**Contextual Text Extraction from PDF files Using Machine Learning Practices**

CAPSTONE PROJECT REPORT

By

Team 16

E Saikiran [22R15A0514]

J Sinduja [21R11A0574]

Naveen Rampa [21R11A0589]

Lakshanya S [21R11A05H3]

A Bhavya Sri [21R11A05A5]



**Department of Computer Science and Engineering**

**Accredited by NBA**

**Geethanjali College of Engineering and Technology**

**(UGC Autonomous)**

(Affiliated to J.N.T.U.H, Approved by AICTE, New Delhi)

Cheeryal (V), Keesara (M), Medchal.Dist.-501 301

March 2024

**ABSTRACT**

Contextual text extraction from pdf files using machine learning is the very useful for extracting required information from pdf files. We can save a lot of time. Accessibility and readability of PDF files are very necessary for those who have vision issues or have trouble reading small or blurred text, useful for legal situations, data analysis, and research. Some instances where extraction is required include using text or image content from PDF files in other documents to save time and avoid mistakes. We are going to implement this using text extraction techniques using ML applications.**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Chapter Name** | **Page No** |
|  | **ABSTRACT** |  |
| **1** | **INTRODUCTION** | 4 |
| **2** | **PROBLEM STATEMENT** | 4 |
| **3** | **OBJECTIVES** | 5 |
| **4** | **METHODOLOGY** | 6 |
|  | **4.1 Data Source** | 6 |
|  | **4.2 Exploratory Data Analysis** | 7 |
| **5** | **Literature Review** | 8 |
| **6** | **Final Implementation** | 14 |
|  | **Conclusion** | 17 |

1. **INTRODUCTION**

The capstone project is focused on developing a solution to efficiently extract essential information from PDF documents. The goal is to automate the extraction process, which is crucial for tasks like accessibility improvement, legal document processing, data analysis, and research. The project specifically aims to extract the Name of the Director, Director type (Independent/Executive), and DIN (Director Identification Number) from PDFs using machine learning techniques.

The challenge lies in dealing with the unstructured nature of PDF data, where information is often scattered across pages and presented in various formats. Manual extraction is time-consuming, error-prone, and not scalable for large volumes of documents. By leveraging machine learning models, the project aims to overcome these challenges by training algorithms to recognize patterns, understand document layouts, and extract relevant data accurately.

The extracted information, such as director names, their DIN numbers, and whether they are independent or executive directors, will be used for various applications, including regulatory compliance, corporate governance analysis, and decision-making processes. The project's success will streamline data extraction workflows, improve data accuracy, and enhance overall efficiency in handling PDF documents.

**2. PROBLEM STATEMENT**

The extraction of specific information from PDF documents, such as corporate announcements, board meeting minutes, or regulatory filings, is a common challenge faced by many organizations. Manually extracting data from these documents is time-consuming, error-prone, and not scalable for large volumes of documents. The unstructured nature of PDF data, varying document layouts, and the need to extract specific information like director names, DIN (Director Identification Number), and their independent/dependent status pose significant challenges.

**3. OBJECTIVES**

Objective:

The objective of this capstone project is to develop an automated solution using machine learning techniques to extract essential information from PDF documents. Specifically, the project aims to extract the Name of the Director, Director type (Independent/Executive), and DIN from PDF files. The solution should be robust, accurate, and scalable to handle a variety of PDF document formats and layouts.

Challenges:

Unstructured Data: PDFs contain unstructured data that makes it challenging to extract specific information accurately.

Document Variability: PDF documents may have varying layouts, fonts, and structures, requiring the model to generalize well across different formats.

Data Accuracy: Ensuring high accuracy in extracting director names, DIN numbers, and their independent/dependent status is crucial for reliable results.

Scalability: The solution should be scalable to process large volumes of PDF documents efficiently.

Model Interpretability: The machine learning model should provide insights into how it makes predictions to ensure transparency and trustworthiness.

**4. METHODOLOGY**

**4.1 Data Source**

The data sources are the pdf files given. The information in the pdf file consist of Directors Details.

* File attachment [equitas.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/ae4a795e-e79f-45b4-b69e-323197c8c565/equitas.pdf) ( 329 KB; Apr 7, 2024, 8:27 pm )
* File attachment [Triveni2.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/eeab3b1d-2e6d-4370-b12f-9eb5e247b35b/Triveni2.pdf) ( 872 KB; Apr 7, 2024, 8:28 pm )
* File attachment [varun.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/0f2c41ce-b0be-42ab-a784-572f03911ebf/varun.pdf) ( 3 MB; Apr 7, 2024, 8:28 pm )
* File attachment [PolyPlex.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/d9f85fe4-42ba-4dbb-b65b-75facc15764c/PolyPlex.pdf) ( 471 KB; Apr 7, 2024, 8:27 pm )
* File attachment [Triveni1.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/67d16034-2396-404f-8546-77de027ef9cc/Triveni1.pdf) ( 9 MB; Apr 7, 2024, 8:27 pm )

The data souces are in the form of pdf . The datasource contains:

1. Name of the Director.

2. Director type : Independent / executive

3. DIN

Name of the Director: This refers to the full name of an individual serving as a director in a company or organization. The name typically includes both the first name and the last name of the director, such as "John Doe" or "Jane Smith."

