International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Document Plagiarism Checker Based on Similarity

Shanmathi V[1]*, Sowmiya K*[1]*, Aishwarya SH*[1]*, Sakthivel M*[2]

***UG Student*[1]*, Assistant Professor*[2]**

[1,2]Department of Computer Science Engineering

Veltech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi.

ABSTRACT:

Plagiarism is the copying of previously published material in an altered manner, or occasionally the original work in its original form. This is rather typical among scholars, researchers, and students. This has had a significant impact on the research community and raised awareness among academics about the need to stop this form of misconduct. Plagiarism is the act of passing off someone else's idea as your own. It is theft of intellectual property. Plagiarism, however, can involve more than just taking credit for someone else's work—it can also involve your own original composition. The cosine formula, which is primarily used to determine the degree of similarity between two documents, is used to calculate plagiarism. By identifying the similarities between the two texts, one can prevent plagiarism by lowering the percentage between the two documents.

Keywords: Plagiarism, intellectual property**,** Cosine similarity lexical, Word tokenize, Bag of words, Vectors, word frequencies, Document Vectors

1. Introduction:

Plagiarism is the theft of intellectual property; this means spreading someone else's idea as your own. But plagiarism is more than stealing someone else's work; you can copy your own work. Copy large sections of text from a single source without citing the source. Take notes from different sources, combine them and make the work your own. Copy from the source but change some words and phrases to hide plagiarism. Plagiarism checkers use advanced software to scan for matches between your text and existing texts. Universities use plagiarism to evaluate student work. You can also use the plagiarism checker to check your work before submitting it.

2. Methodology:

1. Text Based Plagiarism:

This type of plagiarism involves using vector space models to identify similarities between texts. It can also determine how many times it appears in a document and how redundant it is. Based on this information, the fingerprints of the document are compared with other documents to determine whether there is any similarity. This method is suitable to avoid being a part of plagiarism. The program gives very good results when an article contains partial plagiarism. Otherwise it takes all the data and uses the space vector to match the text. This includes "scraping and printing" in the original text of online books and magazines, "editing or changing a few words" and "newspapers, research, magazines, personal information or ideas".

1. First stage of Plagiarism Detection Process involves “the student or the researcher to upload their assignments or works to the web engine, the web engine acts as an interface between the students and the system.”

1.2 Stage Two Analysis:

The second step is to complete the data transfer or work through the similarity search engine to check whether the data is similar to other data. There are two types of common carpels; the first type is the intracarpelic car and the second type is the extracarpelic car. - The intracarpellar motor works by returning an ordered list of all similar pairs.

1.3 Stage Three Confirmations:

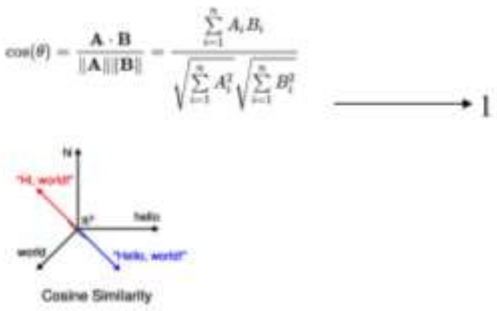
The task of this stage is to detect whether there are some specific texts copied from the original texts or to determine whether there are similarities between the texts and other texts.

1.4 Stage Four Investigation:

This stage determines whether a particular relevant text has been copied from the original text or whether there is a high degree of resemblance between the source text and any other text.

2. Cosine Similarity:

Cosine similarity measures the similarity between two vectors in the internal object space. It is measured as the cosine of the angle between two vectors and determines whether the two vectors point approximately in the same direction. It is often used to evaluate data similarity in text analysis.



$$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \longrightarrow 1$$

Fig 1: Example for cosine similarity

2.2 Document Similarity:

Similarity between two documents can be achieved by converting the words or expressions in the document or sentence into vector representation. The vector representation of the data can be used in a similar cosine model to obtain a corresponding quantity. A cosine similarity of 1 means the two data are the same, and a cosine similarity of 0 means there is no similarity between the two data.

2.3 Working Of Cosine Similarity:

Document 1: Deep Learning can be hard
Document 2: Deep Learning can be simple

WORD	**DOCUMENT 1**	**DOCUMENT 2**
Deep	**1**	**1**
Learning	**1**	**1**
Can	**1**	**1**
be	**1**	**1**
hard	**1**	**0**
simple	**0**	**1**

Document 1: [1,1, 1,1,1,0] let's refer to this as A

Document 2: [1, 1, 1, 1, 0, 1] let's refer to this as B

Above we have two vectors (A and B) that are in a 6 dimension vector space

Calculate the dot product between A and B:

1.1 + 1.1 + 1.1 + 1.1 + 1.0 + 0.1 = 4

Calculate the magnitude of the vector A:

$\sqrt{}1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 = 2.2360679775$

Calculate the magnitude of the vector B:

$1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 1^2 = 2.2360679775$

3. The Overall Process of Work:

In this study, we developed and expanded the application of plagiarism ratings. Add calculations to your final project document to make it easier for students and teachers. Identify and monitor final project documents for levels of plagiarism. The design stages of this study are represented in a diagram. in Figure 4.

