SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 3 | March 2021 - Peer Reviewed Journal

PLAGIAT : A CODE PLAGIARISM DETECTION TOOL

V. Lakshmi Lalitha[1]
[1]*Dept. of Computer Science & Engineering, K L University, Vijayawada, India, 522502*

K. Sree Varsha[2]
[2] *Dept. of Computer Science & Engineering, K L University, Vijayawada, India, 522502*

S. Mahesh[3]
[3] *Dept. of Computer Science & Engineering, K L University, Vijayawada, India, 522502*

R. Sri Lekha[4]
[4] *Dept. of Computer Science & Engineering, K L University, Vijayawada, India, 522502*

B. Sri Lakshmi Navya[5]
[5] *Dept. of Computer Science & Engineering, K L University, Vijayawada, India, 522502*

V. Niteesh Kumar[6]
[6] *Dept. of Computer Science & Engineering, K L University, Vijayawada, India, 522502*

ABSTRACT

The plagiarism and it's detection has become a major problem in today's Edu – tech world. The advantage of being unique is not less important than any other valuable property. Accordingly, the plagiarism itself is a key factor in terms of the development; as it contributes to the uniqueness and differentiation through experimentations between two python source codes or two java source codes. In this paper, we discussed about a new model for the source code detection which works on the concepts of Machine Learning. The machine learning algorithms are profound to get the most accurate results and thereby Naïve Bayes algorithm, K – Nearest Neighbor and ADA Boost Meta Learning Algorithm are deployed in a combined manner. The final source code can be compiled and tested using the designed model to get the desirable results. The conception, interpretation, experimentation and results are discussed elaborately.

KEYWORDS : *Plagiarism, Source Code, Machine Learning, Algorithms, Experimentation.*

SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 3 | March 2021 - Peer Reviewed Journal

INTRODUCTION

The plagiarism is meant to be the replicating activity of one's words or any other material. It is also termed as paraphrasing. It is generally an immense challenge and a tough task to familiarize the content with minimum values of plagiarism. Since the task of manually detecting plagiarism in a large document database is very tedious and time-consuming, efforts are continuously being made to automate the process. There are two main categories of techniques for source code plagiarism detection: attribute-counting-based and structure-based comparison. Attribute-counting-based techniques consider the number of occurrences of different attributes in a file following certain criteria and different similarity measures are used to obtain the similarity between files.. Attribute-counting algorithms are simple to implement and execute faster. Structure-based methods, on the other hand, are more reliable since they gather details of program structure for comparison of programs. However, structure-based methods are computationally expensive. Hence, the aim of this research is to develop a new strategy which combines the advantages of both the categories.

LITERATURE SURVEY

Farhan Ullah., et al., worked on the solving of the plagiarism problem in terms of the academicians. A new technique was generated to detect the cloned or plagiarized content between the C++ and java source codes [1]. Yahya Ali., et al., worked on the frame work design to detect the plagiarism in Arabic language. The detection of the plagiarism is a difficult task due to the structure and principles of the Arabic language. The work is intended to the detection of the plagiarism of the Arabic documents by using the logical representation as paragraphs, words and sentences [2]. Saed Alrabee., et al., proposed a technique known as SIGMA to identify the reused functions in binary code in terms of matching traces of a binary code in its novel representation form named as the Semantic integrated Graph (SIG). The results exhibited that the opted approach has yielded promising results [3]. Upul Bandra., et al., worked on the new plagiarism detection method which is developed based on the attribute counting technique. The proposed method is unique from the other methods as it is made based on a meta – learning algorithm in order to enhance the accuracy rate of the plagiarism detection system [4]. Berry M W., et al., explained the content on understanding the search engines. The search engines are a way vague topics that we're unfamiliar about. The search engines are deployed in such a way that they supply the person with the information they want instead of what they asked for [5]. Xin Chen., et al., worked on a metric based Kolmogorov complexity among the information shared between two sentences. The metric model designed is applied to calculate the amount of the shared information among the programs. A new practical system termed to be software integrity diagnosis system was implemented to work on the application effectively [6]. G Cosma., et al., extended the work on surveying over teaching of programming on computing courses by the UK academicians. The survey was intended to understand the context of source code plagiarism [7]. Sabestiaan de Klerk., et al., discussed about the status of MBPA at a point of time. It is used as an assessing method that incorporates the multimedia like video, audio, graphs etc.; to simulate the work environment of the student [8]. Muhammad Farhan., et al., presented a methodology of qualitative assessment along with an algorithm for measuring and the correlation. The results are interpreted using statics generated in terms of the descriptive and graphical methods [9]. Muhammad Farhan., et al., worked on the comparison of the computational performance of developed algorithms with different video bites along with their processed frames per second by analyzing it as per their corresponding bins. The mean, median and max values are also compared for the processed frames [10].