Director Type (Independent/Executive): This parameter categorizes directors based on their role and relationship with the company's management:

Independent Director: These directors are not involved in the day-to-day operations of the company. They provide an independent perspective and are often part of the board's audit, nomination, and remuneration committees. They are expected to bring objectivity and impartiality to board decisions.

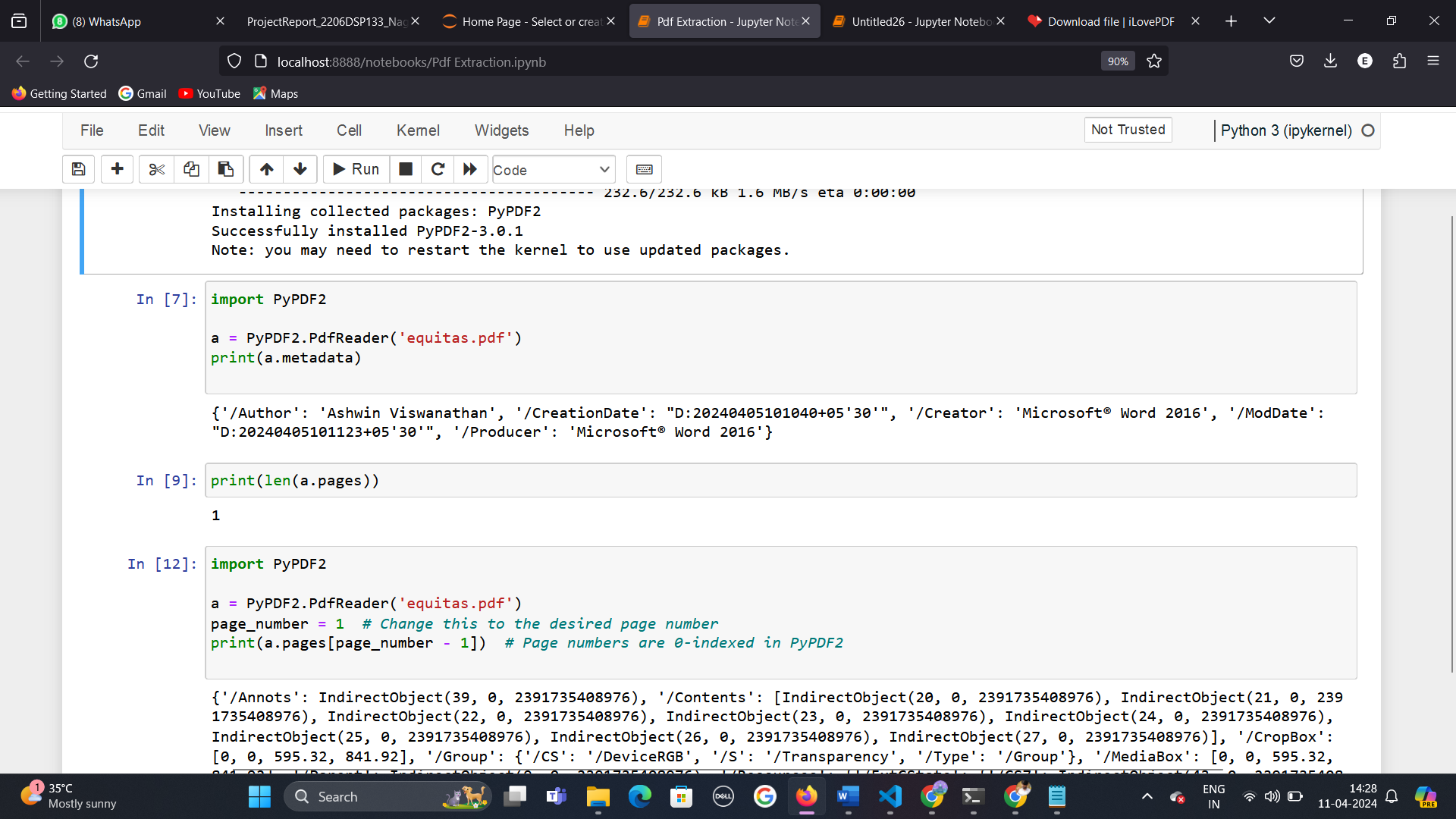
Executive Director: These directors are actively involved in the management and operations of the company. They may hold executive positions such as CEO, CFO, COO, etc., and are responsible for strategic decision-making and overseeing business functions.

DIN (Director Identification Number): DIN is a unique identification number assigned to individuals who serve as directors in companies. It is issued by the Ministry of Corporate Affairs in India and serves as a way to track and identify directors across different companies. DIN helps in maintaining a database of directors and their activities within the corporate sector.

**4.2 Exploratory Data Analysis**

While EDA in PDF extraction projects may not follow the traditional structured data EDA workflow, it serves the purpose of understanding the underlying data characteristics, preprocessing requirements, and quality assessment needed for successful extraction and model performance evaluation.

I took the pdf named ‘equitas.pdf’ to represent metadata and about pdf type of information.

