IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 5, May 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105087 |

A Study on Plagiarism Checker for Duplication Avoidance in Student's Project

Syeda Farha Jabeen[1], Diksha Choughale[2], Vaishnavi Kanmade[3] , Madhu Mohod[4]

U.G. Student, Department of Computer Engineering, SKN Engineering College, Vadgaon(bk), Pune, India[1]

U.G. Student, Department of Computer Engineering, SKN Engineering College, Vadgaon(bk), Pune, India[2]

U.G. Student, Department of Computer Engineering, SKN Engineering College, Vadgaon(bk), Pune, India[3]

U.G. Student, Department of Computer Engineering, SKN Engineering College, Vadgaon(bk), Pune, India[4]

ABSTRACT: The current plagiarism detection system was found to be only for text files checking. Also it is too slow and takes too much time for checking. There is no proper system which can check plagiarism in academic projects.Hence it is very difficult for the college authority to select unique projects every time. The big challenge is to provide plagiarism checking in projects with appropriate algorithm in order to improve the percentage of finding result and time checking. The important question for the plagiarism detection problem in this study is whether it is possible to apply new topic modelling techniques such as Lattent Dirichlet Allocation Algorithm to handle plagiarism problems for projects. Many projects are available on the internet and are easy to access. Due to this availability, users can easily create a new project by copying and pasting from these resources. Sometimes users can reword the plagiarized part by replacing the topic name with their synonyms. Motivation of the paper is to find the most plagiarism content that should be copied from anywhere identified in the efficient manner. Further it helps the staff to maintain the uniqueness of the projects under their guidance**.**

KEYWORDS: Plagarism Checking, academic projects plagiarism, unique projects, Latent Dirichlet Allocation, data mining.

I. INTRODUCTION

Plagiarism defined as "unacknowledged copying of documents or programs". It can occur in many sectors. Plagiarism cases are an everyday topic, for example, in academics, journalism, and scientific research and even in politics. Most empirical studies and analysis were undertaken by the academic community to deal with student plagiarism. With the explosive growth of content found throughout the Web, people can find nearly everything they need for their written work, but detection of such cases can become a tedious task. For these reasons society needs to tackle this problem with computerassisted approaches, and consequently, multiple studies in the field are being conducted. In order to discriminate plagiarized documents from non-plagiarized documents, a correct selection of text features is a key aspect. There are many types of plagiarism mentioned by Hermann Maurer et al., such as copy and paste, redrafting or paraphrasing of the text, plagiarism of idea, and plagiarism through translation from one language to another. Nowadays, many documents are available on the internet and are easy to access. Due to this availability, users can easily create a new document by copying and pasting from these resources. Sometimes users can reword the plagiarized part by replacing the word with their synonyms. This kind of plagiarism is difficult to be detected by the traditional plagiarism detection system. Various methods can be implemented, ranging from document- comparison algorithms and systems to scan the Web, to approaches that utilize language-specific features, for example for the authorship-attribution task

II. RELATED WORK

Data mining is used for finding the useful information from the large amount of data. Data mining techniques are used to implement and solve different types of research problems.This system wants to verify project title using data mining and Text Extraction techniques. Extracting Text is one of the most important tasks when working with text. Readers benefit from keywords because they can judge more quickly whether the text is worth reading. Website creators benefit from keywords because they can group similar content by its topics. Algorithm programmers benefit from keywords because they reduce the dimensionality of text to the most important features. Multiple Methods and Algorithms for

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105087 |

Text extraction and String Matching. But we use machine learning algorithm for Text Attractions. Plagiarism detection became so essential topic in researches area. Plagiarism can be involve in different fields likes research papers, art area and program code. In digitalized future, everything will be in online mode nothing will be there on pen- paper basis. So the plagiarism checking application will be definitely most helpful in the future.

III. METHODOLOGY

Latent Dirichlet Allocation (LDA) Algorithm: It is one of the most popular topic modelling methods. Each document is made up of various words, and each topic also has various words belonging to it. The aim of LDA is to find topics a document belongs to, based on the words in it. LDA describes how the documents in a dataset were created using a generative version. Consider that a dataset is a collection of D documents which, in turn, can be considered as a collection of words. Hence, this model indicates how each document acquires its words. Initially, let's assume that K topic distributions for our dataset, meaning K multinomial containing V elements each, where V is the number of terms in the corpus. Let i represent the multinomial for the i topic. The size of i is V : —i— = V . Given these distributions, the LDA generative process is as follows: Steps: 1. For each document: Randomly choose a distribution over topics (a multinomial of length K) for each word in the document: Probabilistically draw one of the K topics from the distribution over topics obtained in (a), say topic j (ii) Probabilistically draw one of the V words from j.

IV. EXPERIMENTAL RESULTS

Figures shows the results of plagiarism checking from a set of project reports. Fig 1.1is the image showing probability of most used words in each project, Fig 1.2 shows the detailed report of plagiarism checking.