OBJECTIVE OF THE WORK

The grail of the presented work is to concept and design an new model of the plagiarism detection using the machine learning algorithms. The final detection tool is used to detect the plagiarized content between two python source codes or two java source codes.

DETAILED PROCEDURE

The detection of the plagiarized content between two source codes is the most difficult task to do manually. Thereby an integrated plagiarism detection tool was deployed using the machine learning algorithms. This tool can be deployed to find the cloned content between the source codes with much ease and less effort. The similarity in the source code is the crucial factor to be counted in terms of the academics and the software industry. The plagiarism in the source code, the similarity content can be suspicious. These types of codes share the similar type of code segments and are differentiated in the logic, functionality and method. This intent of similarity is availed as the substantial evidence in plagiarism of a source code. There is a specific feature for every language as they differ in their syntax and sematic structure. When a student clones the logic of a source code and converts it to a required format in the targeted language, that individual never gets the ability of thinking on his own. There are even some tools available for the direct conversion from one language to another

SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 3 | March 2021 - Peer Reviewed Journal

language.

The plagiarism detection tool was concepted using the ML algorithms rather than the old conventional methods which include the attribute counting, graph – based analysis and structural methods. So, thereby we used the algorithms like Naïve Bayes Algorithm, K – Nearest Algorithm, and ADA Boost Meta Learning Algorithm. These algorithms all together in a common model produce the best results when compared to the other existing models

a. Naïve Bayes Algorithm : It is a classification algorithm intended to work on the Bayes theorem of probability. It is independent from the general conventional predictors. It is mainly used to build a model containing large sets of data.
b. K – Nearest Neighbor Algorithm : the KNN – Algorithm is a non – parametric method intended for the classification and regression operations. An input when given will be combined with the K – closest training examples in the feature space. The property values which are counted as the output values will be an average one in the result.
c. ADA Boosting : The adaptive boosting is a meta algorithm and also considered as a machine learning algorithm. The ADA algorithm is used to enhance the performance of the detection tool to get the desirable results. The group of outputs all together is said to be the final output of the file based on the classifier. Upon adjusting the configurations and parameters, they achieve the optimal performance over a data set. Thereby, the ADA meta boosting algorithm is termed as the best – out – of – the – box classifier.

EXPERIMENTATION

The conception and design are adhered to the machine learning algorithms. The aim to clear the higher phases of plagiarism. The entire detection tool was designed using the python programming language. The machine learning algorithms combined with the python programming language are termed as the best collaboration so as to get the desired results of detecting the plagiarized content. As a single algorithm alone cannot produce the final output in a desired way, the combination of the machine learning algorithms together are deployed in this. There are two main categories of techniques for source code plagiarism detection: attribute-counting-based and structure-based comparison. Attribute-counting-based techniques consider the number of occurrences of different attributes in a file following certain criteria and different similarity measures are used to obtain the similarity between files. Structure-based techniques derive information on program structure and obtain similarity scores based on this information. Attribute-counting algorithms are simple to implement and execute faster. Structure-based methods, on the other hand, are more reliable since they gather details of program structure for comparison of programs.

