The Generative-AI powered Resume Analyzer

Introduction

The **Advanced Resume Analyzing System** is designed to automate the extraction of key information from resumes, which can be in the form of PDF or DOCX files. This system processes resumes stored in local directories or Google Drive links and outputs a structured CSV file with the parsed data. The system leverages natural language processing (NLP) techniques, regular expressions, and predefined keywords matching to extract important fields such as name, contact information, education, experience, skills, and other relevant data points.

Key Features

1. **Supports Multiple File Formats**: It can parse PDF and DOCX files, which are common formats for resumes.
2. **Google Drive Integration**: The system can download resumes directly from a Google Drive folder, making it easier to process resumes stored in cloud storage.
3. **Named Entity Recognition (NER)**: Utilizes spaCy's NLP model to identify and extract personal information like names from the resume text.
4. **Regex for Contact Information**: Extracts emails and phone numbers using regular expressions to match common patterns.
5. **Keyword Matching for AI/ML and Gen AI**: Allows the user to provide keywords related to AI, ML, and Gen AI, matching them within the resume text to assign a score.
6. **Education and Experience Extraction**: Identifies and extracts educational qualifications, years of experience, and the disciplines of study.
7. **Sporting and Extracurricular Activities**: Extracts details about certifications, internships, conferences, and other extracurricular activities.
8. **Unique File Naming**: Ensures that the output files do not overwrite existing ones by appending a counter to the base name.
9. **Detailed Output**: The processed resumes are saved in a CSV file with detailed columns including all the extracted fields for easy analysis.

Methodology

Step 1: File Processing and Text Extraction

- **PDF Handling**: The system uses pdfplumber to extract text from PDF files. This library allows the extraction of text from each page of the PDF, which is then consolidated into a single string.
- **DOCX Handling**: For DOCX files, python-docx is used to extract text from paragraphs within the document.
- **Google Drive Handling**: If the input is a Google Drive link, the system uses the Google Drive API to authenticate the user, access the folder, and download the files to a temporary local directory for processing.

Step 2: Data Extraction

- **Text Preprocessing**: The extracted text is processed for key information:
    - **Email and Phone**: Regular expressions are used to extract email addresses and phone numbers.
    - **Personal Name**: Named Entity Recognition (NER) from spaCy is used to extract the name of the applicant.
    - **Education**: A predefined list of degrees (e.g., B.Sc, M.Tech, PhD) is matched to extract educational qualifications.
    - **Experience**: Years of experience are extracted using a regex pattern that looks for phrases like "5 years" or "3 yrs".
    - **Discipline**: A predefined list of engineering and science disciplines (e.g., Computer Science, Mechanical Engineering) is used to identify relevant fields of study.
    - **Skills**: A list of common programming and soft skills (e.g., Python, Java, Teamwork) is used to detect the relevant skills mentioned in the resume.
    - **CGPA/Percentile**: CGPA values (e.g., 8.5) or percentiles (e.g., 85%) are detected using regex patterns.
    - **Extracurricular Information**: Keywords such as "certification," "internship," and "conference" are used to detect extracurricular and volunteering activities.

Step 3: Keyword Matching

- **Gen AI and AI/ML Keywords**: The system allows the user to provide a list of Gen AI and AI/ML related keywords. These keywords are matched against the resume text to calculate a score based on the number of matching terms.

Step 4: Data Storage

- All extracted information is stored in a structured format in a list of lists. Each list represents a single resume's data. This data is then converted into a DataFrame using pandas and saved as a CSV file.

Step 5: Output

- The final output is a CSV file that contains columns for:
    - Serial number
    - Name of applicant
    - Years of experience
    - Email and phone number
    - Education details
    - Discipline of study
    - Passing year
    - Key skills
    - CGPA/percentile
    - Certification/Sporting/extracurricular activities/internship
    - Gen AI keyword score
    - AI/ML keyword score
    - Matching Gen AI keywords
    - Matching AI/ML keywords

The CSV file is saved with a unique name to prevent overwriting of existing files.