****

Author: Ashwin Viswanathan

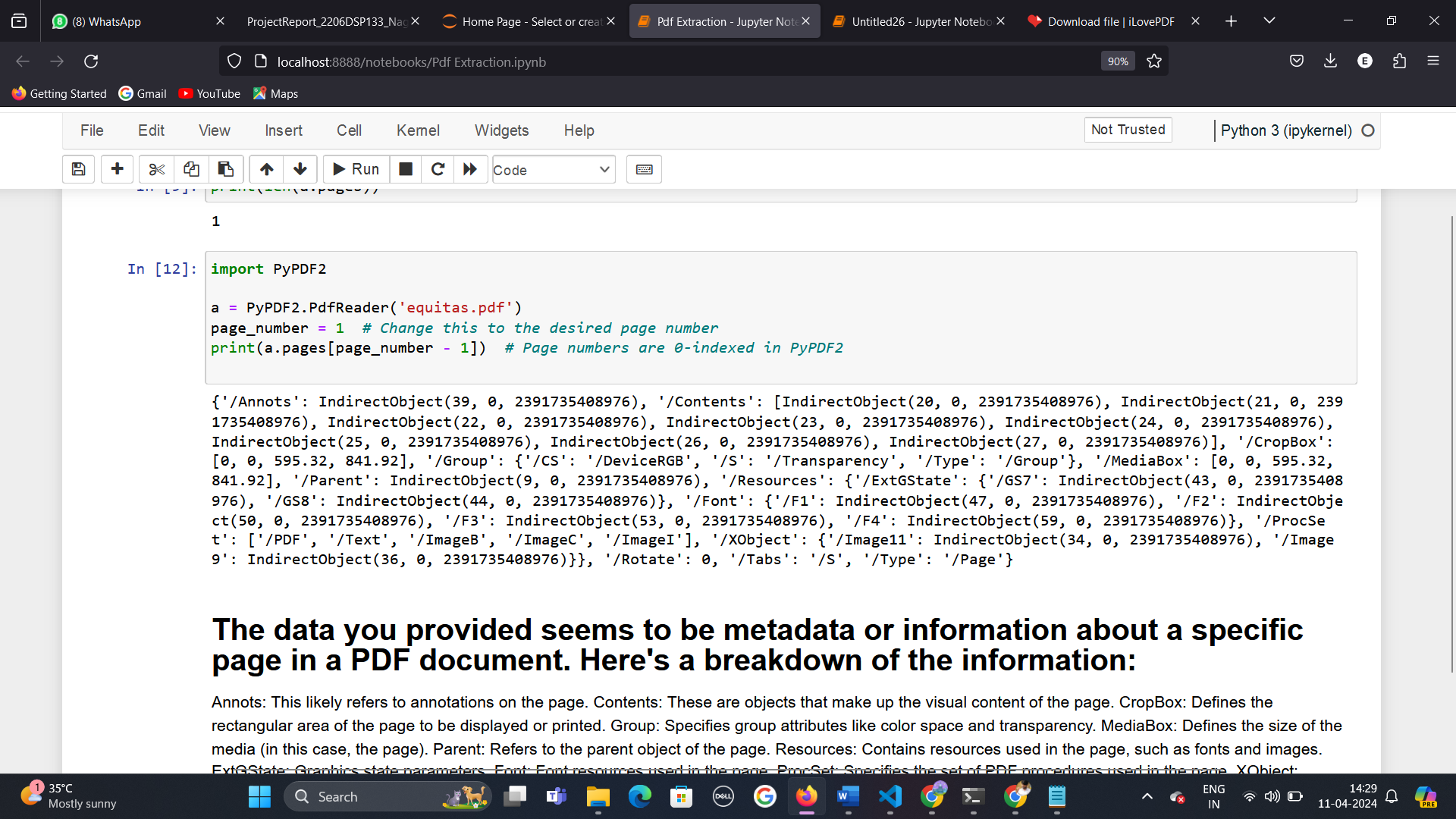
Creation Date: April 5, 2024, at 10:10:40 AM UTC+05:30