RESULTS AND DISCUSSIONS

The new plagiarism detection tool has been concepted and designed in a way that is identifies the cloned content between two source codes related to the java and python programming languages i.e. either between two java source codes or two python source codes. The final model deployed using the Naïve Bayes Algorithm, K – Nearest Algorithm, and ADA Boost Meta Learning Algorithm all together contributed for the best results. The codes are compiled for the corpus demo, corpus demo example and the main code along with the paraphraser & view code.

SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 3 | March 2021 - Peer Reviewed Journal

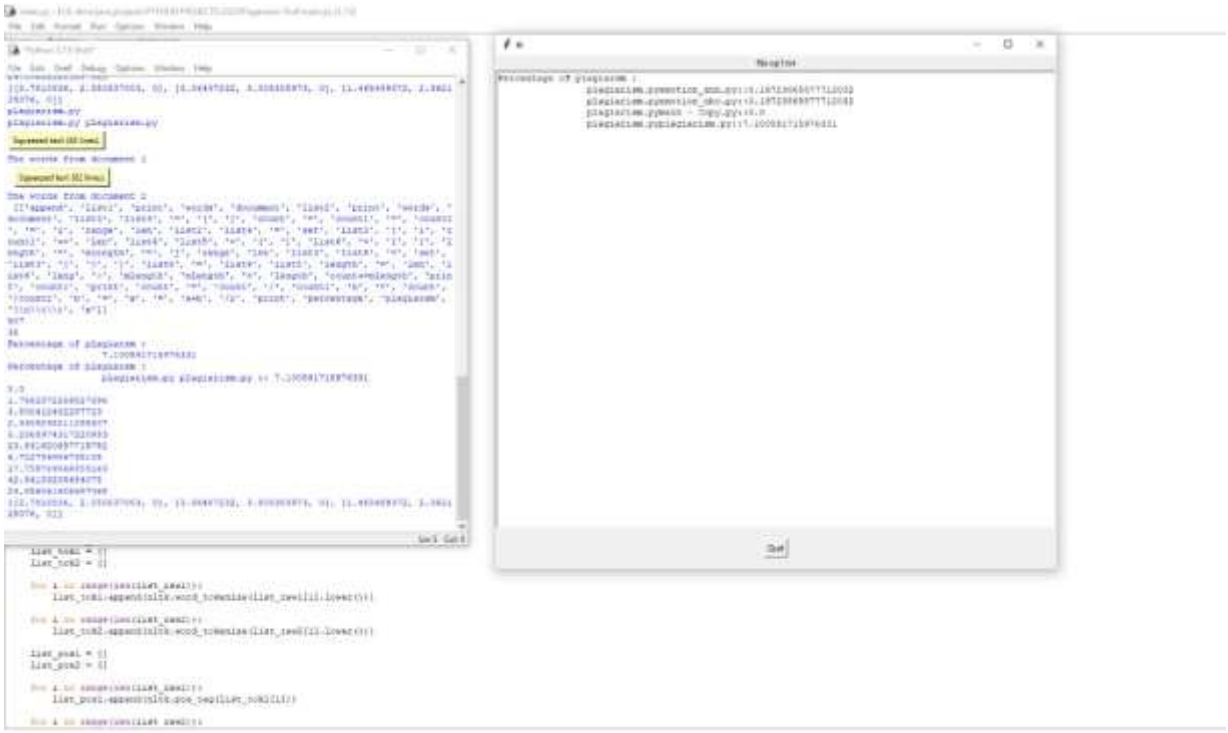

Fig. 1 : The end result when compiled and tested

The result is generated as displayed in the picture. The output is appreciable and in the required form. When the inputs are given in the required format, the code analyses the inputs. The analyzed input is processed accordingly and then the output is generated in the percentage form of the plagiarism. The amount of content cloned or plagiarized is denoted in the result in terms of percentage. Thereby making it the optimum model of computing the plagiarism in a source code.

