PDFX(PDF Plagiarism Checker)

Submitted in partial fulfillment of the requirements

of the degree

BACHELOR OF ENGINEERING

IN INFORMATION TECHNOLOGY

By

Mohammad Ahmed Ansari	**21101B0031**
Yash Gupta	**21101B0043**
Yash Dhanawade	**21101B0054**
Sahil Ukarde	**21101B0027**

Supervisor

Prof. Rasika Ransing

VIT Vidyalankar Institute of Technology

Accredited A+ by NAAC

Department of Information Technology

Vidyalankar Institute of Technology

Vidyalankar Educational Campus,

Wadala(E), Mumbai - 400 037

University of Mumbai

(AY 2022-23)

CERTIFICATE

This is to certify that the Mini Project entitled **“PDFX(PDF Plagiarism Checker)”** is a bonafide work of **Mohammad Ahmed Ansari (21101B0031), Yash Gupta (21101B0043), Yash Dhanawade (21101B0054), Sahil Ukarde (21101B0027)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of “**Bachelor of Engineering**” in “**Information Technology” .**

Prof. Rasika Ransing
Supervisor

Dr. Vipul Dalal
Head of Department

Dr. S. A. Patekar
Principal

Mini Project Approval

This Mini Project entitled **“PDFX (PDF Plagiarism Checker)”** by **Mohammad Ahmed Ansari (21101B0031), Yash Gupta (21101B0043), Yash Dhanawade (21101B0054), Sahil Ukarde (21101B0027)** is approved for the degree of **Bachelor of Engineering** in **Information Technology.**

Examiners

1..
(Internal Examiner Name & Sign)

2..
(External Examiner name & Sign)

Date:

Place:

Contents

Abstract ii

Acknowledgments iii

List of Abbreviations iv

List of Figures v

List of Tables vi

List of Symbols vii

1 Introduction 1

1.1 Introduction

1.2 Motivation

1.3 Problem Statement & Objectives

1.4 Organization of the Report

2 Literature Survey 4

2.1 Survey of Existing/Similar System

2.2 Limitation Existing/Similar system or research gap

2.3 Mini Project Contribution

3 Proposed System (A Thorough Solution for Plagiarism Detection) 7

3.1 Introduction

3.2 Architecture/ Framework

3.3 Algorithm and Process Design

3.4 Details of Hardware & Software

3.4 Experiment and Results

3.5 Conclusion and Future work

References 16

Abstract

We developed a PDF plagiarism checker in Python that can compare multiple PDF files and detecting the percentage of plagiarism present in them. Our software includes both student login and teacher login options, allowing students to upload their PDF files and teachers to check them for plagiarism.

As developers, we utilized advanced algorithms to compare the uploaded PDF files and highlight any areas where plagiarism has occurred. Our software provides a detailed report showing the percentage of similarity between the uploaded documents and any existing sources.

We also included a feature to assign marks based on the percentage of plagiarism detected, with higher marks given to those who submit original work with no plagiarism. In addition, our software ensures that students cannot access other students' work or the teacher's login, ensuring confidentiality and security.

Overall, our PDF plagiarism checker is an essential tool for educational institutions to maintain academic integrity and prevent students from submitting plagiarized work.

Acknowledgments

We would like to express our special thanks of gratitude to our professor's Prof. Rohit Barve, Prof. Rasika Ransing, Prof. Vinita Bhandiwad, Prof. Dilip Motwani as well as Our Honorable H.O.D Dr. Vipul Dalal who gave us the golden opportunity to do this wonderful Project on the topic **"PDFX (PDF Plagiarism Checker)"** which also helped me in doing a lot of Research and gaining some precious knowledge. We would like to express our special thanks of gratitude to our professors.

We would like to express our sincere gratitude to our college faculty for providing us with the opportunity and support to undertake this project. We would also like to thank our project guide for providing us with valuable feedback and guidance throughout the development process.

Finally, we would like to thank the college for providing us with the necessary infrastructure and resources to complete this project. This project has been a valuable learning experience for us, and we are grateful for the opportunity to work on it.

List of Abbreviations

- PDF: Portable Document Format
- Python: A high-level programming language
- %: Percentage
- GUI: Graphical User Interface
- API: Application Programming Interface
- DBMS: Database Management System
- UI: User Interface
- OS: Operating System

Chapter 1

1.1 Introduction

Plagiarism is a serious issue in academics, and detecting it is of utmost importance to maintain the integrity of the education system. The PDF plagiarism checker is an innovative tool that allows teachers to assess the level of similarity between the assignments submitted by their students. This tool has several features that make it convenient and user-friendly. The software contains both student login and teacher login, allowing for easy assignment submission and plagiarism checking.

Teachers can create an assignment and can be seen by the students up on the assignment section, who can then submit their assignments in PDF format. The tool then compares the submitted PDFs and calculates the percentage of similarity between them. Based on the plagiarism percentage, students receive marks for their assignments.

The plagiarism checker is designed to ensure that teachers can only check for plagiarism between student's assignments, and not against other external sources. The tool stores all marks with the corresponding student name in a database, making it easy for teachers to keep track of student performance over time. This tool is a valuable addition to any academic institution, as it allows for efficient and accurate detection of plagiarism, promoting academic integrity and honesty.

