Project Title: **Result Analysis**

Name: Hitesh Taneja

Roll No: 18058570018

University: University of Delhi

Course: Bachelor of Computer Science [Hons.]

Subject: Software Engineering

# Abstract

This thesis presents "Result Analysis" an innovative software tool designed to automate the extraction and analysis of academic results from PDF documents, specifically those published by the University of Delhi. The project addresses the challenge of manually processing extensive result data, a task that is traditionally time-consuming and prone to errors. Utilizing Python, the software efficiently converts PDF content into structured CSV format, enabling easy manipulation and analysis of data.

The core functionality of the tool involves parsing PDF files to extract student results, followed by cleaning and organizing the data into a CSV format. This process is made possible through a series of Python scripts, including 'Extract\_Combine.py' for data extraction and combination, 'Search\_Delete.py' for data cleaning, and 'getCollege.py' along with 'parse.py' for targeted data retrieval based on college codes. The culmination of the process is the generation of a pie chart through 'pie\_chart.py', which visually represents the performance distribution of students.

Significantly, the project not only streamlines data processing but also opens avenues for detailed academic result analysis. The potential applications of this tool extend to automated scorecard generation and broader adaptability to various PDF formats. This thesis details the architectural design, development process, and the practical utility of "Result Analysis" demonstrating its effectiveness in simplifying and enhancing the analysis of academic results.

Table of Contents

[Abstract 2](#_Toc154465195)

[Introduction 4](#_Toc154465196)

[Literature Review 5](#_Toc154465197)

[Data Visualization in Academic Research 5](#_Toc154465198)

[Introduction to Data Visualization 5](#_Toc154465199)

[Guidelines for Effective Visualizations 5](#_Toc154465200)

[Applying Visualization Principles 5](#_Toc154465201)

[The Role of Technology in Educational Administration 5](#_Toc154465202)

[Technological Transformation in Education 5](#_Toc154465203)

[Challenges and Opportunities 5](#_Toc154465204)

[Benefits of Technology in Education 6](#_Toc154465205)

[PDF Data Extraction Technologies 6](#_Toc154465206)

[Background on PDF Data Extraction 6](#_Toc154465207)

[Review of Existing Methods and Tools 6](#_Toc154465208)

[Challenges in PDF Data Extraction 6](#_Toc154465209)

[Methodology 7](#_Toc154465210)

[Project Architecture 7](#_Toc154465211)

[Structure Chart 7](#_Toc154465212)

[Integration and Workflow 9](#_Toc154465213)

[Results 10](#_Toc154465214)

[Key Functionalities Demonstrated: 10](#_Toc154465215)

[Code Snippet: 10](#_Toc154465216)

[**The Use of 123:135 in extract\_text** 10](#_Toc154465217)

[Discussion 12](#_Toc154465218)

[Effectiveness of Design: 12](#_Toc154465219)

[Meeting Stated Requirements: 12](#_Toc154465220)

[Conclusion 13](#_Toc154465221)

[Future Work 13](#_Toc154465222)

[Potential Enhancements: 13](#_Toc154465223)

# Introduction

Background and Motivation

In the era of digital information, the efficient processing and analysis of data have become paramount, especially in educational institutions where handling extensive academic records is a regular task. The University of Delhi, like many universities, publishes student results in PDF format, which, while universally accessible, poses challenges in data extraction and analysis. The traditional approach of manually transcribing this data into more manageable formats is not only time-consuming but also prone to errors, leading to inefficiencies in academic administration and result analysis.

Objective of the Project

The primary objective of "Result Analysis" is to address these challenges by developing a tool that automates the extraction of result data from PDF documents and facilitates its analysis. The specific goals of the project are to:

* Convert the PDF results data into a structured CSV format for ease of manipulation.
* Provide a mechanism to filter and clean the data based on specific criteria, such as college codes.
* Visualize the performance distribution of students through pie charts, enabling a clearer understanding of the overall academic achievement within the university.

Significance of the Project

This project holds significant value in the context of academic data management. By automating the process of data extraction and analysis, "Result Analysis" not only saves time and reduces human error but also provides a platform for more nuanced and detailed analysis of student performance. The software's ability to transform cumbersome PDF files into analyzable formats can be a game-changer for administrative departments in educational institutions. Additionally, the visualization aspect of the project, through pie charts, offers an intuitive and immediate way to interpret the data, which is beneficial for quick decision-making and reporting.

Project Overview

The project is structured as a series of Python scripts, each serving a specific function in the process of extracting, cleaning, and visualizing data. The tool begins by scanning PDF files for relevant result data, then parses and converts this data into CSV format. Subsequent scripts are used to clean the data and generate pie charts, providing a comprehensive analysis of the results. This thesis will detail the architectural design, implementation, and functionality of each component of "Result Analysis".

