Assignment Plagiarism Checker

By

Arya Kumar Johary

(B.Tech 3rd Year, Enrollment No.: 12021002026028)

Abhay Prasad

(B.Tech 3rd Year, Enrollment No.: 12021002026024)

Under the Supervision of

Asst. Prof. Sagarika Ghosh &

Asst. Prof. Hriday Banarjee

Dept. of Computer Science & Engineering

University of Engineering & Management, Jaipur

Outlines

- Introduction
- Literature Review
- Objectives
- Proposed Model
- Result Analysis
- Conclusion & Future Scope
- Reference
- Acknowledgement

Introduction

Problem Statement- To create a tool for writers, which can efficiently detect the plagiarism of a text document with the application of TF-IDF vectorizer and cosine similarity.

The proposed solution is a plagiarism checker program that leverages TF-IDF vectorizer and cosine similarity to compare text documents and identify potential plagiarism. The TF-IDF vectorizer transforms a collection of text documents into a matrix of numerical values, representing the significance of each word in each document. Cosine similarity measures the angle between two vectors, in this case, the TF-IDF vectors of two documents. By comparing the cosine similarity of different document pairs, we can gauge their similarity in terms of word usage and content. This program can help ensure originality and integrity in writing, proving beneficial for students, teachers, researchers, and writers aiming to check their work for plagiarism and avoid academic dishonesty.