1.2 Motivation

The motivation behind the PDF plagiarism checker is to address the growing concern of academic dishonesty and plagiarism. In recent years, instances of plagiarism in academic institutions have increased significantly, creating an need for effective plagiarism detection tools.

By using the PDF plagiarism checker, teachers can detect plagiarism more efficiently and accurately, which helps to maintain the integrity of the education system. This tool motivates students to submit original work, reducing the temptation to plagiarize, and creating a level playing field for all students.

Moreover, the PDF plagiarism checker provides an objective measure of student performance, which helps teachers to evaluate the effectiveness of their teaching methods. With the help of this tool, teachers can identify areas where students struggle the most and provide additional support where necessary.

Overall, the motivation behind the PDF plagiarism checker is to create a fair and transparent academic environment, where students are encouraged to submit original work and are rewarded based on their individual effort and performance.

1.3 Problem statement and Objectives

Problem Statement:

With the increase in the use of digital content, plagiarism has become a major concern in the academic community. It is becoming increasingly challenging for teachers to detect plagiarism effectively, leading to a rise in academic dishonesty. Therefore, there is a need for a reliable and efficient tool to detect plagiarism in digital assignments.

Objectives:

- To develop a PDF plagiarism checker that can accurately compare a given number of PDF documents.
- To provide a user-friendly interface with separate login systems for students and teachers.
- To enable teachers to create assignments and notify students of upcoming deadlines.
- To calculate the percentage of plagiarism in each submitted assignment and assign marks accordingly.
- To notify students of their assigned marks and provide them with a detailed report on their plagiarism percentage.
- To store all marks with the corresponding student name in a database for easy access and record keeping.
- To ensure that teachers can only check plagiarism between students and not against other sources.
- To create a fair and transparent academic environment where students are encouraged to submit original work and are rewarded based on their individual effort and performance.

The primary objective of this project is to develop a reliable and efficient tool for detecting plagiarism in digital assignments that promotes academic honesty and integrity. The software should be user-friendly and provide accurate and objective measures of student performance, allowing teachers to identify areas where additional support may be needed. The project aims to create a fair and transparent academic environment where students are rewarded based on their individual effort and performance, leading to a better learning experience for all students.

2. Literature Survey

2.1 Survey of Existing/Similar System

Turnitin - Turnitin is one of the most popular plagiarism detection tools used in academic institutions. It provides a comprehensive solution that includes grading, feedback, and plagiarism detection. Turnitin compares student submissions to an extensive database of academic and online sources and provides a report indicating the percentage of plagiarism detected.

Turnitin is a cloud-based plagiarism detection and academic writing solution widely used in educational institutions worldwide. Turnitin is primarily used by teachers and professors to check the authenticity and originality of students' papers, assignments, and other written work. Turnitin works by comparing submitted documents against an extensive database of academic papers, journals, and other sources to detect instances of plagiarism.

2.2 Limitation Existing/Similar system or research gap

Limited support for PDF format: Many plagiarism detection tools do not support the PDF file format, which is widely used in academic institutions for submitting assignments. Our proposed PDF plagiarism checker addresses this limitation by providing support for the PDF file format.

Limited customization options: Many existing tools do not provide enough customization options for teachers to create and manage assignments. Our proposed tool provides a user-friendly interface for teachers to create, manage and grade assignments as well as notifications to students.

Limited integration with institutional systems: Many plagiarism detection tools do not integrate well with institutional systems, making it difficult for administrators to manage the tool. Our proposed tool is designed to integrate seamlessly with existing institutional systems, making it easier for administrators to manage the tool.

With the team collaboration feature, students and teachers can create their teams based on the course, subject, or project. They can also add and remove Assignments, Submissions etc.

Furthermore, having a team collaboration feature will foster a sense of community and promote peer-to-peer learning, which can enhance the overall learning outcomes for students.

2.3 Mini Project Contribution

Our project began with brainstorming sessions on how to approach the task at hand. On this first level, we were discussing about so many information that we felt it was time that we pulled those ideas together and start working on the project. Each of us did our share in the project and later a meet up was initiated to discuss and compiled our information.

Mohammad Ahmed Ansari with the help of group members helped to contribute the Front End of The GUI application and User Interface.

Yash Gupta with the help of group members helped to contribute to the integration of Back-End to the Graphical User Interface.

Yash Dhanawade and all members contributed in creation of the architecture of Database (MYSQL).

Sahil Ukarde with the help of group members contributed in optimizing the Algorithm for Plagiarism checking.

The overall Project was completed before the deadline and the attendance of all the members was satisfactory.

3. Proposed System

3.1 Introduction

The proposed system is a comprehensive solution that combines plagiarism detection, assignment management, and team collaboration features into a single platform for academic institutions. It aims to provide a user-friendly and customizable tool for teachers to create, manage, and grade assignments while also giving students the ability to submit assignments in PDF format. The plagiarism detection feature of the system uses advanced algorithms and techniques to provide accurate and reliable results and includes a comprehensive database of scholarly sources and online repositories to ensure that all sources are properly identified and credited.