Workflow Diagram

1. **Input Source**: Google Drive or Local Directory
2. **File Extraction**: Extract text from PDF or DOCX using libraries like pdfplumber or python-docx.
3. **Text Processing**: Extract key details (name, email, phone, education, etc.) using NLP and regex.
4. **Keyword Matching**: Match provided AI/ML and Gen AI keywords to calculate scores.
5. **Store Data**: Convert the extracted information into a structured CSV file.
6. **Output**: Save the CSV file in the specified output directory.

Unique Features

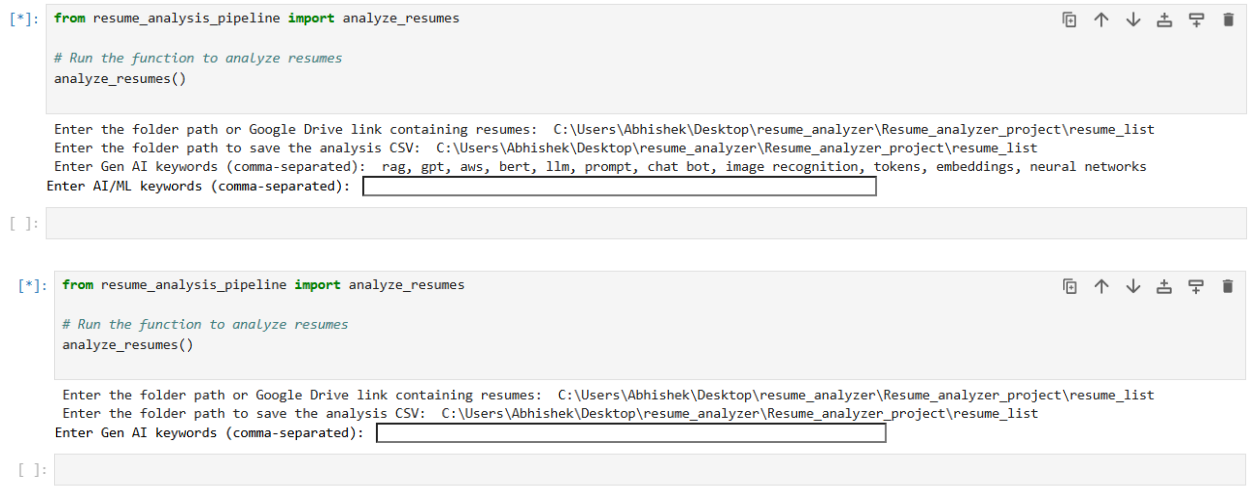
1. **Customizable Keyword Matching**: Users can input their own list of keywords related to AI/ML or Gen AI, making the system flexible for different job profiles.
2. **Google Drive Support**: Ability to download and process resumes directly from a Google Drive folder.
3. **NLP-based Name Extraction**: Uses spaCy's Named Entity Recognition (NER) to accurately extract the applicant's name.
4. **Detailed CSV Output**: The system outputs a comprehensive CSV file that includes extracted details as well as keyword matching scores for AI/ML and Gen AI.
5. **Text Extraction from Multiple Formats**: Supports both PDF and DOCX formats, which are widely used for resumes.
6. **Flexible Input and Output Paths**: Can process resumes from either local folders or cloud storage (Google Drive), and outputs to a specified directory.
7. **Skipping the unnecessary files**: It will skip all other format resumes except pdf and doc & **continue the process by skipping those files** similarly in output also if any file exists with same name it also **modified name by adding numerical to the end** without showing any error .
8. **Easy handling & customizable as Required**: it is too simple to execute and does not need any prior coding skills and similar to less code **Gen Ai resume analyzer without using any Api keys,** so it is also cost effective with high accuracy. It is also **customizable as per skill set so it can be used for multiple job profile screening** with different skillsets without changing any internal code.

Working Procedure

Steps

- 1st create a environment with the latest python version or more than 3.11.
- Then install the necessary libraries with the compatible version as given in Software_package. Some libraries should install using pip command and other high end framework and libraries should be installed using “Anaconda prompt”.All the command with installation guide were given in Software_package.ipnyb file
- Then create two folders, one for storing the resume and another for getting the output in Excel sheet.
- After that run **resume_analyzer_main.ipnyb** file and give the necessary user input and give the folder path or Google drive link **(while giving the folder path for both input and output don’t give the path inside “” else you will get error keep it avoiding if it exist remove that)**
- After executing run **resume_analyzer_main.ipnyb** then execute pipeline_code.ipnyb , which will create the pipeline save the code .py format with the name **resume_analysis_pipeline.py**
- Then run the code using the running code given in **running_code.ipnyb .** after that give the necessary key word as desired then folder path without “” and get the result in given output folder in excel sheet.
- While seeing the result in in excel sheet apart from the given column in question **additional 2 columns given which shows no of Gen Ai and no of Ai Ml matching keyword found** in resume.