CONCLUSIONS

The detection of the plagiarism has been bigger challenge in the student's assessment of the programming assignments to be submitted over the latent semantic analysis method. The source code which is plagiarized can be a modified version or the indirect modifier of the original code in different programming languages. The user gets the code copied to the source and then changes it accordingly to the targeted language. The new model for the source code plagiarism detection, which uses the concepts of machine learning in order to fight with the higher phases of plagiarism was designed. The main language for the conception and design is preferred to be python. Conventional methods like structural methods, attribute counting method and graph – based analysis don't produce the results with accuracy. Machine learning algorithms produce most accurate results with continuous learning from the training modules. The three algorithms used are Naive Bayes Algorithm, K – Nearest Neighbor and ADA Boost Meta Learning Algorithm. Since, no single algorithm can produce the result with accuracy, thus combining the algorithms help to produce results more accurately. The designed code has been compiled and tested accordingly to get the desirable results. The conception, interpretation, experimentation and results are discussed elaborately.

REFERENCES

1. *Farhan Ullah., Junfeng Wang., Muhammad Farhan., Sohail Jabbar., Zhiming Wu., Shehzad Khalid., (2018). Plagiarism detection in student's programming assignments based on semantics : multimedia based e – learning based smart assessment methodology. Multimedia tools application, Springer Publications.*
2. *Abdelrahman YA, Khalid A, Osman IM (2017) A method for arabic documents plagiarism detection. Int J Comput Sci Inf Secur 15(2):79 .*
3. *Alrabaee S et al (2015) Sigma: a semantic integrated graph matching approach for identifying reused functions in binary code. Digit Investig 12:S61–S71 3. Bakker T (2014) Plagiarism detection in source code. PhD dissertation, Universiteit Leiden, 7, pp 1–35 .*
4. *Bandara U, Wijayrathna G (2012) Detection of source code plagiarism using machine learning approach. Int J Comput Theory Eng 4(5):674.*
5. *Berry MW, Browne M (2005) Understanding search engines: mathematical modeling and text retrieval. SIAM.*
6. *Buddrus F, Schödel J (1998) Cappuccino—A C++ to Java translator. In Proceedings of the 1998 ACM symposium on Applied Computing. ACM.*

SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 3 | March 2021 - Peer Reviewed Journal

7. *Chen X et al (2004) Shared information and program plagiarism detection. IEEE Trans Inf Theory 50(7): 1545–1551.*
8. *Cosma G, Joy M. (2006) Source-code plagiarism: a UK academic perspective.*
9. *Cosma G, Joy M (2012) An approach to source-code plagiarism detection and investigation using latent semantic analysis. IEEE Trans Comput 61(3):379–394.*
10. *de Klerk S, Eggen TJ, Veldkamp BP (2014) A blending of computer-based assessment and performancebased assessment: Multimedia-Based Performance Assessment (MBPA). The introduction of a new method of assessment in Dutch Vocational Education and Training (VET). Cadmo, pp 39–56. doi:https://doi. org/10.3280/CAD2014-001006.*
11. *Farhan M, Aslam M, Jabbar S, Khalid S (2016) Multimedia based qualitative assessment methodology in eLearning: student teacher engagement analysis. Multimed Tools Appl 77:4909–4923.*
12. *Farhan M, Aslam M, Jabbar S, Khalid S, Kim M (2017) Real-time imaging-based assessment model for improving teaching performance and student experience in e-learning. J Real-Time Image Proc 13(3):491–504..*
13. *Farhan M, Jabbar S, Aslam M, Ahmad A, Iqbal MM, Khan M, Martinez-Enriquez AM (2017) A real-time data mining approach for interaction analytics assessment: IoT based student interaction framework. Int J Parallel Prog 12:1–18.*
14. *Farhan M et al (2018) IoT-based students interaction framework using attention-scoring assessment in eLearning. Futur Gener Comput Syst 79:909–919.*
15. *Jhi Y-C et al (2015) Program characterization using runtime values and its application to software plagiarism detection. IEEE Trans Softw Eng 41(9):925–943 Multimed Tools Appl.*