# Literature Review

## Data Visualization in Academic Research

### Introduction to Data Visualization

Data visualization plays a pivotal role in academic research, enabling complex data to be presented in a comprehensible and visually appealing manner. Edward Tufte, a pioneer in the field, emphasizes the importance of displaying information clearly and efficiently using various visual techniques. His principles have guided the creation of visual representations that communicate data succinctly and effectively, making them indispensable in academic settings where complex information needs to be conveyed to a diverse audience.

### Guidelines for Effective Visualizations

In line with Tufte's principles, Luís Cruz (2021) outlines several guidelines for creating effective data visualizations in academic papers. These guidelines stress the importance of maximizing the data-ink ratio, ensuring every visual element serves a purpose in conveying information. Cruz advocates for the judicious use of visual variables such as color, shape, and shade, and advises against the use of pie charts for representing complex data sets due to difficulties in comparing multiple slices [32-49].

### Applying Visualization Principles

Applying these principles in academic contexts, particularly in projects like "Result Analysis" is crucial for effective communication. These guidelines help in transforming raw data extracted from educational PDFs into insightful visualizations, making complex data patterns understandable and engaging for academic and administrative purposes.

## The Role of Technology in Educational Administration

### Technological Transformation in Education

Advances in technology have significantly impacted educational administration. The International Society for Technology in Education (ISTE) highlights the creation of high-demand jobs influenced by technological advancements, underscoring the necessity for educators to adapt and equip students with skills for future careers [58].

### Challenges and Opportunities

Integrating technology in education presents both challenges and opportunities. According to BuiltIn and the World Economic Forum, while the educational technology market is expected to grow significantly, challenges in implementation and access remain. Concerns include the effective use of emerging technologies like AR and AI, and ensuring equitable technology access across different educational settings [59, 60].

### Benefits of Technology in Education

Despite these challenges, the benefits of technology in education are manifold. As noted by American University, technology enhances collaboration, communication, and personalized learning opportunities. It also plays a vital role in improving the quality of education and engaging students in creative ways, thereby fostering a deeper understanding and curiosity in various subjects [61-63].

## PDF Data Extraction Technologies

### Background on PDF Data Extraction

PDF data extraction is increasingly relevant in both academic and professional realms. The process involves retrieving information from PDF files, a common format for disseminating academic papers, reports, and administrative documents.

### Review of Existing Methods and Tools

Various methods and tools have been developed for PDF data extraction, ranging from simple text extraction to more complex data mining techniques. These technologies vary in their approach, accuracy, and complexity, catering to different types of PDF documents, from text-based to scanned images.

### Challenges in PDF Data Extraction

One of the primary challenges in PDF data extraction is dealing with the variety of formats and the inconsistent structuring of data within PDFs. Current technologies address these challenges with varying degrees of success, employing advanced algorithms and machine learning techniques to improve accuracy and efficiency.

# Methodology

## Project Architecture

The "Result Analysis" project is designed to automate the process of extracting and analyzing academic results from PDF files. The architecture of the project is built around a series of Python scripts, each fulfilling a specific function in the extraction, cleaning, and visualization of data. The project operates on a modular approach, with each module or script designed to handle different aspects of the data processing pipeline.

## Structure Chart

The structure of the program is conceptualized through a series of interconnected modules, representing the workflow from the initial input to the final output. Here is an overview of the structure chart:

* PDF-File Input: The process begins with the input of a PDF file, typically containing university results.
* get\_College Script: This script searches the PDF file for specific college codes provided as input, identifying relevant pages or sections.
* Parse Script: The identified data is then passed to the parse script, which converts and extracts the data into a preliminary CSV format.
* Data Cleaning: The CSV file undergoes cleaning through the 'Delete Specific Rows' script, which removes unwanted data, and the 'Get specific Columns' script, which isolates relevant columns.
* Pie Chart and Result CSV Generation: Finally, the cleaned data is used to generate two outputs:
* A pie chart visualizing the data distribution.
* A refined result CSV file containing the processed data.

This structure chart [figure 1] visually represents the flow and interconnection of the various components of the "Result Analysis" project.

