Plagiarism Detection in Students' Assignments

[1]Aishwarya Abhaykumar Bhavsar, [2]Kalyani Madhukar Kude, [3]Aditi Shirish Adakmol, [4]Sayali Bhaskar Barhate

BE Students
[1]Department of Computer Engineering
[1]SSBT's College of Engineering and Technology, Bambhori, Jalagon, India

***Abstract* -Plagiarism means "Passing of someone else's work, whether intentionally or unintentionally, as your own for your own benefit". Plagiarism is on the increase in higher education. In practical and theory assignments there are various ways in which students attempt to cheat .The common method is coping contents from other students and making minimal changes in it. By using algorithm for plagiarism detection when multiple student's files exist and allowed assignment is present. The proposed system is tested on assignments. Using this technique, subject teachers received the result regarding the originality of assignments written by student, which is whether the sources had been properly acknowledged and consequently also had the desirable effect of relieving some of the burden for teaching staff in checking student work prior to submission.**

I.Introduction

The issue of plagiarism is not new, however increased ease of access to electronic material via the web is always a concern among the academic community. Plagiarism is using someone else's work without indicating that it is not owns work or crediting the original author. Plagiarism has a number of negative acts on education. Firstly, it limits the thought, research and critical thinking involved in developing an original paper or report, which negatively impacts the overall educational experience of a college/university student. Second, it damages the relationship between peers and instructors, due to the loss of trust. Finally widespread systematic plagiarism can damage the reputation of the academic institutions and devalue their awarded degrees. Therefore it is vitally important to detect cases of plagiarism and apply appropriate punishments in order to deter students.

II. Related Work

There are numerous tools or systems available to find if there is any plagiarism between the files or not. DANIEL R. WHITE and MIKE S. JOY provides one "Sentence-Based Natural Language Plagiarism Detection System"[1]. In tongue, the sentence is thought of in concert of the building blocks for the communication of concepts. With this idea as a beginning point, they were planned the Novel rule to check documents at a sentence level, as plagiarism are doubtless to occur by transforming a supply text on a sentence by sentence basis. The rule is meant for a radical pair-wise comparison of a group of documents, whether they comprise a specific category essays, the comments from a group of programming assignments, or a piece of writing alongside sources that it's doubtless to possess plagiarized. The results generated highlight places wherever 2 documents are terribly similar, in a manner that makes it simple for a teacher/academic to determine whether an investigation of alleged plagiarism is worth pursuing. This system contain three main stages as follows:

Preprocessing

Thepreprocessing stage of the rule is to scan within the documents that are being compared and break down them into Document objects that contain an inventory of Sentence objects, which, in turn, every contain an inventory of words that were found within the original sentence within the supply text. The list of words contained during a Sentence object is then subject to 3 filters. First, all words are reborn to minuscule to save lots of time in later comparisons. Second, a list of words that are thought of too common to be helpful are never keep within the processed type of the document. This list is passed to the computer program as a parameter of the detection engineand will usually include words that add no aiming to the sentence, like the 'a' or "that".

Documents

At this stage, alongside the first source-texts there'sa set of objects containing a specialized, processed kind of the originals. These documents are compared pair wise by examination each sentence within the one document to each sentence in the different. The comparison is finished by computing a similarity score, supported the word count metric elaborate in Culwin and Lancaster [2003]. The score is that the average similarity between the sentences, computed as a function of words in common and therefore the lengths of the sentences being compared.

Document Scores

A score is appointed to every document therefore that pairs of documents is compared to everydifferent. Those with high scores are then the ones that it might be most helpful to look at. There are two possible ways of calculating the document overall score in the software. The first is to assign the score because the price of the similarities between it and its most similar document. The second is to multiply the whole similarity score by the percentage of the whole that came from its most similar document.

Novel Algorithm

Document[] docs = readDocsFromDisk();for each Document, i, in docs {for each document, j, following i in docs{compareSentences(docs[i], docs[j]);
}
}
compareSentences(Document doc1, Document doc2){
for each sentence,i, in doc1{
for each sentence,j, in doc2 {
int common = number of shared words;
int score=similarityScore(i,j,common);
if(score > SIM THRESHOLD ||
common> COM THRESHOLD)
storeLink(sent1, sent2, score);}
}}

III. Proposed System

As the DANIEL R. WHITE and MIKE S. JOY gives the Plagiarism Detection System which detects plagiarism between the text documents on the basis of sentence based. To detect the sentence based plagiarism they gave the algorithm i.e Novel algorithm. Which contains 3 main steps: **1. Preprocessing:** In this step all the unwanted letters and data are remove for ex."a", "the", "that". **2. Comparing Documents:** At this stage, alongside the original source-texts there is a set of objects containing a specialized, processed form of the originals. These documents are compared pair wise by comparing every sentence in the one document to every sentence in the other. The score is the average similarity between the sentences, computed as a function of words in common and the lengths of the sentences being compared. **3. Document Scores:** A score is assigned to each document so that pairs of documents can be compared to each other. Those with high scores are then the ones that it would be most useful to examine. There are two possible ways of calculating the documents overall score in the software. There is to assign the score as the value of the similarities between it and its most similar document. The second is to multiply the total similarity score by the percentage of the total that came from its most similar document. But in the existing system there is some drawback i.e it only compare the documents on the basis the sentences. If some student do something like they break the one big sentence into one or more sentences then this is not consider as plagiarism in the existing system. Or student can make the single sentence by using two different sentences by using conjunction. So to overcome this drawbacks this system adds some additional steps to the existing algorithm. The system detect the plagiarism by the following techniques**:**

1. First it check the validity of the assignments uploaded by student's i.e the assignments is valid or not. **2**. After separating the valid and invalid assignments it checks for the similarity between two assignments by comparing each with another assignment. **3**. After comparing two documents if the similarity score is greater than the threshold then the link between those two assignments is shown as a result of system.