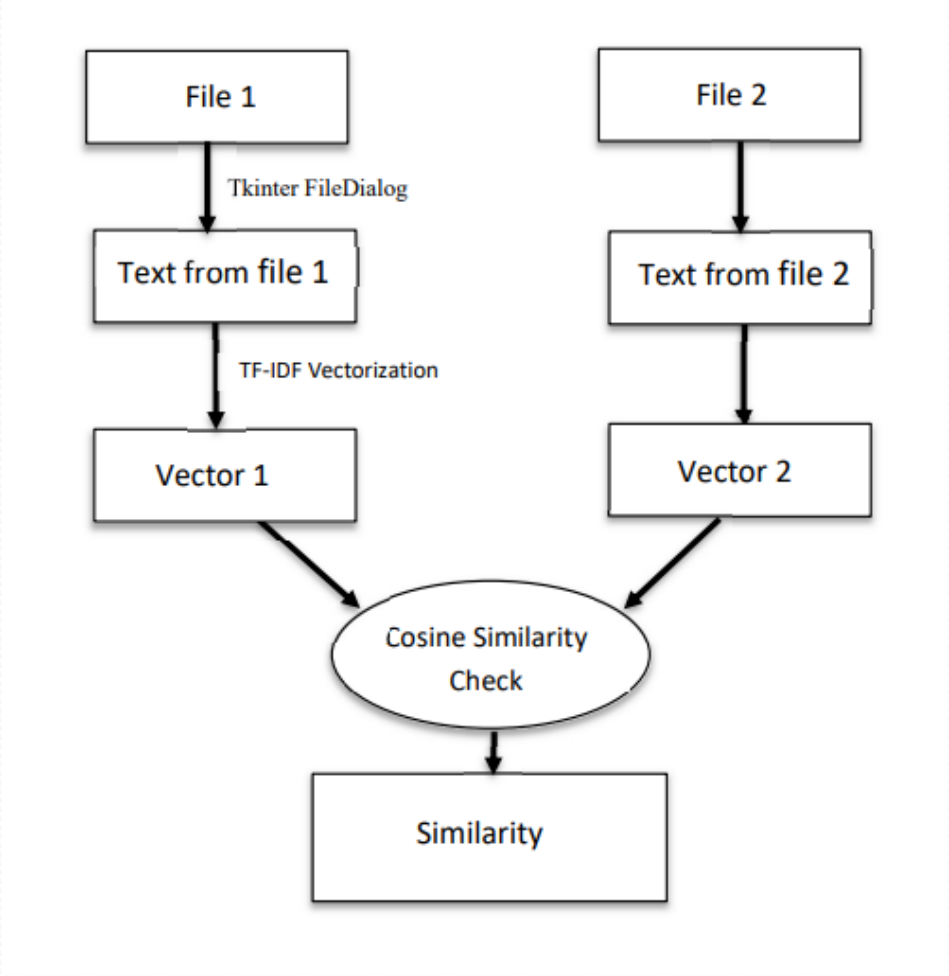
Literature Review

Author(s)	Date	Description
Martins, Vítor T., Fonte, Daniela, Henriques, Pedro Rangel Cruz and Daniela da	2014	Proposed three methodologies to compare source codes: attribute-based (size metrics), token-based (hashing and fingerprinting), and structure-based (IR, AST, or PDG)
Hakkun Elmunsyah, Hary Suswanto, Khoirudin Asfani and Wahyu Hidayat	2018	Addressed the impact of IT on education and evaluated a Plagiarism Checker's effectiveness, showing over 85% success in detecting similarities in students' scientific papers.
Angelos Rodafinos	2018	Summarized the potential and challenges of Python-based plagiarism checker systems, and suggests future research directions, such as using cloud computing and big data analytics.

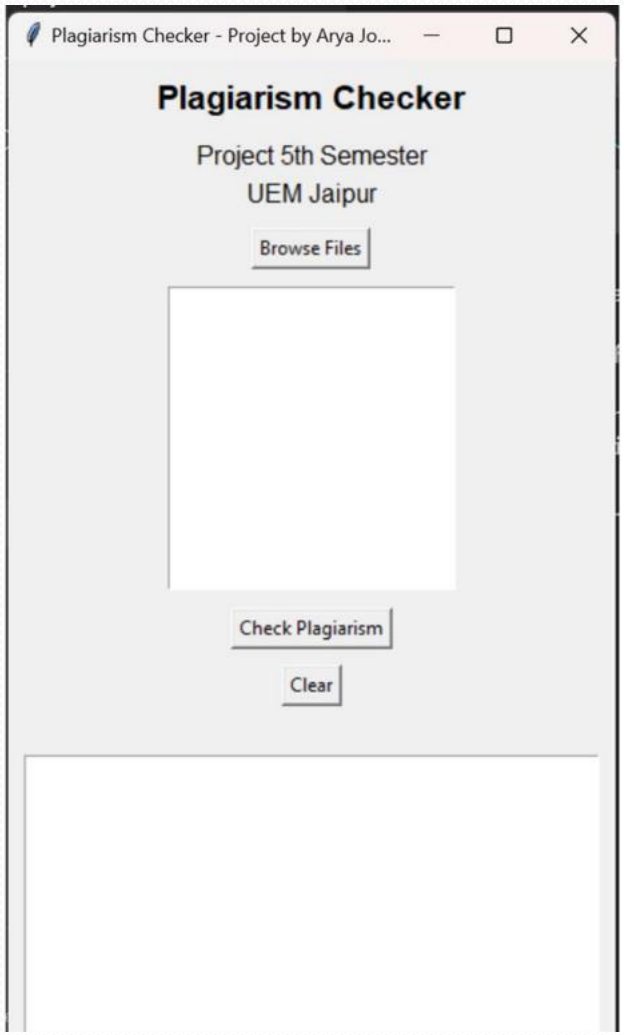
Objectives

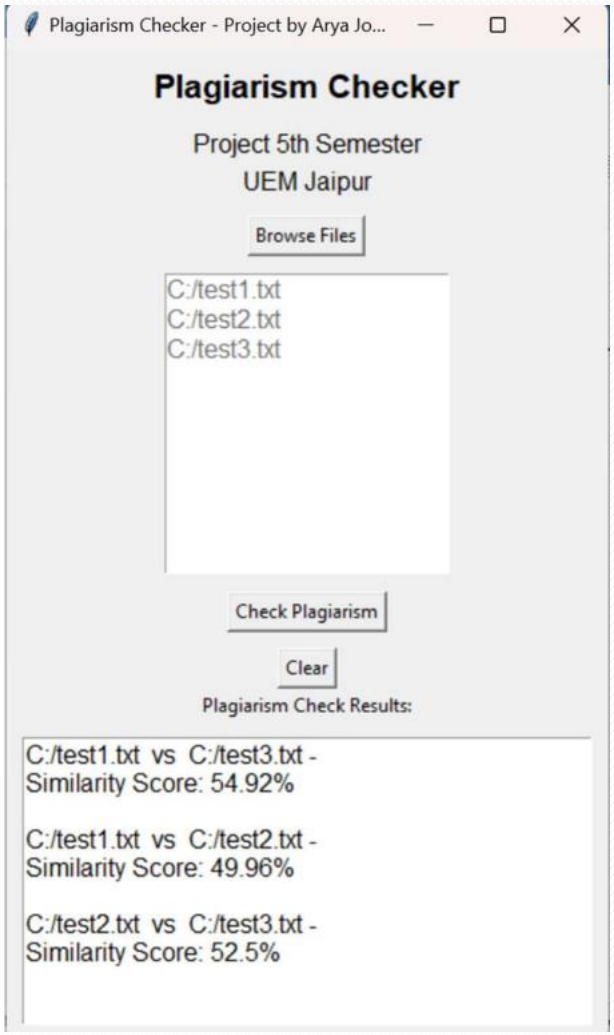
- Detect Plagiarism
- Text Processing
- Similarity Score
- Scalability
- Efficiency
- User Interface
- Report Generation
- User Interaction
- Testing
- Documentation

