Plagiarism Detection through Data Mining Techniques

Rajashekar Nennuri¹, M Geetha Yadav², M Samhitha³, S Sandeep Kumar⁴ G Roshini⁵
Dept. of CSE, Institute of Aeronautical Engineering¹²³⁴⁵, Dundigal - Hyderabad
rajasekharnennuri@gmail.com¹ geethayadav22@gmail.com²
samhitham2000@gmail.com³ sandymoses741@gmail.com⁴
gaddam.padmasrinivas@gmail.com⁵

Abstract
Plagiarism is a problem that is becoming more prevalent as technology advances and the use of computer systems grows in comparison to previous generations. Plagiarism is the unauthorized use of another person's work. Since manual plagiarism detection is difficult, this method should be automated. Plagiarism detection can be done using a variety of methods. Some of the research focuses on intrinsic plagiarism, while others focus on extrinsic plagiarism. Data mining is an area that can assist in both detecting plagiarism and improving the reliability of the operation. Plagiarism can be detected using a variety of data mining techniques. Text mining, clustering, bi-grams, tri-grams, and n-grams are some of the techniques that can assist with this. In this paper we will use the data mining techniques to increase the efficiency of detection of plagiarism.

Keywords: k-nearest neighbor’s algorithm (KNN), Plagiarism Detection, Efficiency, Data Mining.

1. Introduction
In modern world with the advent of the internet and the widespread availability of computers around the world, it is now much easier to access other people's work, resulting in plagiarism. Plagiarism is described
as the act of using another person's work without their permission. With the advent of technology, the use of computers has increased dramatically, and they can now be found in colleges, institutes, and businesses. Students' assignments are increasingly being sent electronically. Since e-forms are simple to use, it is appropriate for both teachers and students, but it opens the door to simple plagiarism.

With the widespread availability of information around the world, it is very convenient to copy data from a variety of sources, such as the internet, magazines, online books, newspapers, and so on, and paste it into a single work without giving any credit. Sources should be acknowledged. Students are unable to learn as a result of these behaviors. As a result, there is a need to identify plagiarism in order to increase and enhance student learning. Plagiarism may occur in a variety of fields, including books, source code for programs, research articles, and so on. Furthermore, there may be instances where students from various institutes copy data from the internet, various books, journals, and other sources without citing any sources. The problem isn't only limited to written text; it also includes programming codes. Different small pieces of code are copied from the source and used as required without referencing the original authors. According to a survey conducted by the University of California, Berkley, the percentage of plagiarism has increased over the last four years. The percentage for the years 1993-1997 has been raised to 74.4 percent. According to other reports, it is estimated that over 90.0 percent of high school students are involved.

2. Proposed Model

This project's proposed model is to identify plagiarism. With the advent of technology, the use of computers has increased dramatically, and they can now be found in colleges, institutes, and businesses. Students' assignments are increasingly being sent electronically. Using the KNN method to increase plagiarism detection performance results in high efficiency.

2.1 KNN (k-nearest neighbors algorithm) Method

One of the simplest machine learning algorithms for pattern recognition is the k-Nearest Neighbor Algorithm. In certain cases, the k-Nearest Neighbor (kNN) algorithm is used. In most pattern detection implementations, it works admirably. In the kNN algorithm, "k" is a parameter. Several tests with different k values are needed to select the right k value for the kNN algorithm. K-NN assigns the "k" closest neighbors in the text data set's neighbors kNN simply memorizes all of the documents in the training dataset and compares the assigned test document to them. When we do classification, certain primary words in a text can be assigned to the wrong group or wrongly categorized out of a training package. As a result, establishing such a training set is virtually impossible. There are some disadvantages, such as the need for relevant documentation and contents. So, by using a specific text classifier, k-NN can be used to efficiently and successfully solve the above disadvantages. And training set has its own unique characteristics. To define and distinguish the paper material, they use taxonomic criteria and their own definition stages. The k-NN algorithm has been used in a variety of fields so far. Because of its versatility and precision, it is used in text categorization.

3. Methodology

a) Collecting data and transferring files: Electronic assignments are stored as three separate data sets. Since each task has a different format, they are all translated to the same format.
b) Pre-processing: This is a crucial phase in the detection of plagiarism. In this stage, data is transformed into a format that can be used in the detection process. The papers submitted are in a variety of formats, including lower and upper case letters. As a result, all records are translated to the same format, i.e. lower case, to eliminate the sensitivity. Figures, numbers, and images are also removed.

c) Train and Deploy the Model: Data from the train/test function should be uploaded. Create a KNN model for binary classification and a training script. SageMaker is used to train and deploy the model. Evaluate the classifier that has been deployed.

d) KNN Model usage: The KNN algorithm selects a number of K neighbors, finds the term we need, completes the data point correlation, and brings us expected data that is observed.

e) Predict: After the data has been identified, it is now forecast so we need precision, and this accurate result will help us figure out how to improve the KNN algorithm process.