Creator: Microsoft® Word 2016

Modification Date: April 5, 2024, at 10:11:23 AM UTC+05:30

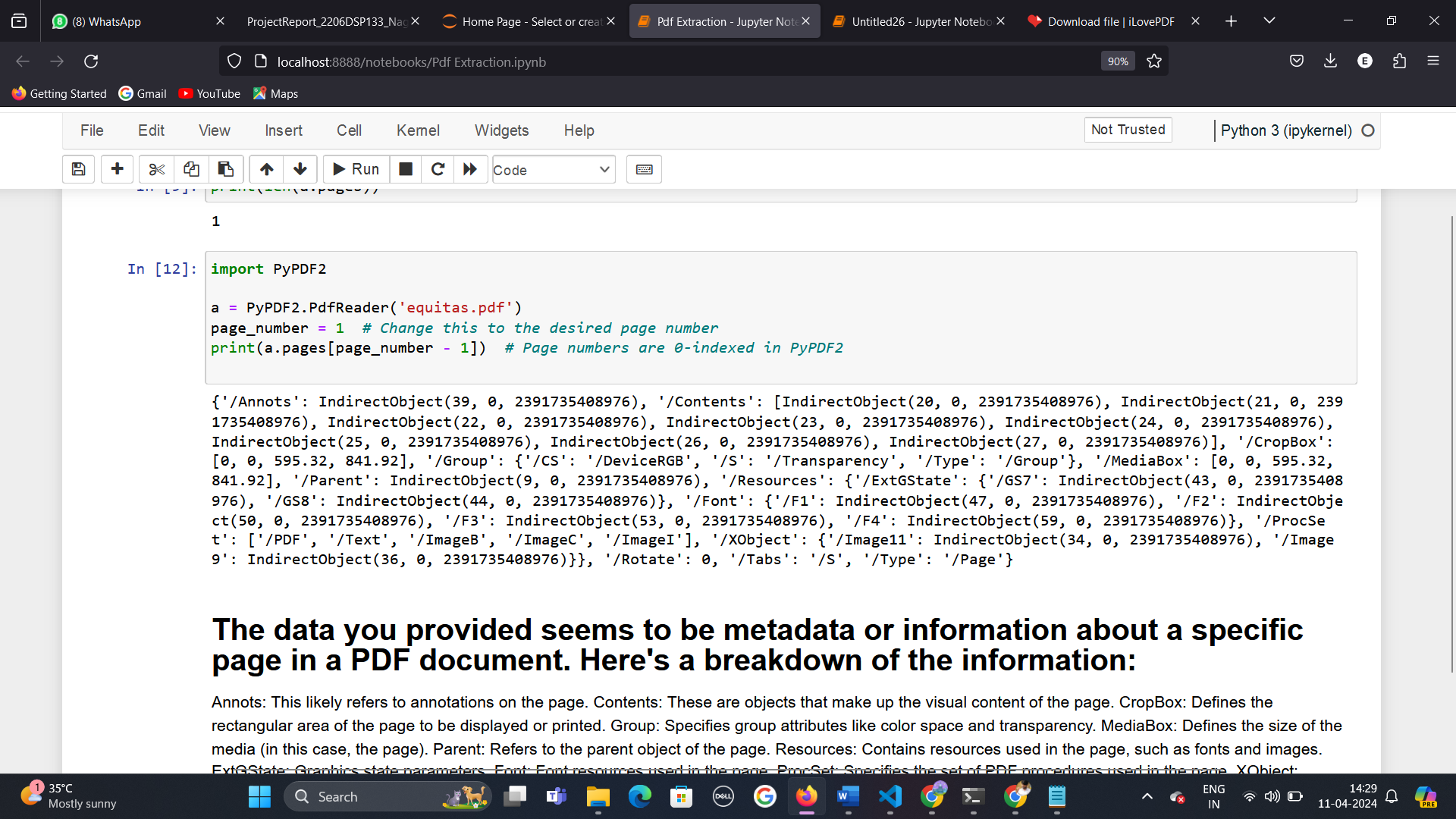
Producer: Microsoft® Word 2016

This metadata provides insights into who created the PDF, when it was created and modified, and which software was used to create it. It can be useful for tracking document history and understanding its origin and editing processes.

****

Annots: This likely refers to annotations on the page. Contents: These are objects that make up the visual content of the page. CropBox: Defines the rectangular area of the page to be displayed or printed. Group: Specifies group attributes like color space and transparency. MediaBox: Defines the size of the media (in this case, the page). Parent: Refers to the parent object of the page. Resources: Contains resources used in the page, such as fonts and images. ExtGState: Graphics state parameters. Font: Font resources used in the page. ProcSet: Specifies the set of PDF procedures used in the page. XObject: External objects like images used in the page. Rotate: Specifies the rotation angle of the page. Tabs: Indicates the type of tabs used in the page. Type: Denotes the type of object, in this case, a PDF page. This information provides a detailed description of the structure and content of a specific page in the PDF document.

Number of pages in pdf:

****

**5. Literature Review**

5.1 Existing Methods for PDF Extraction

***Manual Extraction Techniques:***

Overview of manual methods used for extracting data from PDFs.

Limitations such as time-consuming, error-prone, and lack of scalability.

***Rule-based Approaches:***

Description of rule-based systems that use predefined patterns or rules to extract information.

***Natural Language Processing (NLP) Techniques:***

Application of NLP algorithms for PDF extraction, such as tokenization, parsing, and entity recognition.

Examples of NLP tools and libraries used in PDF text extraction.

***Commercial and Open-Source Tools:***

Review of popular commercial and open-source tools for PDF extraction, such as *Adobe Acrobat, PyPDF2*, and *Apache PDFBox.*

Comparison of features, usability, and limitations of these tools.

5.2 Machine Learning Techniques in Text Extraction

***Supervised Learning Approaches:***

Overview of supervised learning algorithms used for text extraction, such as Support Vector Machines (SVM), Random Forest, and Neural Networks.

Use cases and performance metrics for supervised learning in text extraction from PDFs.

***Unsupervised Learning Techniques:***

Description of unsupervised learning methods like clustering and topic modeling for text extraction.

***Hybrid Approaches:***

Discussion of hybrid models combining rule-based, NLP, and machine learning techniques for improved accuracy and efficiency.