In addition, the system includes a team collaboration feature that allows students and teachers to communicate and collaborate on assignments, discuss course materials, and share feedback. This feature is designed to enhance the learning experience for students and facilitate better communication and collaboration between students and teachers.

Overall, the proposed system aims to address some of the limitations of existing plagiarism detection tools and provide a more effective and efficient solution for academic institutions. By providing a comprehensive platform that combines plagiarism detection, assignment management, and team collaboration features, the proposed system offers a streamlined and intuitive tool that can improve the academic experience for both students and teachers.

3.2 Architecture / Framework

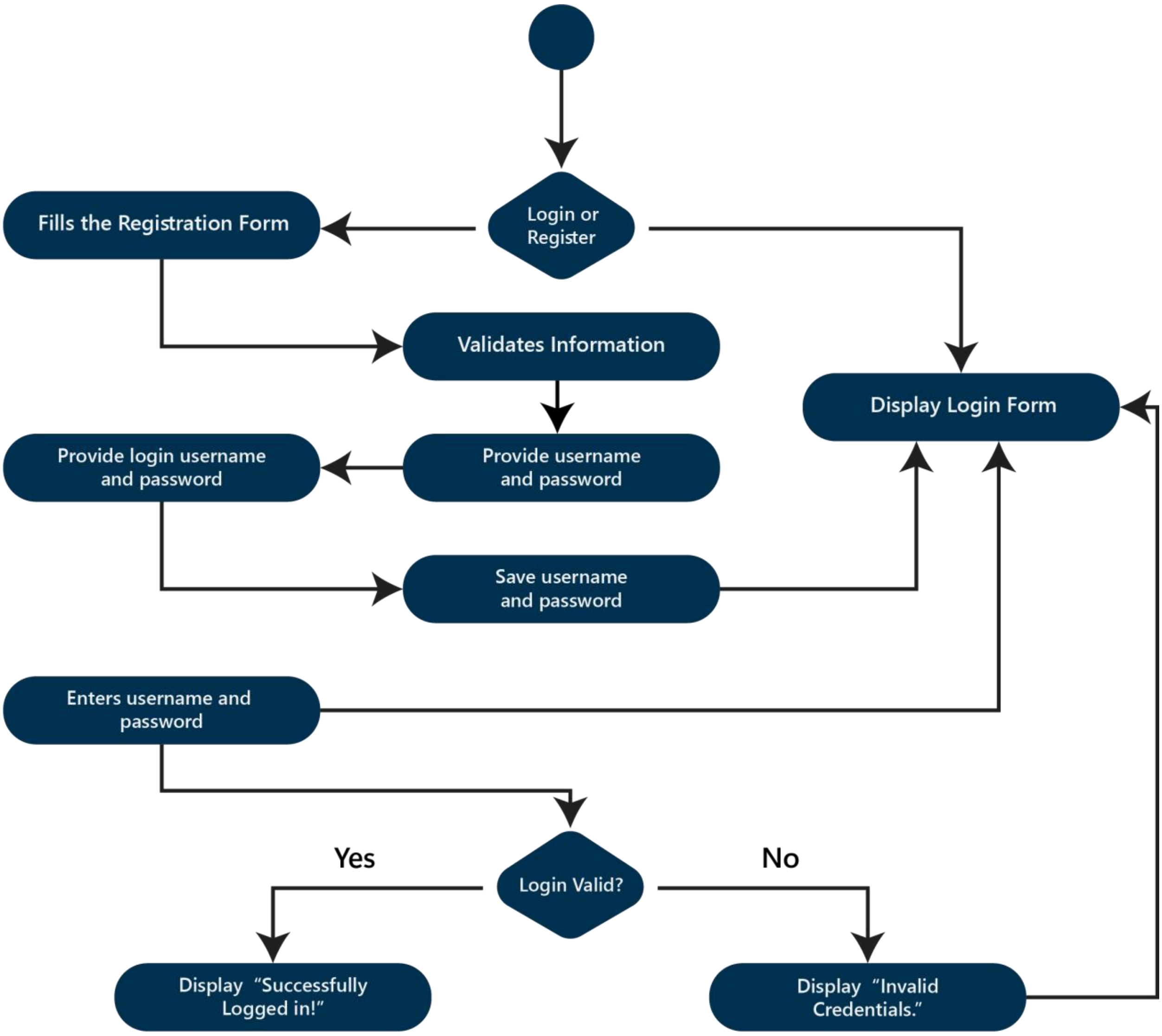
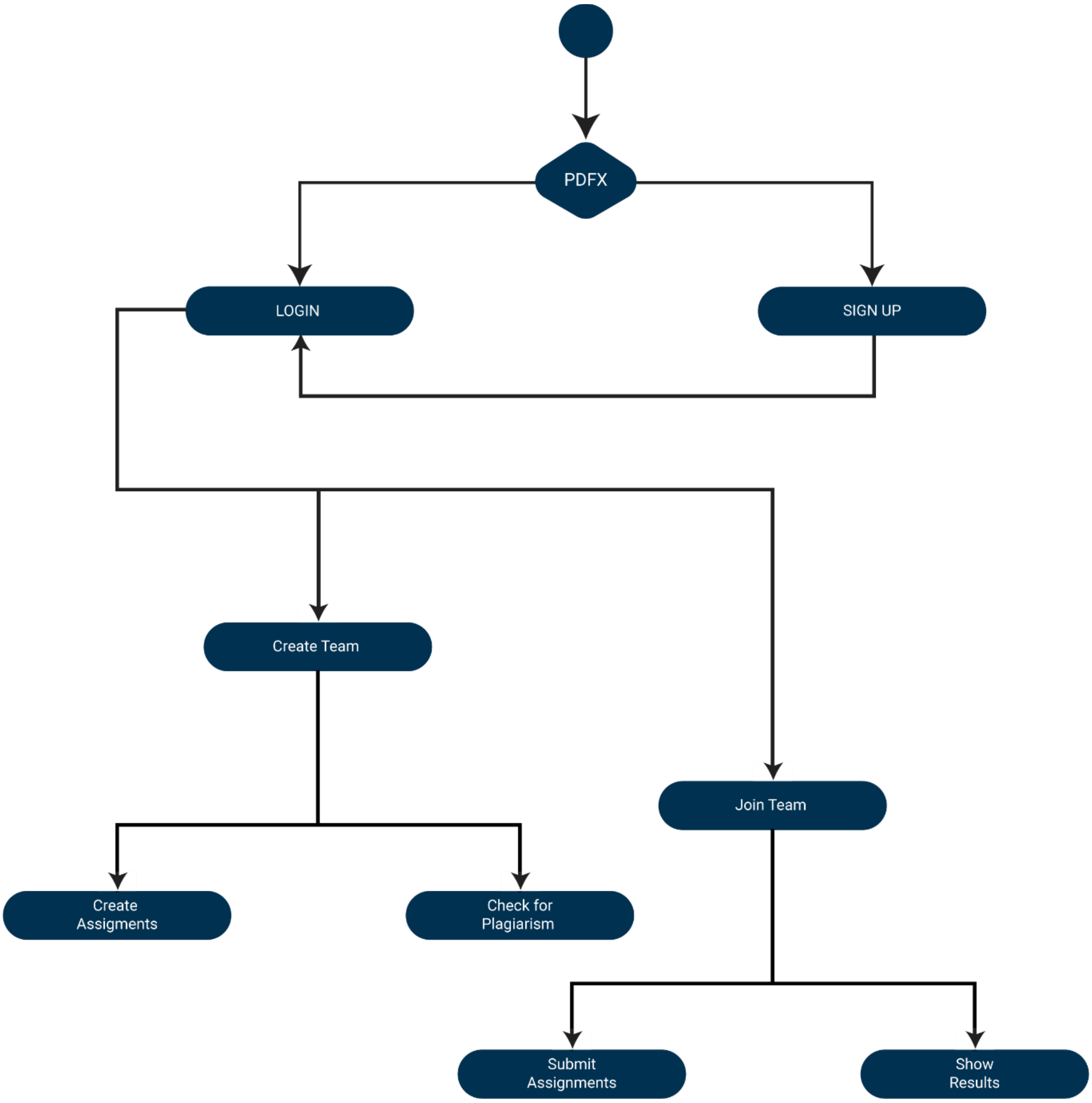