IV. Architecture

As this system is standalone its architecture is very simple. As shown in the figure given below, the top layer is client layer at this layer JAVA is used to provide user interface to user of this system. At the middle layer which is application layer JDBC Drivers are used. The middle layer is nothing but the interface between the top and bottom layer. At the bottom layer the MySQL is used.

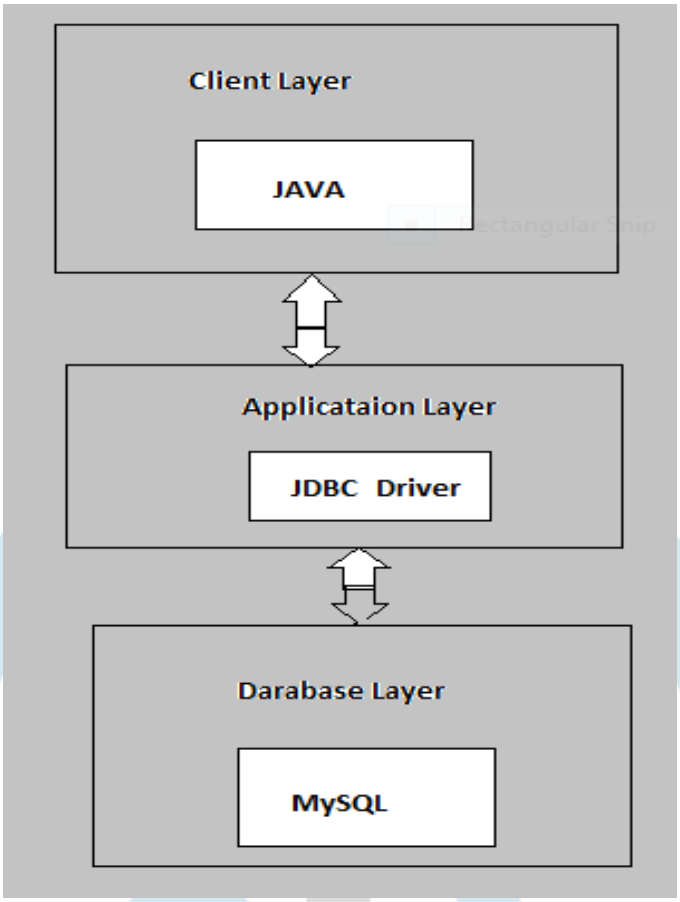

Fig: Architecture of the System

***Proposed Algorithm**:*

1] Check for validity in a student assignment by:

i) Count the number of questions in a model answer and student assignment file.

ii) If number of question match, match each question from model answer with question in student assignment. If number of question don't match skip and goto next assignment. If each question matches continue otherwise skip the assignment and goto step (i).

iii) Each keyword already entered is compared with each other word. If match counter<80% then the file is declared as invalid and gotostep(i) with next assignment.

2] For all valid assignment, each valid assignment is check with all the other remaining assignment on word by word basis.

3] If similarity percentage>75%, then the similarity percentage between two files is display.

VI. Result

The PDISA system is fed with student's assignments by respective students along with the model answer files by the respective subject teachers. Ason output, the PDISA systemprovides the list of invalid assignments along with the reason for the same. Also, it displays the listvalid files and renders the percentage of similarity between the contents of valid files if it isbeyond a certain threshold level.

VII. Conclusion and Future work

The Plagiarism Detection System has achieved the time savingfactor, eliminate the overhead of class teacher, irritation of students due to waiting for checking the assignment, accessibility of assignment of students by all teachers without going to respective class teachers.The system currently supports the text files for the natural languages. The systemcompareamong the student's submitted-assignments and give the rate of plagiarism among them.

As the future work, to implement the work which contain source code files and detect the plagiarism within this files. A further improvement would be an archive of submitted files that it could then check against new submissions. This would allow it to catch plagiarism across different years or from different courses.

References

[1] Daniel R.White and Mike S.Joy "sentence-Based Natural Language Plagiarism Detection", "University of Warwick", August 23 2005.

[2] Timothy C. Hoad and Justin Zobel "Methods for Identifying Versioned and Plagiarized Documents","RMITUniversity",February 1, 2003

[3] T. Lancaster, Effective and Efficient Plagiarism Detection, South Bank University, 2003.

[4] KrisztinMonostori, ArkadyZaslavsky "Using the MatchDetectReveal System for Comparative Analysis of Texts", December 7, 2001.

[5] T. Citron and P. Ginsparg, Patterns of text reuse in a scientist corpus, Proceedings of the National Academy of Sciences, vol. 112, no. 1, pp. 2530, 2015.

[6] M. Errami, J. Hicks, W. Fisher, J. W. D. Trusty, T. Long, and H.Garner, vu - a study of duplicate citations in medline, Bioinformatics, vol. 24, pp. 243249, 2008.

[7] R. Boisvert and M. Irwin, Plagiarism on the rise, Communications of the ACM, vol. 49, pp. 2324, 2006.