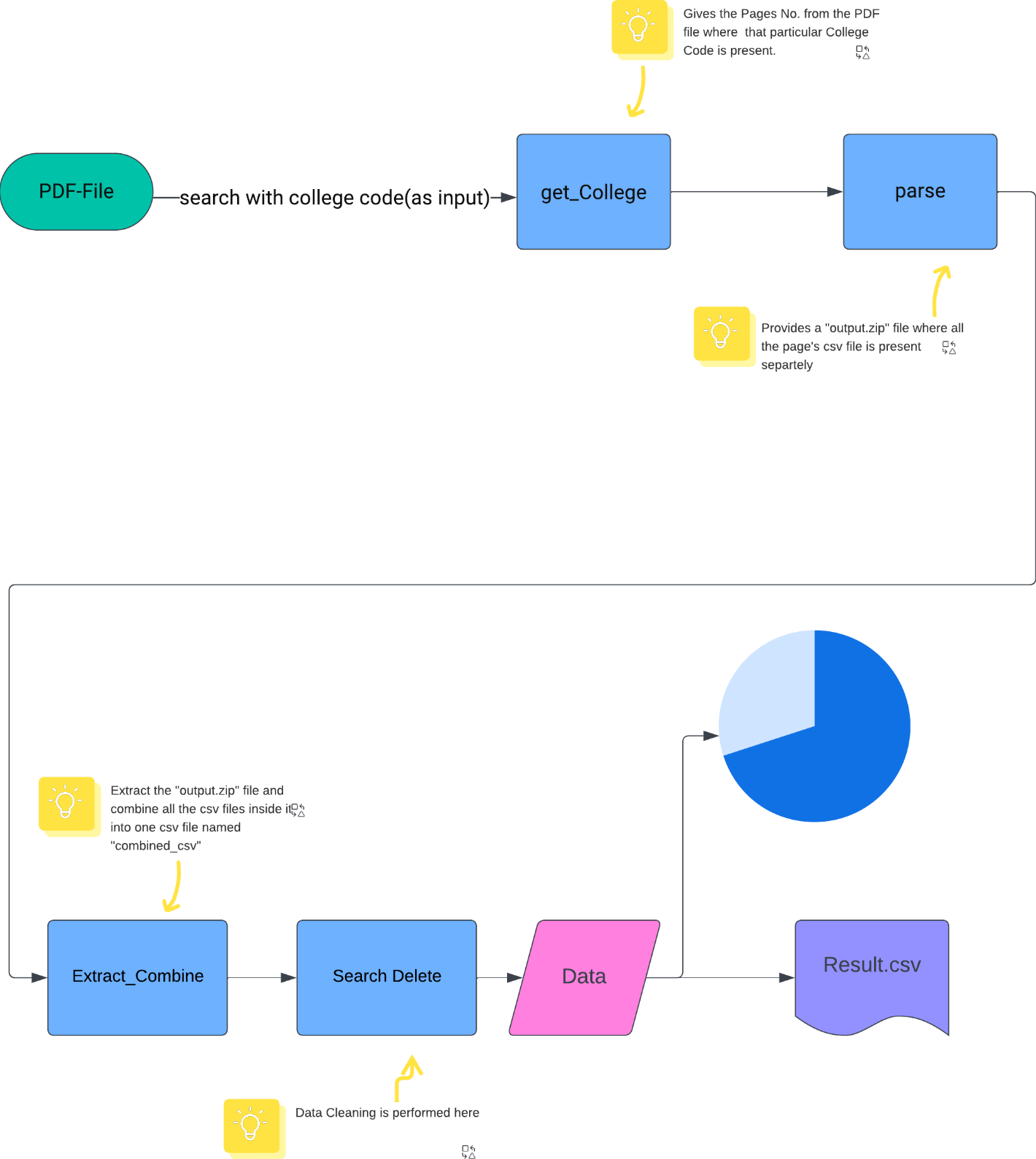


Figure 1

## Integration and Workflow

The methodology also describes how these modules integrate and function together. The workflow is sequential, with each script feeding its output into the next stage of the process. This modular design allows for flexibility and ease of maintenance, as each part of the project can be modified or enhanced independently.

The "Result Analysis" project, through its structured and modular architecture, demonstrates an effective approach to automating the task of data extraction and analysis from PDF files, specifically tailored to the needs of academic result processing.

# Results

The "PDF\_Scrapper-Pie-Chart" project successfully achieved its goal of automating the extraction and analysis of academic results from PDF files. The outcomes of the project are multi-faceted, demonstrating efficiency in data processing, accuracy in extraction, and clarity in visualization.

## Key Functionalities Demonstrated:

* PDF Data Extraction: The project effectively identifies and extracts relevant data from extensive PDF files, particularly those published by the University of Delhi.
* Data Cleaning and Organization: Extracted data is cleaned and structured into a usable format (CSV), facilitating easy manipulation and analysis.
* Data Visualization: The project creates intuitive pie charts that visually represent the distribution of academic performances, making it easier to comprehend and analyze the data.

## Code Snippet:

The getCollege function in the project represents a critical part of the data extraction process:

def getCollege(path, code):

pages = []

try:

with open(path, 'rb') as pdfFileObj:

pdfReader = PyPDF2.PdfReader(pdfFileObj)

for pageNO in range(len(pdfReader.pages)):

pageObj = pdfReader.pages[pageNO]

revInfo = pageObj.extract\_text()[123:135]

if code in revInfo:

print(f"Code {code} found on page {pageNO+1}")

pages.append(int(pageNO)+1)

except Exception as e:

print(f"Error reading PDF file: {e}")

return pages

This function is pivotal for locating and extracting data specific to a college based on the provided code.

