SHRI G. S. INSTITUTE OF TECHNOLOGY AND SCIENCE

DEPARTMENT OF COMPUTER ENGINEERING

BTECH – III YEAR

FORMAT VALIDATION AND DOCUMENT SEGREGATION

Submitted By – Nishant Gupta 0801CS221101	Submitted To – Mr. Surendra Gupta Ms. Ashwini Sharma

1. Introduction

1.1 Purpose

The purpose of this system is to create an automated process for validating and segregating PDF documents based on pre-defined formatting rules and subject codes. The application will:

1. Validate PDF files against specific formatting criteria (font style, size, and sequence).
2. Extract subject codes to determine the branch and year.
3. Segregate the files into branch- and year-specific folders.

1.2 Scope

This document segregation system is focused on:

- Validating PDF files for adherence to formatting rules such as font type and size.
- Ensuring subject codes are correctly formatted and match a specific pattern.
- Organizing files into proper directories based on subject code extraction.

1.3 Definitions, Acronyms, and Abbreviations

- PDF: Portable Document Format.
- Flask: A Python-based web framework used for creating the web interface.
- PyPDF2: A Python library to work with PDF files.

1.4 Overview

This document outlines the system requirements, functionality, and constraints for the document segregation system. The system will provide a user-friendly interface to upload PDF files and receive feedback on format validation.

2. System Overview

2.1 System Features

1. File Upload: Users can upload PDF files through the web interface.
2. Format Validation: The system validates the uploaded file based on predefined format rules:
    - College name must be on the first line, in Arial font, size 18, and bold.

    - Subject code must be on the second line, in Arial font, size 16, and bold, and follow a specific pattern.
    - Body text must use Arial font, size 14 or less.
3. File Segregation: Files are segregated into directories based on extracted subject codes.

3. Functional Requirements

3.1 File Upload Module

- Description: Users will upload PDF files through a web interface.
- Input: A PDF document.
- Output: A message indicating whether the file passed format validation and, if so, the file is moved to the appropriate folder.

3.2 Format Validation Module

- Description: The system will validate the format of the uploaded PDF file. It checks the following:
    - First Line: The college name must be "Shri G. S. Institute of Technology and Science", in Arial font, size 18, and bold.
    - Second Line: The subject code must follow a valid format (e.g., "ABXXXXX"), in Arial font, size 16, and bold.
    - Remaining Body: All body text must be in Arial font, size 14 or less.
- Input: The uploaded PDF file.
- Output: A message indicating if the file passed or failed the format checks.

3.3 File Segregation Module

- Description: After format validation, the file will be moved to a folder based on the subject code's year and branch.
    - Branch codes: e.g., CO (Computer Engineering), EE (Electrical Engineering), etc.
    - Year codes: e.g., 1 (First Year), 2 (Second Year), etc.
    - Even/odd semester folders: Based on the 4th character of the subject code (even/odd).
- Input: A valid PDF file with a correctly formatted subject code.
- Output: The file will be moved to a specific directory based on year, branch, and semester.

4. Non-Functional Requirements

4.1 Performance

- The system should be able to handle multiple PDF file uploads concurrently.
- Format checks should be performed within 2 seconds of file upload.

4.2 Usability

- The web interface should be simple, with clear instructions for users to upload files.
- If a format error is found, an informative message should be displayed to the user.

4.3 Security

- The application should restrict uploads to PDF files only.