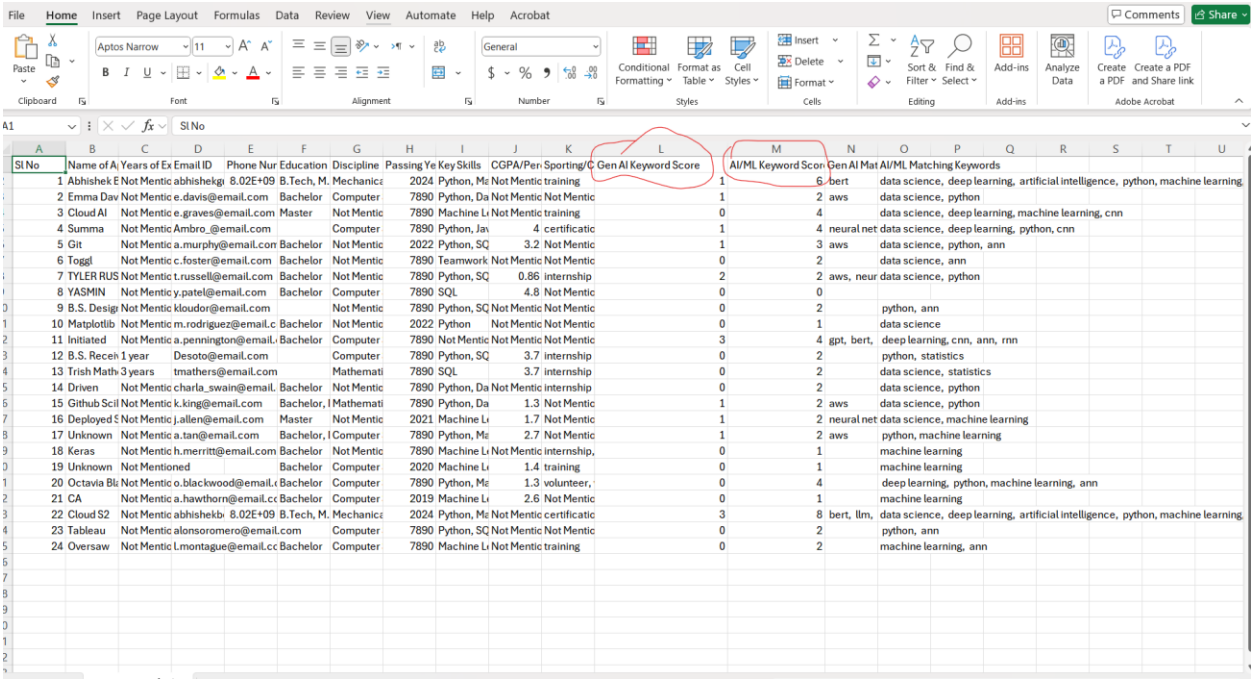
Diagrammatic Demo



```
[*]: from resume_analysis_pipeline import analyze_resumes

     # Run the function to analyze resumes
     analyze_resumes()

Enter the folder path or Google Drive link containing resumes:  C:\Users\Abhishek\Desktop\resume_analyzer\Resume_analyzer_project\resume_list
Enter the folder path to save the analysis CSV: ↑↓ for history. Search history with c-↑/c-↓
```

```
[ ]:
```

```
[*]: from resume_analysis_pipeline import analyze_resumes

     # Run the function to analyze resumes
     analyze_resumes()

Enter the folder path or Google Drive link containing resumes: ↑↓ for history. Search history with c-↑/c-↓
```