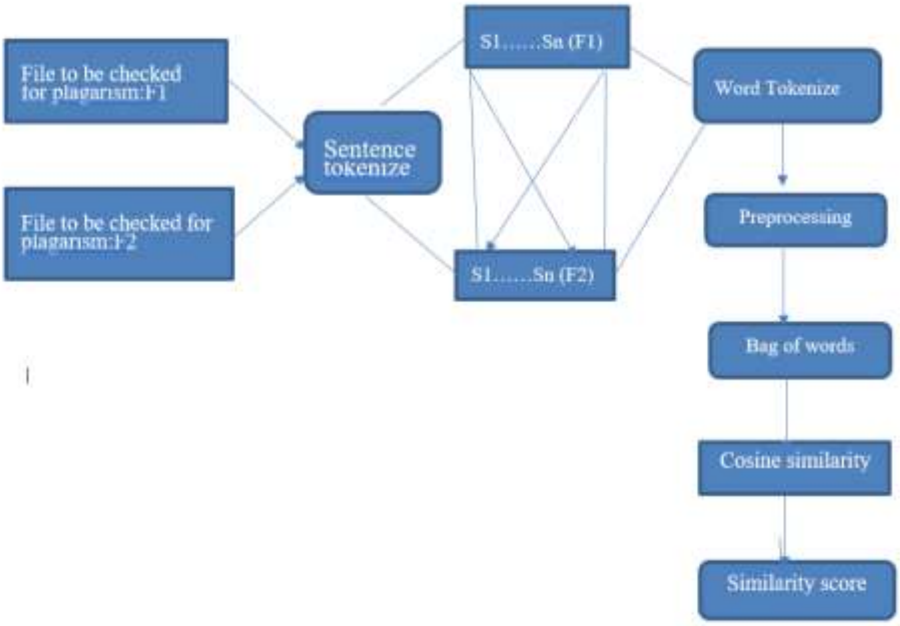


Figure.2 Architecture Diagram

1. Tokenization:

Tokenization is the process of using every word in a document. The output of the tokenization step is data identical to the content of the document. Participates in article preparation or review and revision for intellectual content.

2 Sentence Tokenization:

Sentence tokenization, text is split into sentences. This is useful for tasks that require analyzing or processing individual sentences.

3. Preprocessing:

Preprocessing in text similarity refers to the steps taken to clean and prepare text data before it is used in similarity calculations. These steps help improve the accuracy and efficiency of the text similarity algorithms.

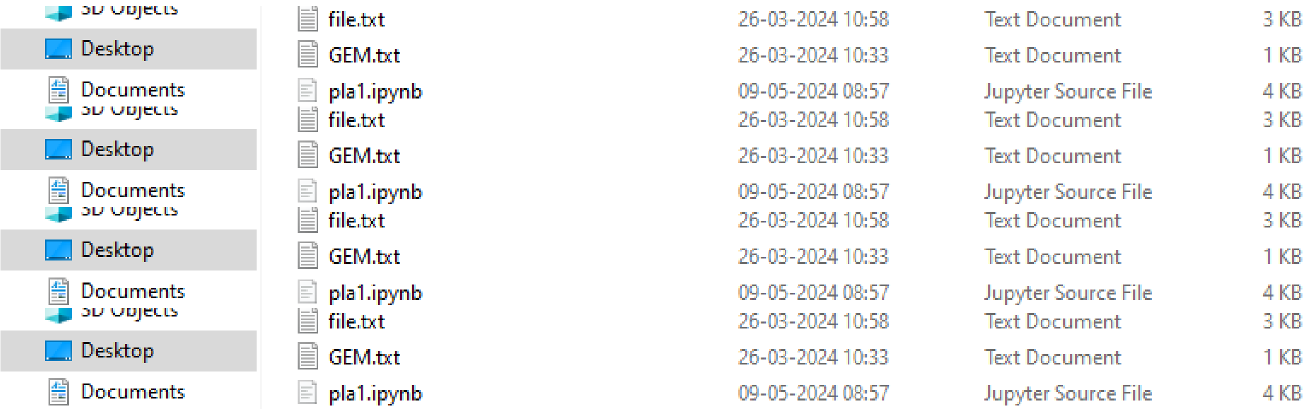
4. Bag of Words:

The Bag of Words (Bow) model is a method for converting text into a numerical representation, which can then be used to measure the similarity between different texts.