5. External Interface Requirements

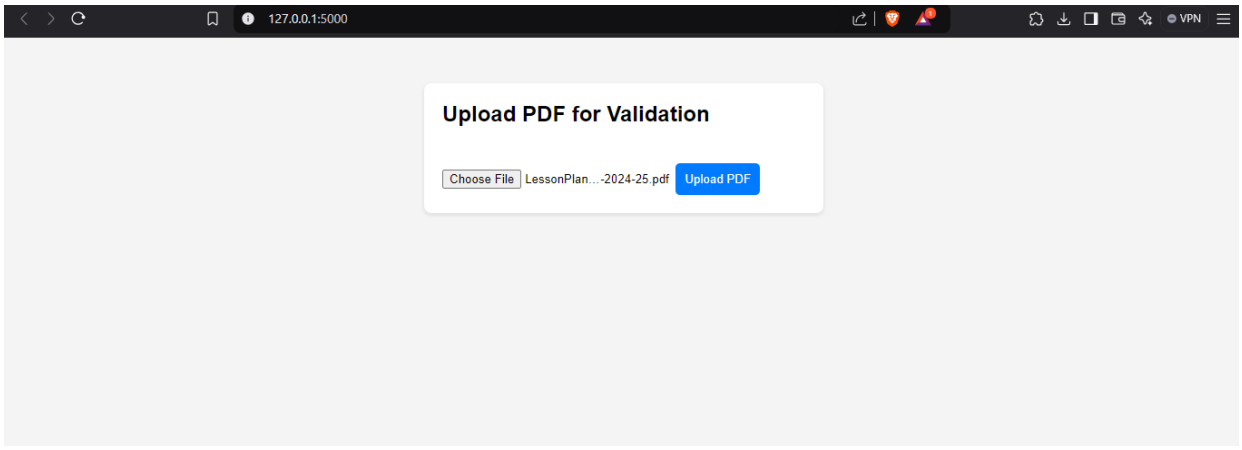
5.1 User Interface

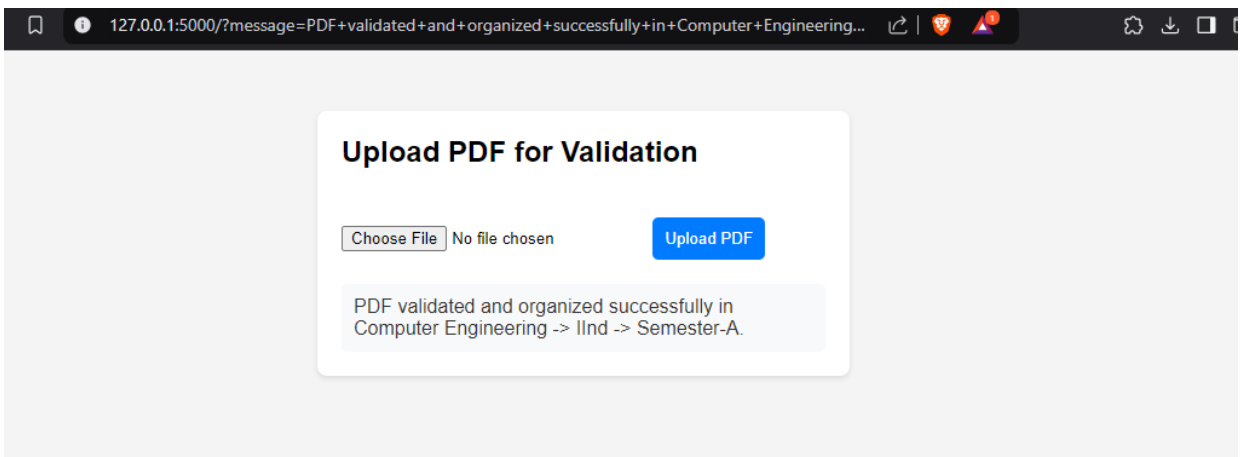
- Upload Form: A web form for users to select and upload PDF files.
- Feedback Messages: Notifications for users to indicate if the file passed validation or failed due to format errors.

5.2 Hardware Interface

No specific hardware interface requirements**.**

5.3 Software Interface





- The system interacts with the file system for storing uploaded PDFs and segregating them into folders.
- PyPDF2 is used to extract and analyse the PDF content.

6. System Design Constraints

- The system should be deployable on any server that supports Python and Flask.
- PyPDF2 will be used for PDF reading and validation.