Login or Register
Fills the Registration Form
Validates Information
Display Login Form
Provide login username and password
Provide username and password
Save username and password
Enters username and password
Yes
Login Valid?
No
Display "Successfully Logged in!"
Display "Invalid Credentials."

Fig. 3.2.1 Registration/Login Process

Flow of Program



3.3 Algorithm and Process Design

We will focus on the following set of requirements while designing the **PDFX**:

- A teacher can create teams for their students after logging into the application.
- The teacher can then create assignments and check plagiarism of uploaded assignments or external PDFs.
- To submit an assignment, a student must log in and join their team.
- Once joined, students can submit their assignments and check their marks.
- The marks are based on the level of plagiarism detected.

3.4 Details of Software and Hardware

Operating Environment:

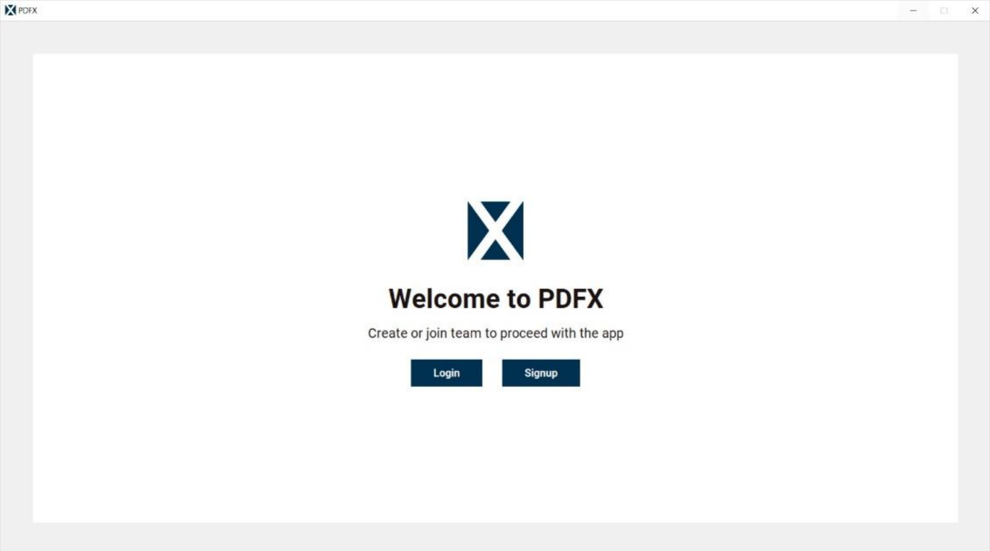
Hardware Requirement –

- 20 GB HDD Free Space
- 128 MB RAM
- P IV or above Processor
- Monitor or Keyboard
- Mouse

Software Requirement –

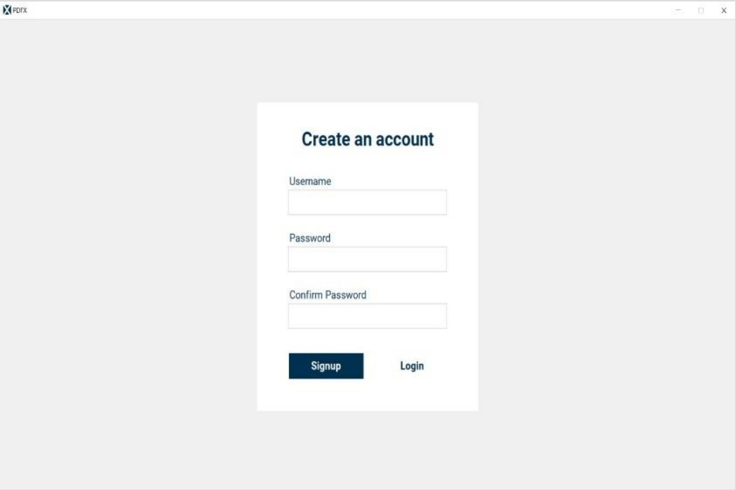
- VS Code
- MySQL
- Platform Used - Windows 10