Fig. 1.1 Detailed Report based on synopsis

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105087 |

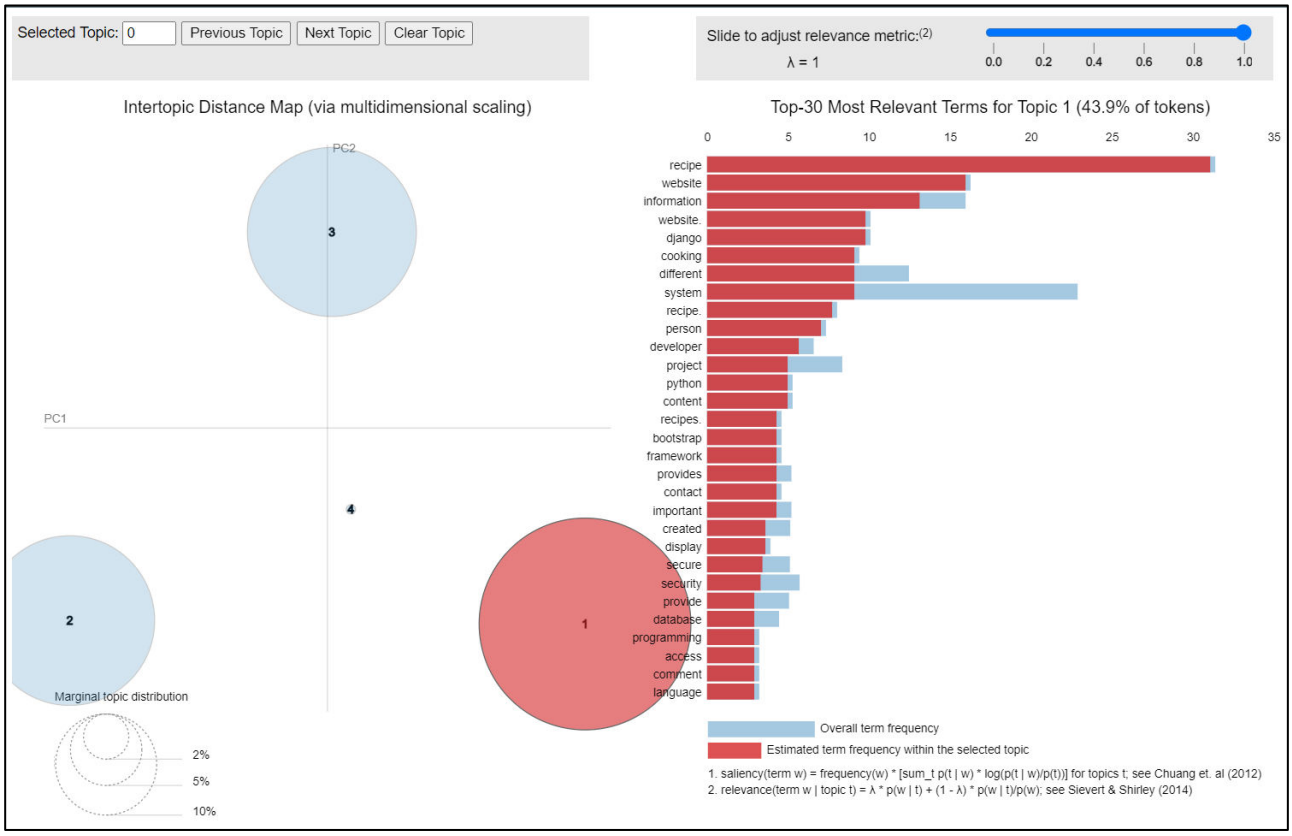

Fig. 1.2 Statistical graph showing the frequency of keywords in each topic

VI. CONCLUSION

The goal of this project is to look at the subject of text plagiarism and the possibility of using computer algorithms to detect it. As a result, strategies and methodologies for detecting digital automated plagiarism have been developed. The collecting of possible sources to compare the suspected papers with is one of the first issues the systems encounter. This is a separate issue, because it is typical that ideal and real sources are not always available, limiting the capabilities of algorithms that compute document-to-document similarity. It was recently put to the test, and research including several writing style markers are currently being conducted. In this study, a self-based information algorithm is investigated, whose primary idea is to utilise a function to measure writing style simply based on the use of words.

REFERENCES

[1] Kamalakar Pallela, Sneha Talari, "Plagiarism : A serious ethical issue for indian students", 2016 IEEE International Symposium on technology and society (ISTAS) 20-22 October 2016

[2] Joshi O D, Pudale A H, Ghorpadde R D, "Text extraction technique applied to plagiarism detector : the semantic analysis for analyzing the writing style" International Journal of Computer science engineering(IJCSE) Mach-2016 ISSN 2319-7323

[3] Dr. S.Vijayarani Ms. R.Janani: String Matching Algorithms for Retrieving Information from Desktop – Comparative Analysis.

[4] Dr. (Ms). Ananthi Sheshasayee1 Ms. G. Thailambal2: A COMPARITIVE ANALYSIS OF SINGLE PATTERN MATCHING ALGORITHMS IN TEXT MINING

[5] Hemlatha A.M, M.Subha, "A study on plagiarism checking with appropriate algorithm in data mining", International Journal of Research In Computer Application and Robotics, Nov 2014, ISSN 2320-7345

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details