5. Similarity Score:

The text similarity score using cosine similarity is a numerical value that quantifies the degree of similarity between two text documents. This score ranges from -1 to 1. Here 1 means both documents are identical. 0 means there is no similarity between documents. -1 indicates that the document is the exact opposite. However, in practice, this value rarely occurs in general text similarity problems because word frequencies are non-negative

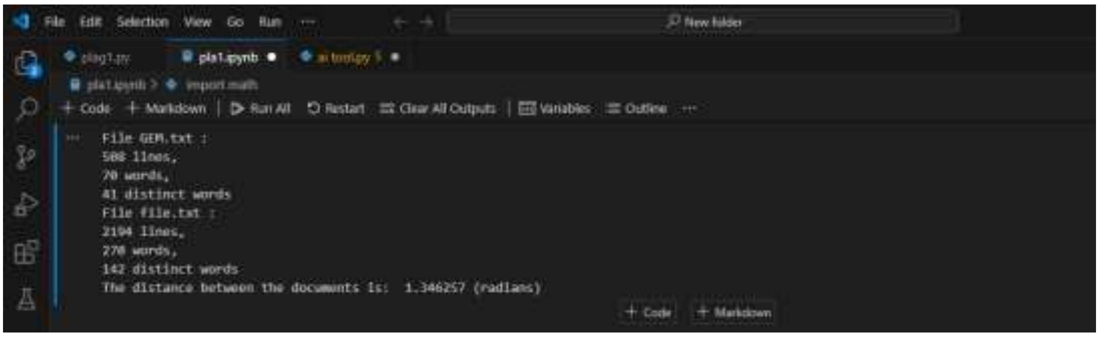
4. Results:



file.txt	26-03-2024 10:58	Text Document	3 KB
GEM.txt	26-03-2024 10:33	Text Document	1 KB
pla1.ipynb	09-05-2024 08:57	Jupyter Source File	4 KB
file.txt	26-03-2024 10:58	Text Document	3 KB
GEM.txt	26-03-2024 10:33	Text Document	1 KB
pla1.ipynb	09-05-2024 08:57	Jupyter Source File	4 KB
file.txt	26-03-2024 10:58	Text Document	3 KB
GEM.txt	26-03-2024 10:33	Text Document	1 KB
pla1.ipynb	09-05-2024 08:57	Jupyter Source File	4 KB
file.txt	26-03-2024 10:58	Text Document	3 KB
GEM.txt	26-03-2024 10:33	Text Document	1 KB
pla1.ipynb	09-05-2024 08:57	Jupyter Source File	4 KB

Import Files

Visual Studio Code



```
File GEM.txt :
588 lines,
70 words,
41 distinct words
File file.txt :
2194 lines,
270 words,
142 distinct words
The distance between the documents is:  1.346257 (radians)
```

Fig 3 : Experimental Result

5. Conclusion:

Plagiarism is a serious ethical and academic concern that undermines the principles of originality, integrity, and intellectual honesty. It involves presenting someone else's ideas, work, or intellectual property as one's own, thereby violating the fundamental principles of academic and creative pursuits. The consequences of plagiarism can be severe, ranging from academic penalties to reputational damage. In future we are planning to implement plagiarism checker for journals, to find image similarity and voice similarity.

References:

1.George Tsatsaronis , Iraklis varlamis , Andreas giannakoulopoulos , Nikolaos kanellopoulos,"Identifying free text plagiarism based on semantic similarity " – 2010

2. M. E. T. Ali, H. M. D. Abdulla, and V. Snasel, "Overview and comparison of plagiarism detection tools," – 2011.

3. D Gunawan, C A Sembiring , M A Budiman," The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents "– 2018.

4.Piska dwi Nurfadila, A. Wibawa, I. Zaeni, A. Nafalski ,"Journal Classification Using Cosine Similarity Method on Title and Abstract with Frequency-Based Stop word Removal"- 2019.

5. Sagarika Pattnaik, Nayak, “Summarization of Odia Text Document Using Cosine Similarity and Clustering “-2019.

6. Suniel kumar P, Athira P shaji,” A Survey on Semantic similarity” – 2020.

7. Sherya saloni verma, Abdullah sarguroh, jyotsana Rawat,” Identification of text similarity based context “- 2021.

8. Rival Fauzi, Muhammad Iqbal, Tita Haryanti,” Design and Implementation of a Final Project Plagiarism Detection System Using Cosine Similarity Method”- 2021.

9. Sayantan Pal, Maiga Chang, Maria Fernandez Iriarte, “Summary Generation Using Natural Language Processing Techniques and Cosine Similarity” - 2021.

10. H.Zhang, M.Huang and W. Li,” .Research on text similarity measurement hybrid algorithm with term semantic information and TF-IDF method “- 2022.