Proposed Model



Result


Plagiarism Checker - Project by Arya Jo...
Plagiarism Checker
Project 5th Semester
UEM Jaipur
Browse Files
Check Plagiarism
Clear


Plagiarism Checker - Project by Arya Jo...
Plagiarism Checker
Project 5th Semester
UEM Jaipur
Browse Files
C:/test1.txt
C:/test2.txt
C:/test3.txt
Check Plagiarism
Clear
Plagiarism Check Results:
C:/test1.txt vs C:/test3.txt -
Similarity Score: 54.92%
C:/test1.txt vs C:/test2.txt -
Similarity Score: 49.96%
C:/test2.txt vs C:/test3.txt -
Similarity Score: 52.5%

Limitation

- Paraphrased, translated, or obfuscated text is hard to detect
- Large and diverse corpus of text documents needed
- Scalability, performance, and security issues may arise
- Responsibility, accountability, and transparency issues may emerge
- Images and other non-text content may be ignored

Conclusions & Future Scope

The Plagiarism Checker project is a Python program that detects similarities among text documents. It uses cosine similarity to measure the likeness between documents and shows the results in a user-friendly interface. Users can browse, select, and analyze multiple files for potential plagiarism. The project supports academic integrity by encouraging originality and ethical writing. The Plagiarism Checker project is a useful tool for education and writing.

Future Scope:

- The project can be extended to detect plagiarism in other types of content, such as images, audio, video, or code, by using appropriate similarity metrics and techniques.
- Using NLP, text mining, and web scraping to improve project.
- Multi-Modal Approaches
- Integrating project with online platforms and content providers.

References

- Cosine Similarity Documentation. (https://scikitlearn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html)
- TF – IDF Vectorizer Documentation. (https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)
- The Effectiveness of Plagiarism Checker Implementation in Scientific Writing for Vocational High School (https://www.atlantispress.com/proceedings/aptekindo-18/25903493)
- Plagiarism Detection: A Tool Survey and Comparison (https://repositorium.uminho.pt/bitstream/1822/64358/1/14.pdf)
- Plagiarism Management: Challenges, Procedure and Workflow Automation (https://www.ijello.org/Volume14/IJELLv14p159-175Rodafinos5036.pdf)

Acknowledgement

The endless thanks goes to Lord Almighty for all the blessings he has showered onto us, which has enabled us to write this last note in our research work. During the period of our research, as in the rest of our life, we have been blessed by Almighty with some extraordinary people who have spun a web of support around us. Words can never be enough in expressing how grateful we are to those incredible people in our life who made this thesis possible. We would like an attempt to thank them for making our time during my research in the Institute a period we will treasure. We are deeply indebted to our research supervisors, Prof. Sagarika Ghosh and Prof. Hriday Banerjee, who gave us such an interesting thesis topic. Each meeting with them added in valuable aspects to the implementation and broadened our perspective. They have guided us with his invaluable suggestions, lightened up the way in our darkest times and encouraged us a lot in the academic life.

Arya Kumar Johary & Abhay Prasad

UEM
श्रद्धावान लभते ज्ञानम्
UNIVERSITY OF ENGINEERING & MANAGEMENT
Good Education, Good Jobs