Case studies or examples of successful implementations of hybrid models in PDF extraction tasks.

***Deep Learning for Text Extraction:***

Exploration of deep learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures for text extraction from PDFs.

Advances, challenges, and future prospects of deep learning in PDF data extraction. Formation about the Features & their data types.

**Methodology followed to achieve the solution for the problem statement:**

* To achieve the goal of this project we followed certain steps:
  + Step-1: Extract text from PDF
  + Step-2: Finding patterns to identify the required text to retrieve from pdf file. By using PyPDF2, re, pdfplumber libraries.
  + Step-3: Created flask app to create user friendly environment. To drop pdf file and required text will print in Director Details page.
* Data Collection (Sample PDFs)

The pdf samples are retrieved with the help of pdfplumber,PyPDF2 libraries.

* Preprocessing Steps

**Text Extraction:**

The pdf information should be extracted from the local drive and the information also extracted in text format.

import PyPDF2

def extract\_text\_from\_pdf(pdf\_path):

# Open the PDF file in binary mode

with open(pdf\_path, 'rb') as file:

# Create a PdfReader object

pdf\_reader = PyPDF2.PdfReader(file)

# Initialize an empty string to store the extracted text

text = ''

# Iterate through each page in the PDF

for page\_num in range(len(pdf\_reader.pages)):

# Get the specific page using getPage() method

page = pdf\_reader.getPage(page\_num)

# Extract text from the page using extractText() method

page\_text = page.extractText()

# Append the extracted text to the overall text string

text += page\_text

return text

# Example usage:

pdf\_path = 'path\_to\_your\_pdf\_file.pdf' # Replace 'path\_to\_your\_pdf\_file.pdf' with your actual PDF file path

extracted\_text = extract\_text\_from\_pdf(pdf\_path)

print(extracted\_text)

**Data Cleaning**

Data cleaning plays an important key role to locate and retrieve the required data. To achieve that we used ***‘Name Entity Recognition on PDF using NLP and spacy’***.

Named Entity Recognition (NER) is a natural language processing (NLP) technique used to identify and classify named entities in text into predefined categories such as person names, organizations, locations, dates, and more. When applied to PDF resumes, NER helps in extracting key information like the candidate's name, contact details, work experience, education history, skills, and other relevant information.

Spacy is a popular NLP library in Python that provides robust support for NER. Here's an explanation of how NER can be used on PDF resumes using Spacy:

**Data Preparation:**

* The first step is to extract text from the PDF resume, as shown in the previous code snippet.
* Once you have the extracted text, you can pass it through a Spacy NLP pipeline for NER.

**Spacy NLP Pipeline:**

* Spacy provides pre-trained models that can recognize named entities out of the box.
* You can load a pre-trained Spacy model that includes an NER component. For example, en\_core\_web\_sm is a small English model that includes NER.

**Entity Recognition:**

* After loading the Spacy model, you can process the extracted text through the NLP pipeline.
* The NER component in Spacy will automatically identify and classify named entities in the text.
* Entities recognized by Spacy include PERSON (person names), ORG (organizations), DATE (dates), etc.

**Extracting Relevant Entities:**

* Once the NER process is complete, you can filter the extracted entities based on your requirements.
* For example, to extract the candidate's name, you would look for entities classified as PERSON.
* Similarly, you can extract other relevant information such as organizations, dates, locations, etc., depending on your needs.

**Handling Custom Entities:**

* Spacy also allows for training custom NER models if you have specific entity types not covered by the pre-trained models.
* You can annotate and train a custom NER model using Spacy's training capabilities, which involves providing labeled data for training.

**Post-Processing and Validation:**

* + After extracting named entities, it's essential to perform post-processing and validation to ensure accuracy.
  + This may involve checking extracted entities against a list of known entities (e.g., validating company names against a database) or applying additional logic to validate extracted information.
* **Observations:**

Total Pages : 1

{

"DATE": [

"April 0 5, 2024",

"400051",

"400001",

"2015",

"1949",

"April 04",

"2024",

"2024 @ 4.33",

"08537123",

"3 ) years",

"April 25",

"April 24, 2024"

],

"ORG": [

"National Stock Exchange of India Limited \nExchange",

"BKC",

"BSE Limited",

"Reserve Bank",

"RBI",

"the Bank \n \nPursuant",

"the Banking \nRegulation Act",

"the Reserve Bank of India vide",

"Bank",

"the Board of the Bank for the",

"Reserve \nBank of India",

"Chair",

"the Bank for a period of three",

"the Board of \nDirectors",

"For Equitas Small Finance Bank Limited"

],

"PERSON": [

"Bandra Kurla Complex",

"Bandra",

"Anil Kumar",

"Arun Ramanathan"

],

"GPE": [

"Mumbai"

],

"NORP": [

"Sharma"

],

"CARDINAL": [

"024-25"

]

}

The extracted named entities from the PDF are as follows:

DATE:

"April 05, 2024", "400051", "400001", "2015", "1949", "April 04, 2024", "08537123", "3) years", "April 25", "April 24, 2024".

