PDF Reader using CHATGPT API —

Upon discovering that ChatPDF employs chatgpt as its underlying technology, I opted to directly utilize the chatgpt API through the langchain framework.

This is how i have used it to complete the given task –

#set the key environment variable

```
import os
os.environ["OPENAI_API_KEY"] = "your api key"
```

#set up dependencies , and loaded pdf documents.

```
from langchain.llms import OpenAI
llm = OpenAI()
from langchain.document_loaders import PyPDFLoader

loader =PyPDFLoader('/content/29062021_Adv_Museum_Assistant_and_Assistant_Conservator.pdf')
pages = loader.load_and_split()
```

set prompt for the llm

```
query="Extract information and answer in this format only ---1) Organization name 2) Post Name 3) No. Of vacancies 4) Location 5) Salary 6) Last date of filling form 7) Qualification required for filling the form If there are multiple jobs then put it in same index and give answer,don't give headings just give answer for the point, i wan't it save it in excel file"
```

#setting prompt chain

```
from langchain.chains.question_answering import load_qa_chain
chain = load_qa_chain(llm, chain_type="stuff")
answer=chain.run(input_documents=pages, question=query)
```

#answer

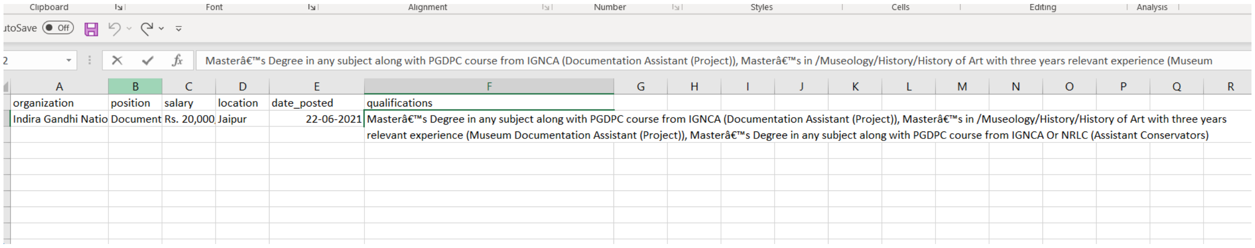
```
print(answer)
```

1) Indira Gandhi National Centre for the Arts
2) Documentation Assistant (Project), Museum Documentation Assistant (Project), Assistant Conservators
3) 04 (Documentation Assistant (Project)), 02 (Museum Documentation Assistant (Project)), 02 (Assistant Conservators)
4) Jaipur
5) Rs. 20,000/ - (Documentation Assistant (Project)), Rs. 30,000/ - (Museum Documentation Assistant (Project)), Rs. 20,000/ - (Assistant Conservators)
6) 22/06/2021
7) Master's Degree in any subject along with PGDPC course from IGNCA (Documentation Assistant (Project)), Master's in /Museology/History/History of Art with three years relevant experience (Museum Documentation Assistant (Project)), Master's Degree in any subject along with PGDPC course from IGNCA Or NRLC (Assistant Conservators)

#stored it in csv file

```
import csv


# Split the text by newline characters and remove any empty lines
lines = [line.strip() for line in answer.split('\n') if line.strip()]


# Extract the data from the lines
organization = lines[0][3:]  # Remove the prefix "1) "
positions = lines[1][3:]    # Remove the prefix "2) "
num_positions = lines[2][3:]
location = lines[3][3:]
salaries = lines[4][3:]
date_posted = lines[5][3:]
qualifications = lines[6][3:]


# Split the salaries into individual positions and amounts
```

```python
positions_salaries = [x.strip() for x in salaries.split(';')]


# Create a list of dictionaries with the extracted data
data = []


data.append({
    'organization': organization,
    'position': positions.strip(),
    'salary': salaries.strip(),
    'location': location,
    'date_posted': date_posted,
    'qualifications': qualifications,
})


# Write the data to a CSV file
with open('data.csv', 'w', newline='') as csvfile:
    fieldnames = ['organization', 'position', 'salary', 'location', 'date_posted', 'qualifications']
    writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
    writer.writeheader()
    for row in data:
        writer.writerow(row)
```



Masterâ€™s Degree in any subject along with PGDPC course from IGNCA (Documentation Assistant (Project)), Masterâ€™s in /Museology/History/History of Art with three years relevant experience (Museum

organization	position	salary	location	date_posted	qualifications
Indira Gandhi Natio	Document	Rs. 20,000	Jaipur	22-06-2021	Masterâ€™s Degree in any subject along with PGDPC course from IGNCA (Documentation Assistant (Project)), Masterâ€™s in /Museology/History/History of Art with three years relevant experience (Museum Documentation Assistant (Project)), Masterâ€™s Degree in any subject along with PGDPC course from IGNCA Or NRLC (Assistant Conservators)

By utilizing this method, we have the capability to develop an automated system for extracting vacancy information from pdfs into a structured format.