PLAGIARISM DETECTION AND REPORTING SYSTEM

A proposal for the COEP Technological University Hackathon by
Team V6

Abhijeet Jadhav – 612203071 Aaryan Tokekar – 612211001 Kapil Tangsali – 612203177
Yash Pawar – 612203142 Soham Vaze – 612203187 Tanmayi Sulakhe - 612203173

PLAGIARISM DETECTION AND REPORTING SYSTEM (PDRS)

Approach with respect to the Problem Statement

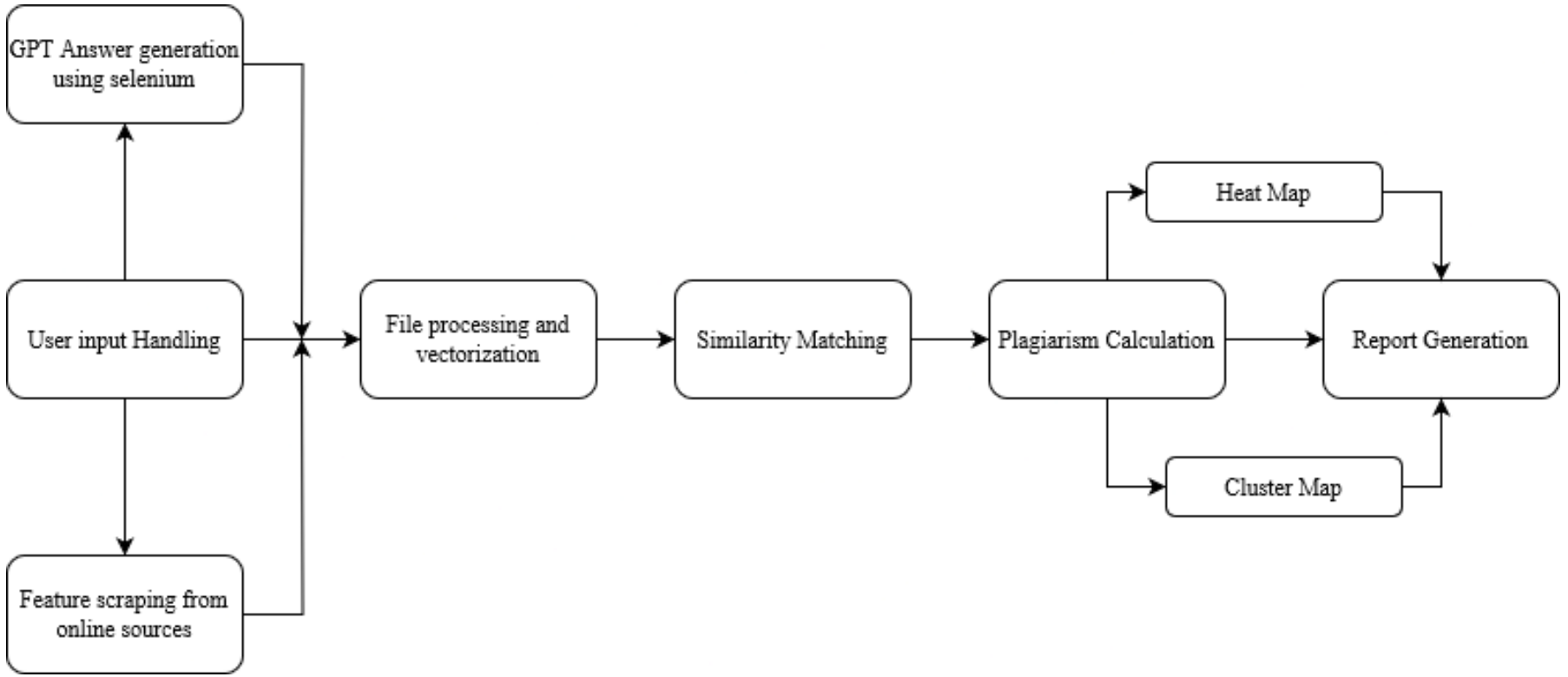
- PDRS is designed to detect plagiarism in a large volume of files submitted to it on top of a pre-existing database
- Capable of cross-referencing a batch of files within themselves for plagiarism but also references them against several online sources
- It provides a comprehensive report to a user which is customizable and save-able for future references

PLAGIARISM DETECTION AND REPORTING SYSTEM (PDRS)

Tech Stack Overview

- Flask: Lightweight web framework for building scalable web applications.
- SQLAlchemy: Python SQL toolkit and Object-Relational Mapping (ORM) library for database management.
- HTML-CSS: Markup and styling languages for creating structured and visually appealing web pages.
- JavaScript: Versatile scripting language for adding interactivity and dynamic content to web applications.
- Python: High-level programming language known for its simplicity and versatility, used for backend logic and effective file processing tasks
- Selenium: A robust automated testing framework, augments our tech stack, enabling us to bypass OpenAI API