ORG:

"National Stock Exchange of India Limited Exchange", "BKC", "BSE Limited", "Reserve Bank", "RBI", "the Bank Pursuant", "the Banking Regulation Act", "the Reserve Bank of India vide", "Bank", "the Board of the Bank for the", "Reserve Bank of India", "For Equitas Small Finance Bank Limited".

PERSON:

"Bandra Kurla Complex", "Bandra", "Anil Kumar", "Arun Ramanathan".

GPE:

"Mumbai".

NORP:

"Sharma".

CARDINAL:

"024-25".

*These entities represent different categories such as dates, organizations, people, locations, nationalities, and numerical values extracted from the text of the PDF document using spaCy's named entity recognition (NER) capabilities.*

1. **Final Implementation**

For the final implementation, a web application is developed using Flask. The libraries used for it’s implementation is re,spacy,PyPDF2,pdfplumber,NLP libraries and ML techniques. Patterns for retrieving ‘Director Name’,’DIN’ and ‘status’ is:

director\_pattern = r"(Mr\.|Mrs\.|Ms\.)\s+([A-Za-z]+)\s+([A-Za-z]+)"

din\_pattern = r"DIN:\s\*(\d+)"

independence\_pattern = r"(?i)Independent"

**Director Pattern:**

Pattern: (Mr\.|Mrs\.|Ms\.)\s+([A-Za-z]+)\s+([A-Za-z]+)

Explanation:

(Mr\.|Mrs\.|Ms\.): This part matches the salutation of the director, which can be "Mr.", "Mrs.", or "Ms.", using the pipe | as an "or" operator.

\s+: Matches one or more whitespace characters (spaces).

([A-Za-z]+): This part captures the first name of the director. [A-Za-z]+ matches one or more uppercase or lowercase letters.

\s+: Matches one or more whitespace characters again.

([A-Za-z]+): This part captures the last name of the director, similar to the first name capture group.

**DIN Pattern:**

Pattern: DIN:\s\*(\d+)

Explanation:

DIN:: Matches the literal string "DIN:".

\s\*: Matches zero or more whitespace characters.

(\d+): This part captures the DIN (Director Identification Number), which consists of one or more digits (\d+).

**Pattern: (?i)Independent**

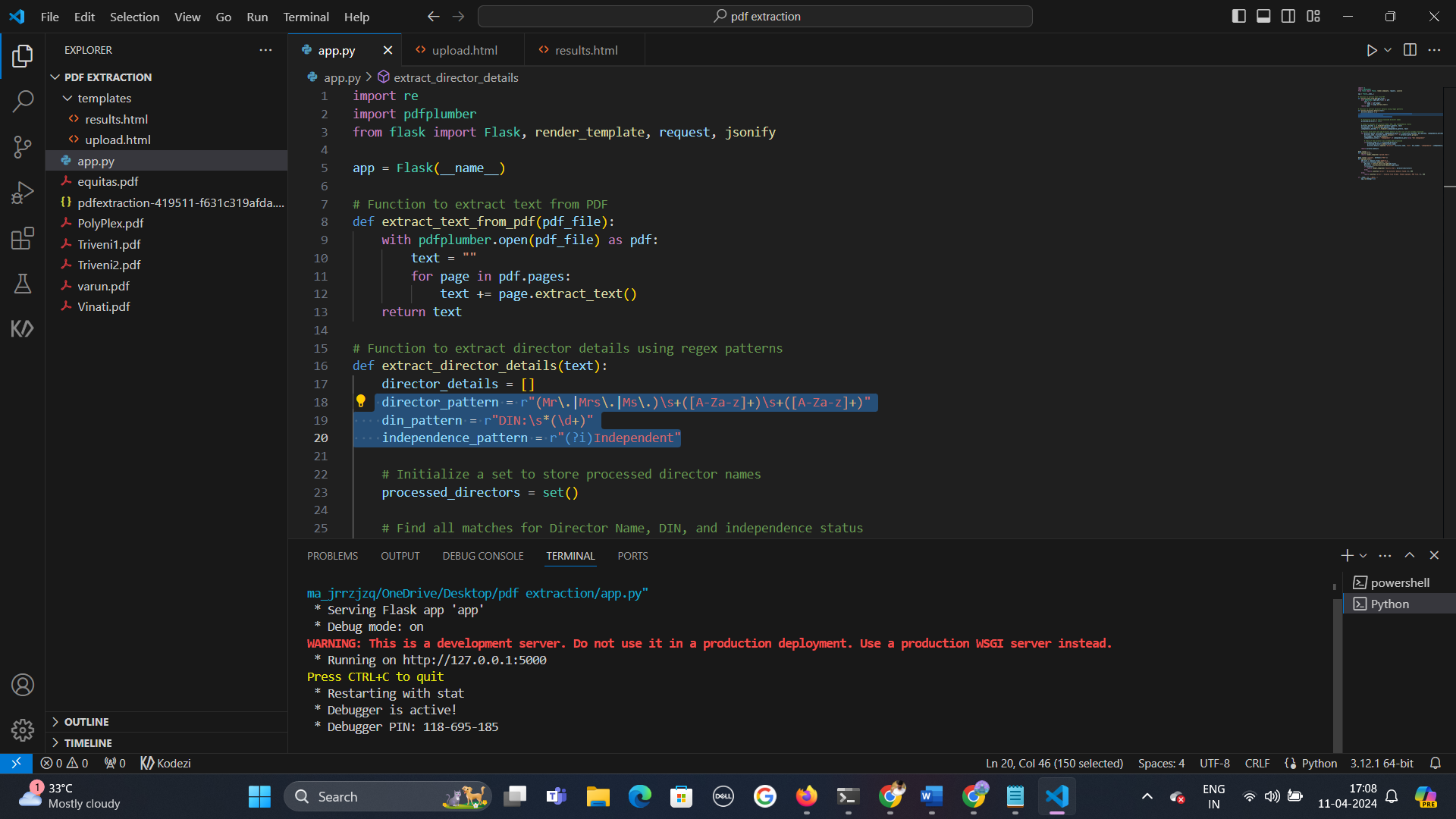
Explanation:

(?i): This is a flag for case-insensitive matching, allowing the pattern to match "Independent" in any case (uppercase or lowercase).

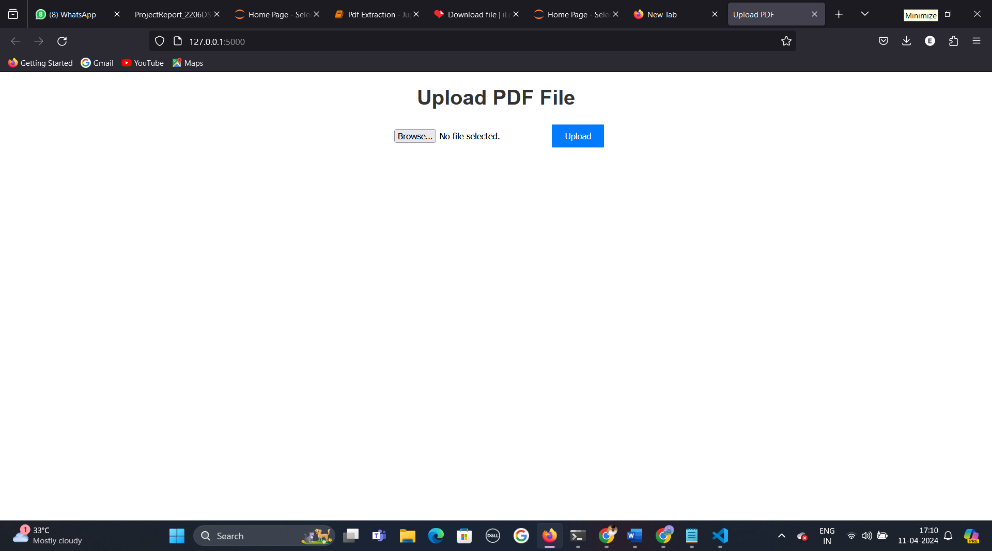
Independent: Matches the word "Independent".

These patterns are used for extracting specific information from text based on predefined formats or keywords. In your case, you can use these patterns to search for and extract director names, DIN numbers, and independence status from text documents.

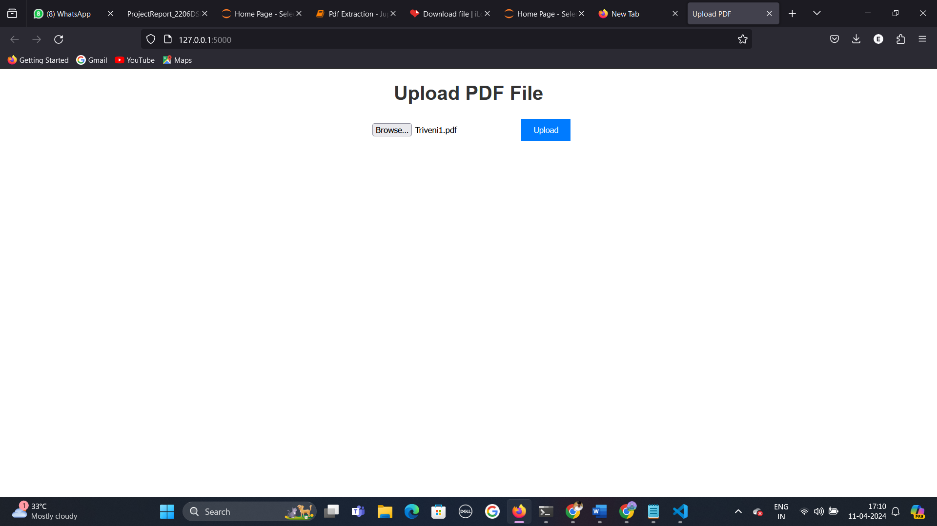
1. Launch the web app by executing the Python code, which will deploy the application on localhost.