### **The Use of 123:135 in extract\_text**

The numbers 123:135 are used to slice a specific substring from the extracted text of each page. This slicing is based on the observation that the relevant information (in this case, the college code) is consistently located in the same position on each page of the PDF.

#### Determining the Slice Range:

This specific range (123:135) was likely determined through an analysis of the PDF's structure.

You might have manually reviewed the PDF and found that the college code consistently appears within this character range in the text extracted from each page.

#### Customized to Specific PDF Structure:

This approach suggests that the PDF has a consistent format where the relevant data appears in the same position on each page.

It is a tailored solution for this specific PDF format and may need adjustment if the structure of the PDF changes or if the function is applied to a different PDF with a different format.

#### Presentation of result.csv

The **result.csv** file serves as the culmination of the data extraction and processing stages of the "PDF\_Scrapper-Pie-Chart" project. This CSV file is structured into several columns, each representing a specific piece of information extracted from the PDF files. The primary columns include 'Roll No', 'Name', and 'CGPA', along with additional columns for each semester or assessment period, such as 'CGPA 1', 'CGPA 2', etc. Each row corresponds to an individual student's academic performance data.



Figure 2

#### Display of the Pie Chart

The pie chart is an integral part of the project's data visualization component. It represents the distribution of grades across a spectrum, categorized into segments such as 'Below 4', 'Below 7 and Above 4', 'Above 7 and Below 9', and 'Above 9 or Equal to 10'. This visual distribution allows for immediate insight into the academic performance of the student cohort, highlighting areas of excellence and those that may require additional support or intervention.

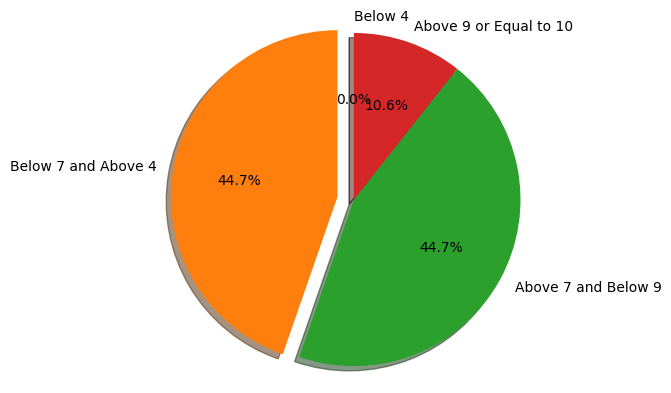


Figure 3

# Discussion

## Effectiveness of Design:

The project's modular design proved effective in handling different stages of data processing, from extraction to visualization.

The architecture facilitated easy modification and troubleshooting, which was particularly useful in adapting the scripts to handle various formats within the PDFs.

## Meeting Stated Requirements:

The project successfully met the requirement of reducing manual data entry and processing time, which is a significant challenge in academic administrations.

The generated pie charts and CSV files provided clear and actionable insights, aiding in the efficient analysis of academic results.

# Conclusion

The "Result Analysis" project represents a significant advancement in the field of academic data management. By automating the process of extracting, cleaning, and visualizing data from cumbersome PDF files, the project not only saves time and reduces human error but also provides a platform for more insightful analysis of academic results. Its successful implementation demonstrates the potential of automated tools in enhancing the efficiency and effectiveness of administrative processes in educational institutions.

# Future Work

## Potential Enhancements:

* Automating Scorecard Generation: Future iterations of the project could include the development of an automated system for generating individual student scorecards, further reducing manual workload.
* Adding a Print Feature: Incorporating a direct print functionality for the generated reports and pie charts could streamline the process of disseminating this information.
* Generalizing the Application: Broadening the scope of the project to handle various PDF formats and structures would make the tool more versatile and applicable to a wider range of documents and use cases.

These enhancements would not only increase the utility of the "PDF\_Scrapper-Pie-Chart" but also expand its applicability to a broader spectrum of data extraction and analysis tasks in academic and other professional settings.

# References

* <https://luiscruz.github.io/2021/03/01/effective-visualizations.html#:~:text=Visualization%20are%20very%20heterogenous.,you%20create%20effective%20information%20visualizations>.
* https://www.exam.du.ac.in/?Results

# Appendices

* <https://github.com/hiteshtanejaa/PDF_Scrapper-Pie-Chart-/blob/main/result_analysis.ipynb>
* <https://github.com/hiteshtanejaa/Result-Analysis>