Implementation and Algorithm Design



Algorithm for better Plagiarism Calculation

- Our algorithm is a self developed algorithm largely based on Cosine Similarity
- It improves upon the effectiveness of calculating Cosine Similarity between two vectors by including multiple words (bigrams) for a single value in the Term Frequency – Inverse Document Frequency calculation
- TF-IDF Vectorization helps us to understand the relevance of a document's feature words relative to the whole corpus
- By including bigrams we can identify semantic relationships between words and identify important phrases in documents

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

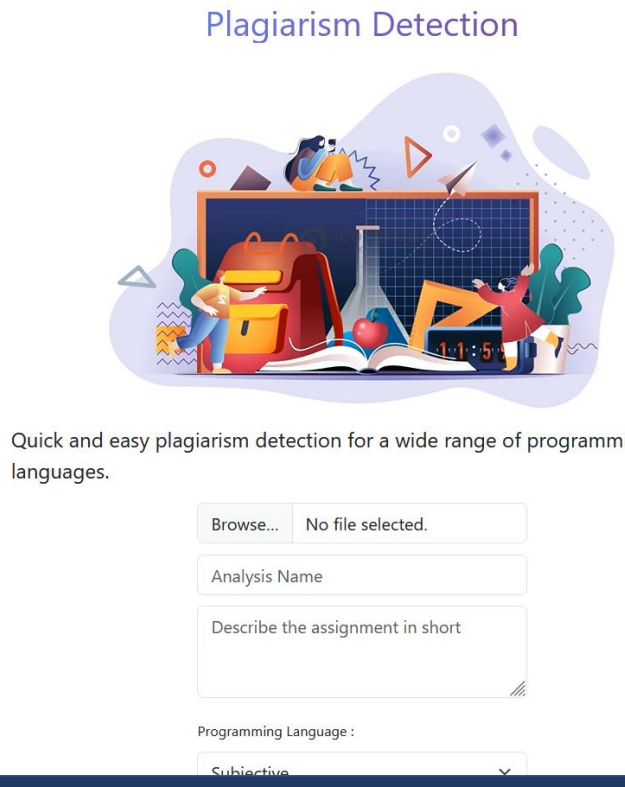
$tf_{x,y}$ = frequency of x in y

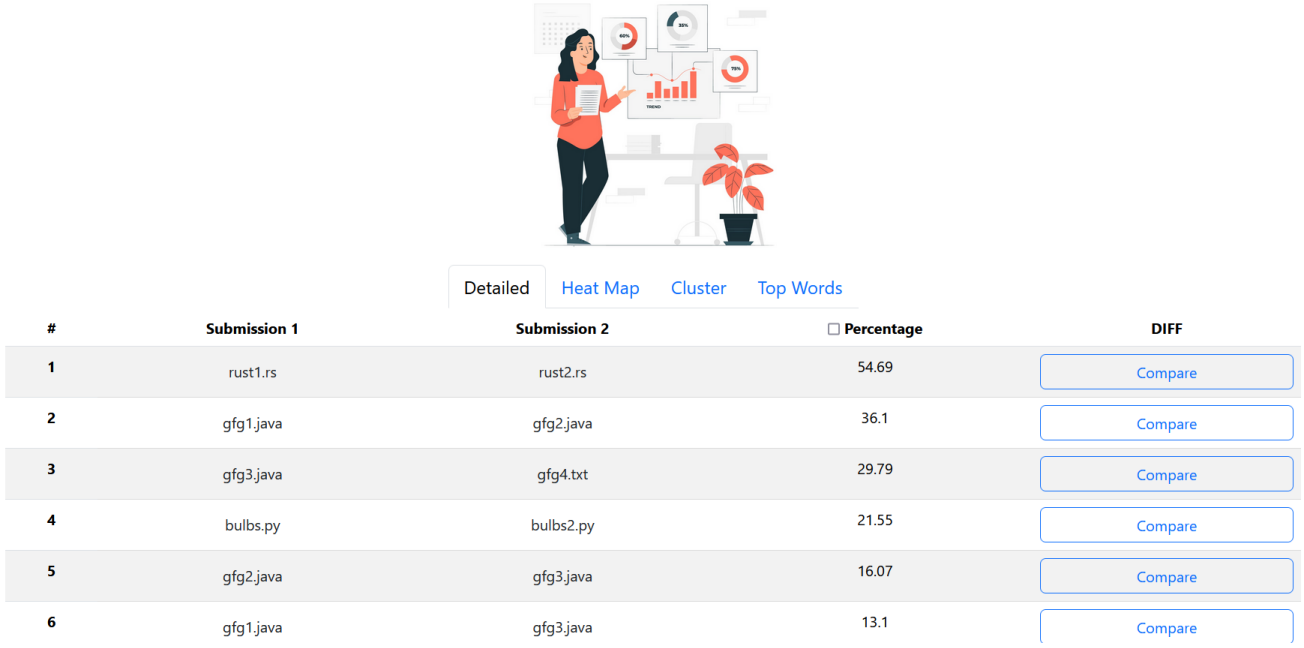
df_x = number of documents containing x

N = total number of documents

PLAGIARISM DETECTION AND REPORTING SYSTEM (PDRS)