Sl No	Name of A	Years of Ex	Email ID	Phone Nur	Education	Discipline	Passing Ye	Key Skills	CGPA/Per	Sporting/C	Gen AI Keyword Score	AI/ML Keyword Scor	Gen AI Mat	AI/ML Matching Keywords
1	Abhishek E	Not Mentic	abhishekg	8.02E+09	B.Tech, M.	Mechanica	2024	Python, Ma	Not Mentic	training	1	6	bert	data science, deep learning, artificial intelligence, python, machine learning
2	Emma Dav	Not Mentic	e.davis@email.com		Bachelor	Computer	7890	Python, Da	Not Mentic	Not Mentic	1	2	aws	data science, python
3	Cloud AI	Not Mentic	e.graves@email.com		Master	Not Mentic	7890	Machine Le	Not Mentic	training	0	4		data science, deep learning, machine learning, cnn
4	Summa	Not Mentic	Ambro_@email.com			Computer	7890	Python, Jav	4	certificatic	1	4	neural net	data science, deep learning, python, cnn
5	Git	Not Mentic	a.murphy@email.com		Bachelor	Not Mentic	2022	Python, SQ	3.2	Not Mentic	1	3	aws	data science, python, ann
6	Toggl	Not Mentic	c.foster@email.com		Bachelor	Not Mentic	7890	Teamwork	Not Mentic	Not Mentic	0	2		data science, ann
7	TYLER RUS	Not Mentic	t.russell@email.com		Bachelor	Not Mentic	7890	Python, SQ	0.86	internship	2	2	aws, neur	data science, python
8	YASMIN	Not Mentic	y.patel@email.com		Bachelor	Computer	7890	SQL	4.8	Not Mentic	0	0		
9	B.S. Desig	Not Mentic	kloudor@email.com			Not Mentic	7890	Python, SQ	Not Mentic	Not Mentic	0	2		python, ann
10	Matplotlib	Not Mentic	m.rodriguez@email.c		Bachelor	Not Mentic	2022	Python	Not Mentic	Not Mentic	0	1		data science
11	Initiated	Not Mentic	a.pennington@email.		Bachelor	Computer	7890	Not Mentic	Not Mentic	Not Mentic	3	4	gpt, bert,	deep learning, cnn, ann, rnn
12	B.S. Recei	1 year	Desoto@email.com			Computer	7890	Python, SQ	3.7	internship	0	2		python, statistics
13	Trish Math	3 years	tmathers@email.com			Mathemati	7890	SQL	3.7	internship	0	2		data science, statistics
14	Driven	Not Mentic	charla_swain@email.		Bachelor	Not Mentic	7890	Python, Da	Not Mentic	internship	0	2		data science, python
15	Github Scil	Not Mentic	k.king@email.com		Bachelor, I	Mathemati	7890	Python, Da	1.3	Not Mentic	1	2	aws	data science, python
16	Deployed S	Not Mentic	j.allen@email.com		Master	Not Mentic	2021	Machine Le	1.7	Not Mentic	1	2	neural net	data science, machine learning
17	Unknown	Not Mentic	a.tan@email.com		Bachelor, I	Computer	7890	Python, Ma	2.7	Not Mentic	1	2	aws	python, machine learning
18	Keras	Not Mentic	h.merritt@email.com		Bachelor	Not Mentic	7890	Machine Le	Not Mentic	internship,	0	1		machine learning
19	Unknown	Not Mentioned			Bachelor	Computer	2020	Machine Le	1.4	training	0	1		machine learning
20	Octavia Bla	Not Mentic	o.blackwood@email.(		Bachelor	Computer	7890	Python, Ma	1.3	volunteer,	0	4		deep learning, python, machine learning, ann
21	CA	Not Mentic	a.hawthorn@email.cc		Bachelor	Computer	2019	Machine Le	2.6	Not Mentic	0	1		machine learning
22	Cloud S2	Not Mentic	abhishekb	8.02E+09	B.Tech, M.	Mechanica	2024	Python, Ma	Not Mentic	certificatic	3	8	bert, llm,	data science, deep learning, artificial intelligence, python, machine learning
23	Tableau	Not Mentic	alonsoromero@email.com			Computer	7890	Python, SQ	Not Mentic	Not Mentic	0	2		python, ann
24	Oversaw	Not Mentic	l.montague@email.cc		Bachelor	Computer	7890	Machine Le	Not Mentic	training	0	2		machine learning, ann

****