3.5 Experiment & Results



Welcome to PDFX

Create or join team to proceed with the app

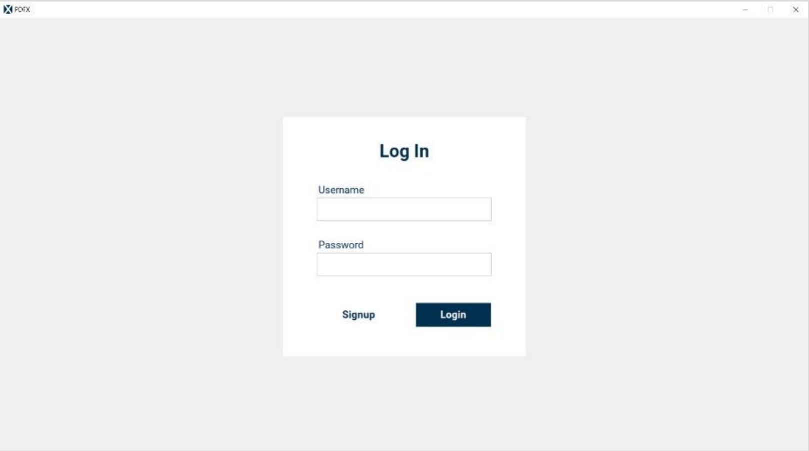


Create an account

Username

Password

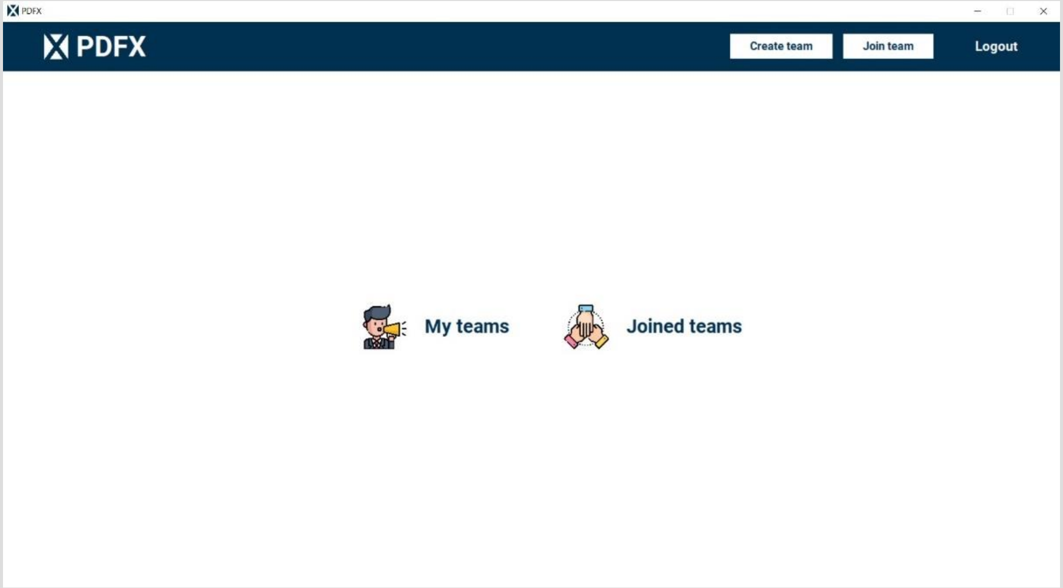
Confirm Password

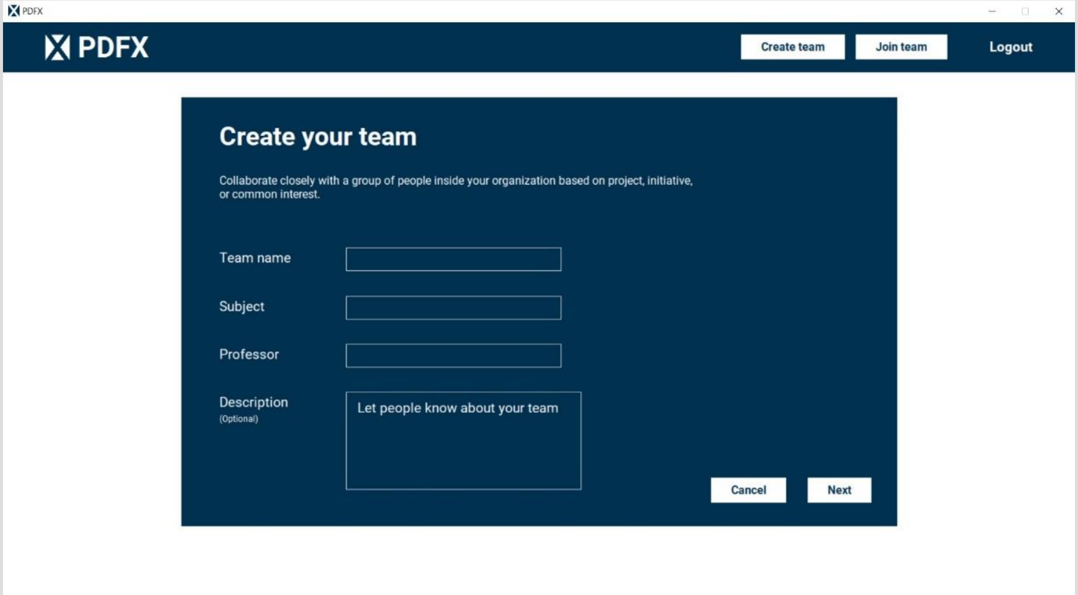


Log In

Username

Password





Create your team

Collaborate closely with a group of people inside your organization based on project, initiative, or common interest.