UI Design and Functionality

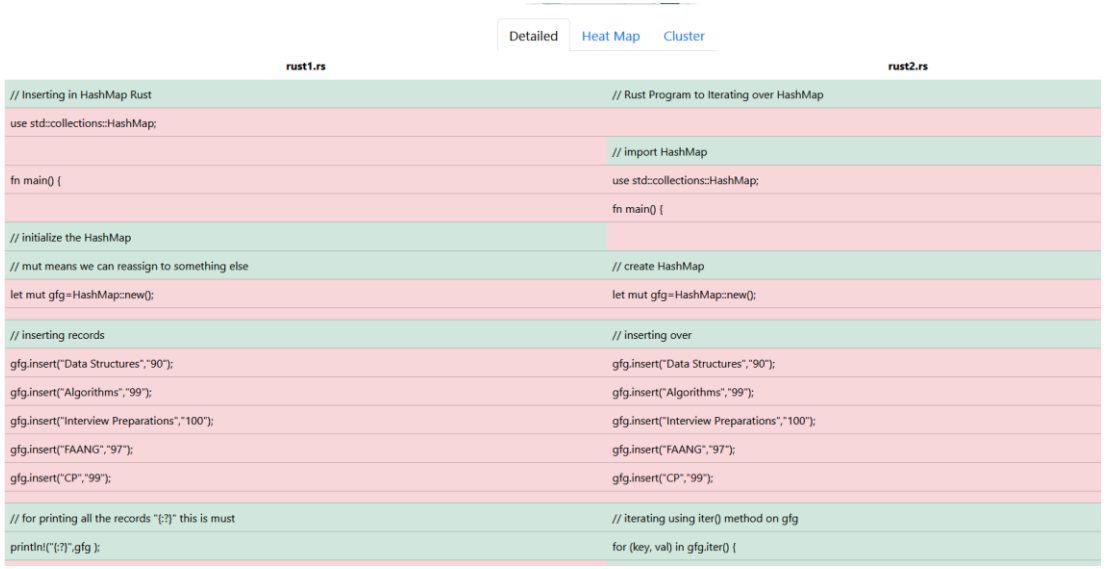


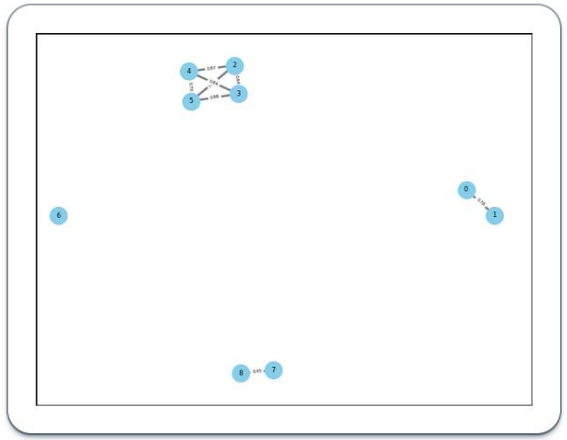


#	Submission 1	Submission 2	Percentage	DIFF
1	rust1.rs	rust2.rs	54.69	Compare
2	gfg1.java	gfg2.java	36.1	Compare
3	gfg3.java	gfg4.txt	29.79	Compare
4	bulbs.py	bulbs2.py	21.55	Compare
5	gfg2.java	gfg3.java	16.07	Compare
6	gfg1.java	gfg3.java	13.1	Compare

PLAGIARISM DETECTION AND REPORTING SYSTEM (PDRS)

UI Design and Functionality





#	Submission 1	Submission 2	Percentage	DIFF
1	rust1.rs	rust2.rs	54.69	Compare
2	gfg1.java	gfg2.java	36.1	Compare

PLAGIARISM DETECTION AND REPORTING SYSTEM (PDRS)

Report Generation

- We have taken major strides to enable easy identification of plagiarism once the corpus has been processed
- Key visual aids are a heatmap correlation matrix, a cluster plot and 1:1 comparison between a pair of documents
- These aids combined with a readily available comprehensive report on the dashboard enable the user to quickly identify and help them make a well informed decision