies iteratively to all the letters from left to right. Its best-case time is linear because it has expected behaviour depending upon the elements in the file. By using the technique stop word removal it automatically treats some items as junk. These checks how many times a word appears in the text.

In this algorithm, we use a duffle module which is the tool for working with differences between files. It generates a total percentage of copied text in the document. To compare the bodies of text, here Differ class is required. This class produces a human-readable text format. After splitting up the lines we go comparing which can be done using command *compare ()*, it takes two arguments.

If there is any noise in the text we can remove it using the n*diff()* function. By default, it removes tab and other spaces. When a user wishes to display only the modified lines instead of the complete document then we can use *unified_diff()* function.

Different users use different technologies in their daily routine. So keeping in mind they also implemented *Html_Diff()* command for producing output in HTML format. For Example, if we take a document as input and it is compared with other files then it produces output as:

This below shows only the basic form of difference between copied and changed text. The copied is represented as red whereas the changed text as green in color. Finally, To measure similarity between two strings: *difflib.SequenceMatcher (None, "abbb", "abc").ratio ()*

To find longest match between substrings we use:
var.find_longest_match(initial,last,next_initial,next_last)

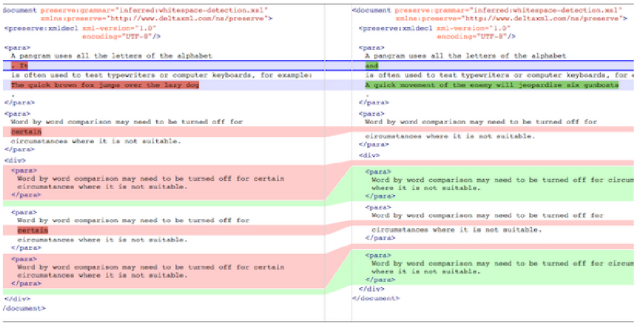

Figure 2: Difference

6. Results

The results of the proposed system must be accurate and reliable to the user. This will show in the form of a progress bar. This bar shows two shades: red i.e., plagiarism found, green i.e., unique text along with percent in center of the circle.

Figure 3: Result

In the above diagram, we can optically discern that the total plagiarism found is about 65% which is in red. Whereas the remaining 35% is a unique text which is in green. The sequence matcher engenders this percent by calculating the percent of every file in the local database.

Figure 4: Report

In this diagram, the total report will be generated. This shows the percent of plagiarism found in every file. And the file that was uploaded by the user will be shown on top of the window.

7. Conclusion

We can conclude that plagiarism is an offense whether it's done intentionally or unintentionally. It is very crucial to detect plagiarism for ameliorating cognizance in different fields of work.

Here authors propose a system that can detect facsimiled text from different sources present in the local database. It is very essential to find the replicated text to stop the breach of the pristine objective.

References

[1]. T. A. E. Eisa, N. Salim, and S. Alzahrani. Existing plagiarism detection techniques: A systematic mapping of the schoolarly literature. *Online Information Review*, 39(3):383–400, 2015.

[2]. H. A. Chowdhury and D. K. Bhattacharyya. Plagiarism: taxonomy, tools and detection techniques.*arXiv preprint arXiv:1801.06323*, 2018.

[3]. J. Kasprzak and M. Brandejs. Improving the reliability of the plagiarism detection system.*Lab Report for PAN at CLEF*, pages 359–366, 2010.

[4]. Y.M. Shkodkina and D. Pakauskas. comparative analysis of plagiarism detection systems.2017.

[5]. A. Bin-Habtoor and M. Zaher. A survey on plagiarism detection systems. *International Journal of Computer Theory and Engineering*, 4(2):185, 2012.

[6]. S. Dahl. Turnitin R: The student perspective on using plagiarism detection software. *ActiveLearning in Higher Education*, 8(2):173–191, 2007.

Author's Profile

Mr. S Phani Kumar received his Bachelors Degree in Computers Science from Acharya Nagarjuna University in 2005, his Master of Computer Applications in 2008 and his Master of Technology in Computer Science and Engineering from Jawaharlal Nehru Technological University Kakinada in 2011.

He had worked in different designations in various colleges under different universities. He had Published 4 papers and he is working on Data Sciences, Data Analytics .Presently he is working as an Assistant Professor in Sasi Institute of Technology and Engineering, Tadepalligudem, Andhra Pradesh.

G Likhitha, at present pursuing IV B.Tech at Sasi Institute of Technology and Engineering, Tadepalligudem, India. She got certified in DBMS, Block Chain Technology, and Data Structures.

CH Mounika, at present pursuing IV B.Tech at Sasi Institute of Technology and Engineering, Tadepalligudem, India. Her interesting subjects are IoT and Data Mining.

M Rama Lakshmi, at present pursuing IV B.Tech at Sasi Institute of Technology and Engineering, Tadepalligudem, India. She is active member in various college clubs like Web Technologies, IoT.