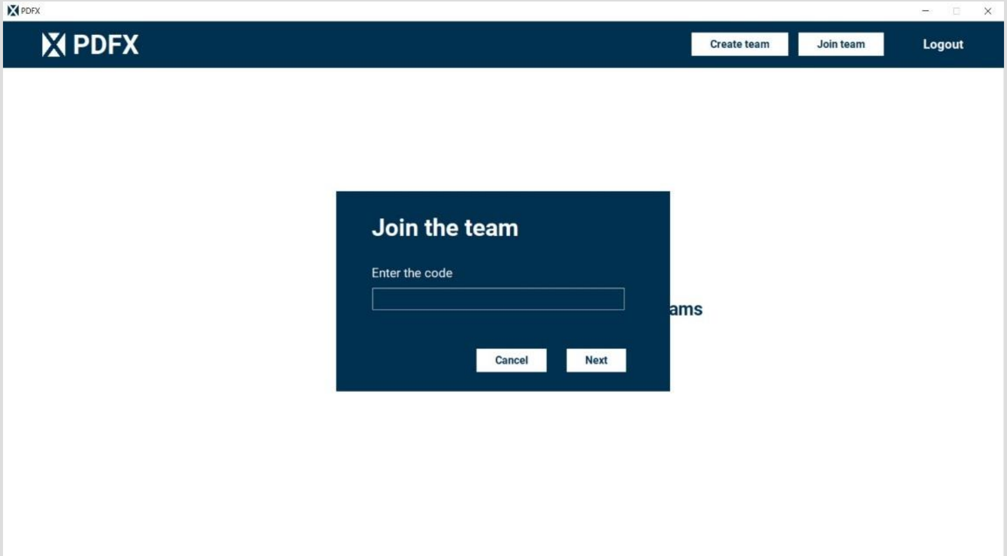
Team name

Subject

Professor

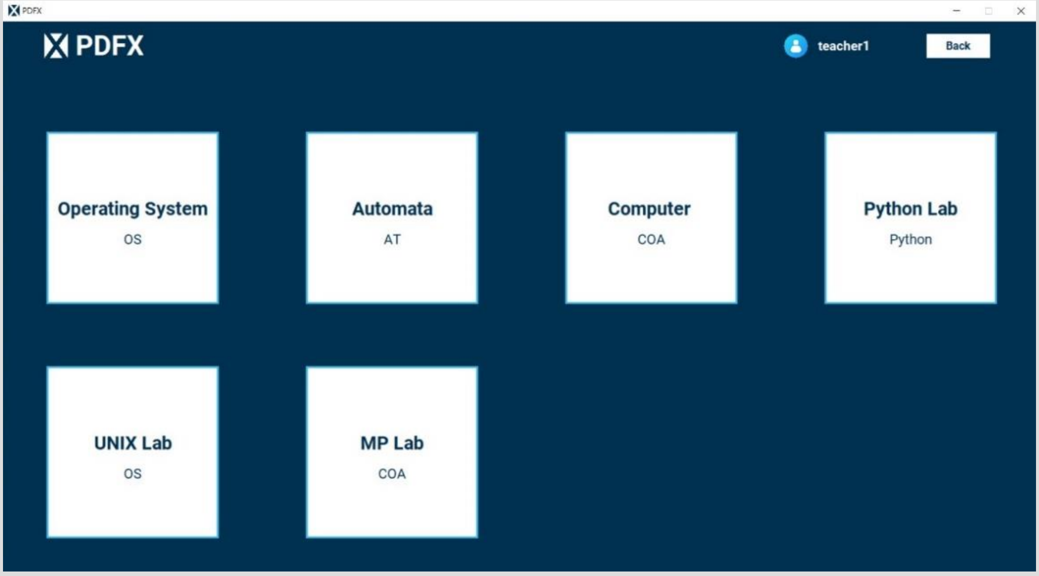
Description
(Optional)