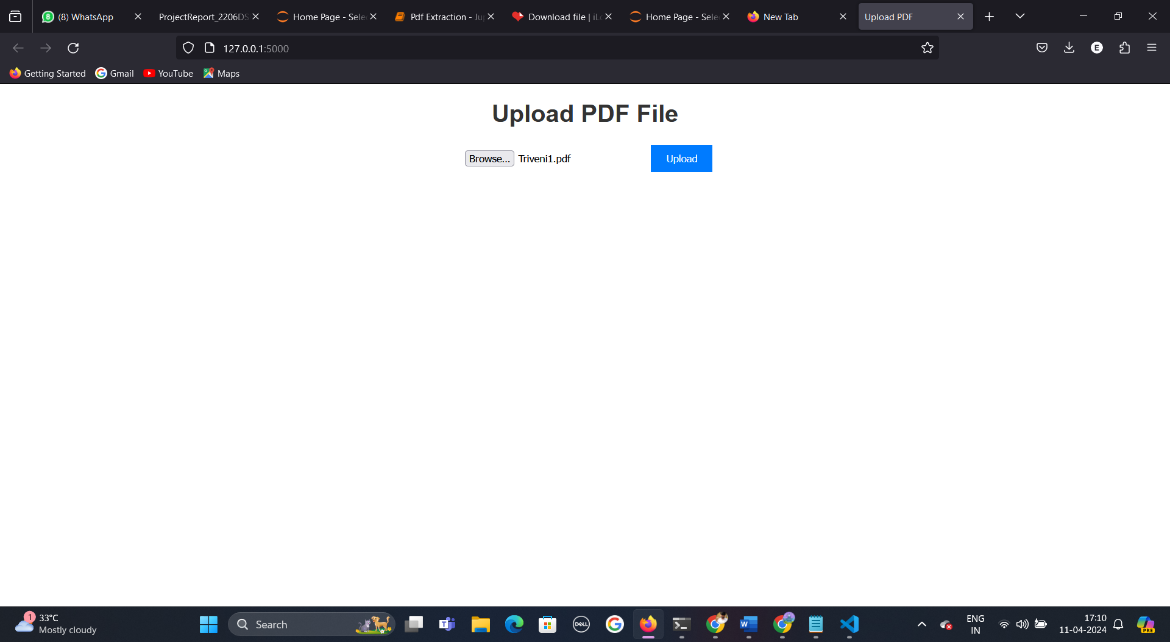
1. Open the deployed web app in a browser. If the deployment is successful the following webpage is rendered.



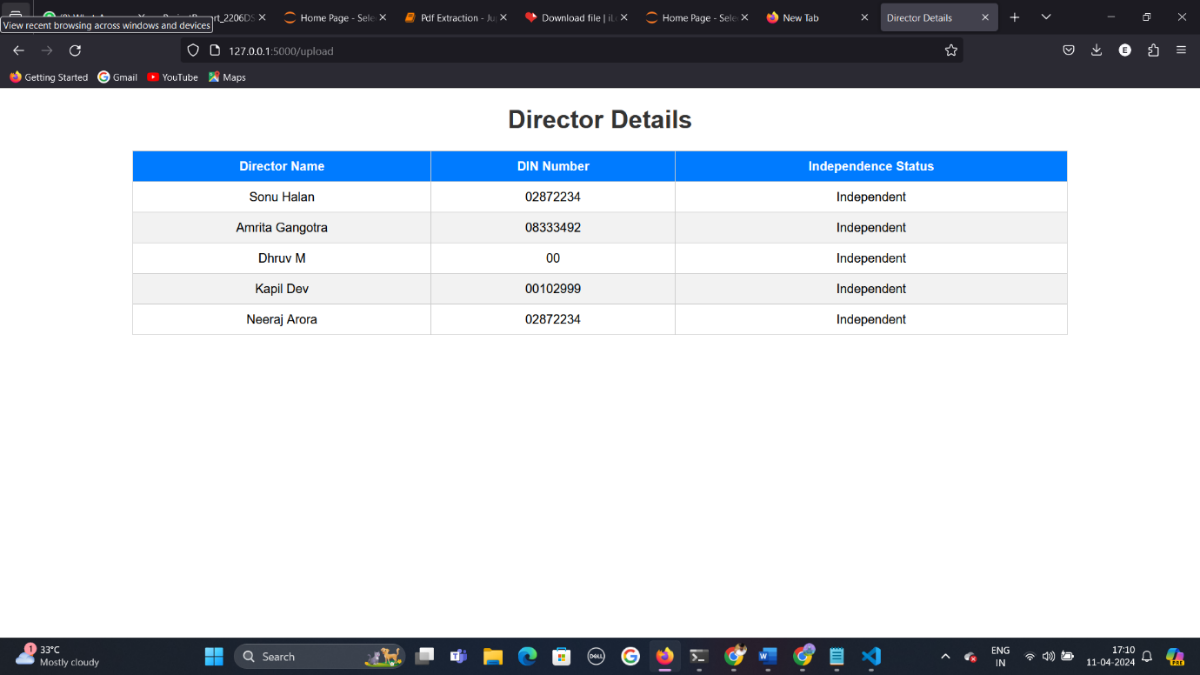
1. Upload the pdf in pdf upload section.



1. Click on the upload button.



5.Results will appear in ‘Director Details Page’.



**Conclusion**

In our study, the project demonstrated the efficiency of machine learning in PDF data extraction, highlighting its potential to streamline processes, enhance data quality, and improve document accessibility for various stakeholders. Because of using regular expression pattern matching we also find out that it will give the accuracy of 80% to retrieve the information of Director name, DIN, whether dependent or executive.