![Flow chart of Methodology](image)

**Figure 1:** Flow chart of Methodology

### 3. Implementation And Result

The implementation of our algorithm in the Java platform is discussed in this section. We begin by parsing the text into its constituent data using the parsing technique. It returns a series of tokens that can be used to align patterns and determine whether two strings are identical.

#### 3.1 Analysis

Computing similarity functions, which quantify how close a given text file is to an original source text, is one way we could go about detecting plagiarism. We are free to create as many features as we wish, but we must specify a few as illustrated in this article. Researchers developed features called containment and longest common subsequence in this article.
3.2 Correlation

A linear association between two variables can be found using correlation analysis. A similar method for determining the relationship between an outcome variable and one or more risk factors or interfering variables is regression analysis.

Output

Plagiarism is detected in this step.
Knn Accuracy

We looked at the k-NN clustering process, which clusters the string and matches each term with its neighbors. We considered the problem of plagiarism identification and attempted to make it simple to locate. To compare the String, we used a method. We used counter to count how many times each string in the text files was paired.

Conclusion

Plagiarism is described as the act of reproducing previously published material in a different format. Plagiarism is now present in nearly every area of human activity, so identifying and detecting it is a top priority. In the case of plagiarism, some preliminary findings suggest that using mixed machine learning approaches improves efficiency in general. The hybrid system, on the other hand, does not always yield better results. As a result, we created a process that uses the k-NN machine learning approach to increase
efficiency. When all methods in this field are compared, we can infer that the k-nearest neighbor approach is extremely useful in pattern recognition as well as detecting plagiarism by locating copied dataset.

References

1. Imam Much Ibnu Subroto and Ali Selamat, "Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine," TELKOMNIKA, Vol.12, No.1, March 2014, pp. 209-218.
2. Upul Bandara and Gamini Wijayrathna, "Detection of Source Code Plagiarism Using Machine Learning Approach," International Journal of Computer Theory and Engineering, Vol. 4, No. 5, October 2012, pp.674-678.
3. Salha Alzahrani, Naomie Salim, Ajith Abraham, and Vasile Palade, "iPlag: Intelligent Plagiarism Reasoner in Scientific Publications," IEEE World Congress on Information and Communication Technologies, 2011.
4. Barrón Cedeño, A., & Rosso, "On automatic plagiarism detection based on n-grams comparison," In Advances in Information Retrieval, Vol. 5478. Lecture Notes in Computer Science, pp. 696–700, Springer.
5. Ahmad Gull Liaqat and Aijaz Ahmad, "Plagiarism Detection in Java Code," Degree Project, Linnaeus University, June 2011, pp. 1-7.
6. Michael Tschuggnall and Gunther Specht, "Detecting Plagiarism in Text Documents through Grammar-Analysis of Authors," pp. 241-255. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 5, ISSUE 04, APRIL 2016 ISSN 2277-8616 114 IJSTR©2016 www.ijstr.org
7. Plagiarism Detection Process using Data Mining Techniques https://doi.org/10.3991/ijes.v5i4.7869
8. Francisco R., Antonio G., Santiago R., Jose L., Pedraza M., and Manuel N., —Detection of Plagiarism in Programming Assignments,‖ IEEE Transactions on Education, vol. 51, No.2, pp.174-183, 2008.
9. MG Yadav, R Nennuri, “ Data Mining based Modern and Advanced Design and Develeopment Applications”, International Journal of Pure and Applied Mathematics 119 (16), 4651-4658
10. R Nennuri, AK Chaitanya, LP Malyala,” Implementation of data frame work system based on model driven architecture for MAS and Web based applications”, International Journal of Engineering & Technology 7 (2.20), 1-4
11. Rajasekhar Nennuri , Prathyusha Malyala , SaiTejaswi Thotakura,” Crime Prediction and Analysis Using Machine Learning”, International Journal of Advanced Science and Technology 29 (special issue), 9
12. YB M Geetha yadav Golla Swetha , Vasista Kumar, “Intrusion Detection Scheme Using Machine Learning”, icrcsit-20
13. KS M Geetha Yadav , S Srihitha , C Tejaswi, “Clustering of Bigdata in Application Review Analysis”，International Journal of Advanced Science and Technology 29 (special issue)