Let people know about your team



Join the team

Enter the code



Operating System
OS

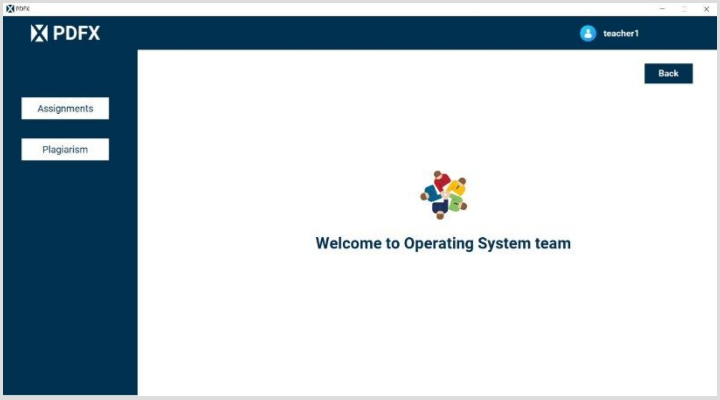
Automata
AT

Computer
COA

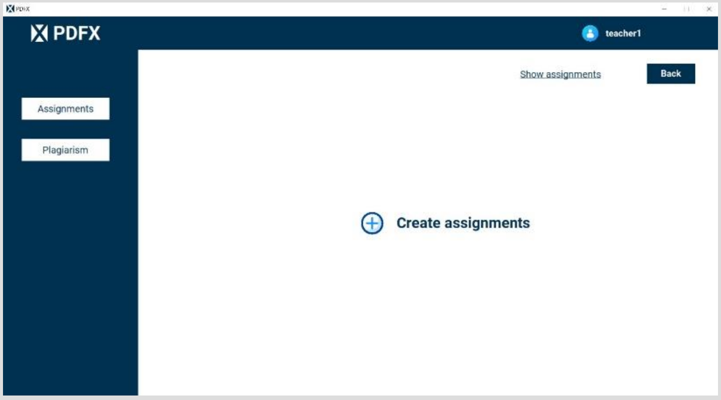
Python Lab
Python

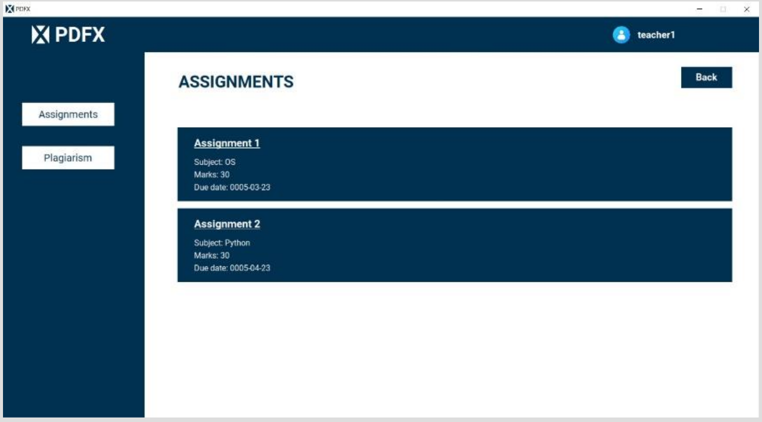
UNIX Lab
OS

MP Lab
COA



Welcome to Operating System team

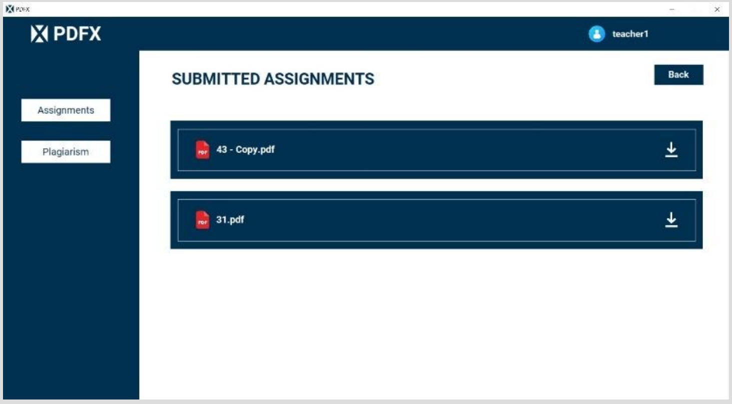




ASSIGNMENTS

Assignment 1
Subject: OS
Marks: 30
Due date: 0005-03-23

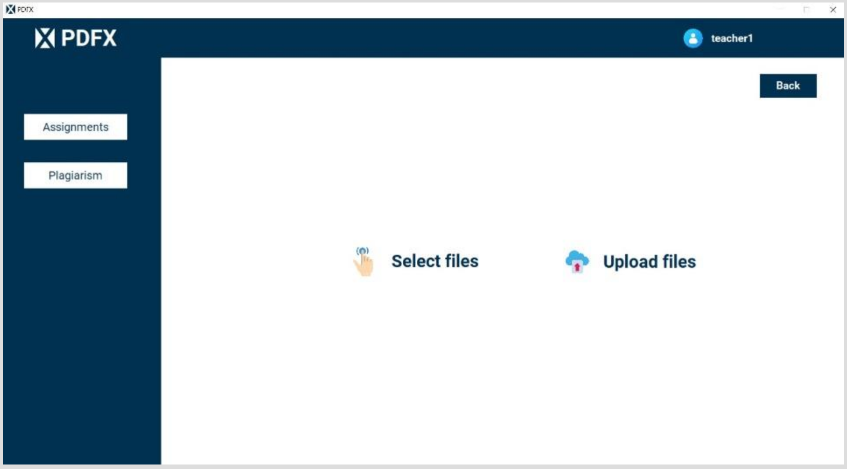
Assignment 2
Subject: Python
Marks: 30
Due date: 0005-04-23

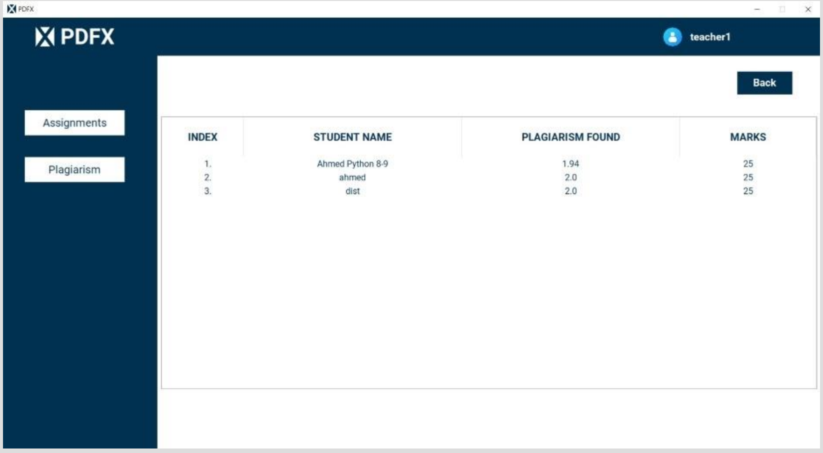


SUBMITTED ASSIGNMENTS

43 - Copy.pdf

31.pdf





INDEX	STUDENT NAME	PLAGIARISM FOUND	MARKS
1.	Ahmed Python 8-9	1.94	25
2.	ahmed	2.0	25
3.	dist	2.0	25

3.6 Conclusion and Future Work

Conclusion:

In conclusion, our PDF plagiarism detection application provides a user-friendly and efficient solution for teachers and students to identify and prevent plagiarism. With features such as team creation, assignment submission, plagiarism checking, and database storage, our application offers a comprehensive and streamlined approach to academic integrity. We believe that our application will prove to be a valuable tool in promoting honest and original work in academic settings.

Future Work:

Advanced plagiarism detection algorithms: Our current plagiarism detection algorithm could be improved to include more sophisticated techniques, such as natural language processing or machine learning, to increase accuracy and reduce false positives.

Web scraping for online sources: In addition to checking PDF files, our application could incorporate web scraping techniques to detect plagiarism from online sources such as websites, articles, and publications. This would enhance the effectiveness of our application and provide a more comprehensive plagiarism detection solution.

References

[1] The Python Software Foundation. [Online]. Available: https://www.python.org/.

[2] Tkinter documentation. [Online]. Available: https://docs.python.org/3/library/tk.html.

[3] PyPDF2 documentation. [Online]. Available: https://pypdf2.readthedocs.io/en/3.0.0/

[4] difflib documentation. [Online]. Available: https://docs.python.org/3/library/difflib.html.

[5] pytesseract documentation. [Online]. Available: https://pypi.org/project/pytesseract/.

[6] os module documentation. [Online]. Available: https://docs.python.org/3/library/os.html

[7] MySQL documentation. [Online]. Available: https://dev.mysql.com/doc/.

[8] Turnitin plagiarism checker. [Online]. Available: https://www.turnitin.com/.