Web Scraping

DR. VEENA R S
ASSOCIATE PROFESSOR
DEPARTMENT OF ISE
DSATM, BENGALURU

10 a) Write a python program to combine select pages from many PDFs

Merging multiple PDFs into a single document is one activity which most of us have to do. The ones which allow you to merge PDFs for free often have some limits. Either based on number of files or the time between every merging operation

Prerequisites

- Must install latest version of Python
- Must install the PyPDF2 tool kit
- Saved the PDF files that you want to merge in Python's working directory. Of course, you can change the directory using Python code. For simplicity of code, place the PDF files on the working directory for these two methods that I am going to present here.

Steps

1. Import the PyPDF2 tool kit which has the tools that we need for playing with PDFs
2. Open each and every file by entering the file name
3. Read each and every file which was opened in Step 2 using PdfFileReader
4. Create a blank PDF file using PdfFileWriter where you can store the merged output
5. Loop through every page in every file which was read in Step 3 using for loop and copy all the information
6. Give a name for the output file and then paste all the copied information in Step 5
7. Close all the files

Program:

```
import PyPDF2
 import os
# Define the PDF files to combine and the page ranges to select pdf_files = ["file1.pdf", "file2.pdf", "file3.pdf"]
page_ranges = {"file1.pdf": [(1, 3), (5, 6)], "file2.pdf": [(2, 4), (7, 9)], "file3.pdf": [(1, 2), (4, 4)]}
# Create a new PDF writer object
pdf_writer = PyPDF2.PdfFileWriter()
#Loop through each pdf file
for pdf_file in pdf_files:
        # Open the PDF file
        pdf_file_obj = open(pdf_file, "rb")
        pdf_reader = PyPDF2.PdfFileReader(pdf_file_obj)
```

```
        # Loop through each page range and add the pages to the writer object
        for page_range in page_ranges[pdf_file]:
            for page_num in range(page_range[0]-1, page_range[1]):
                page_obj = pdf_reader.getPage(page_num)
                pdf_writer.addPage(page_obj)
            # Close the PDF file
        pdf_file_obj.close()
# Save the combined PDF to a new file
output_file = open("combined.pdf", "wb")
pdf_writer.write(output_file)
output_file.close()
print("Done.")
```

#10 a) Alternate Program

```
from PyPDF2 import PdfWriter, PdfReader

num = int(input("Enter page number you want combine from multiple documents "))

pdf1 = open(input('Enter the First Filename'), 'rb')
pdf2 = open(input('Enter the Second Filename'), 'rb')

pdf_writer = PdfWriter()

pdf1_reader = PdfReader(pdf1)
page = pdf1_reader.pages[num - 1]
pdf_writer.add_page(page)

pdf2_reader = PdfReader(pdf2)
page = pdf2_reader.pages[num - 1]
pdf_writer.add_page(page)

with open('d://DSATM//Finaloutput.pdf', 'wb') as output:
    pdf_writer.write(output)
```

10 b) Write a python program to fetch current weather data from the JSON file.

The OpenWeatherMap is a service that provides weather data, including current weather data, forecasts, and historical data to the developers of web services and mobile applications. It provides an API with JSON, XML, and HTML endpoints and a limited free usage tier. Making more than 60 calls per minute requires a paid subscription starting at USD 40 per month. Access to historical data requires a subscription starting at 150 USD per month. Users can request current weather information, extended forecasts, and graphical maps (showing cloud cover, wind speed, pressure, and precipitation).

Program

```
import json
#open JSON file and load the data into python object
with open("d://DSATM//weather.json","r") as f:
        data=json.load(f)
# Extract the current weather data from the object
current_temp = data['main']['temp']
humidity = data['main']['humidity']
weather_desc = data['weather'][0]['description']

# Display the weather data
print(f"Current temperature: {current_temp}°C")
print(f"Humidity: {humidity}%")
print(f"Weather description: {weather_desc}")
```

weather.ison File


```
{
  "coord": {
    "lon": -73.99,
    "lat": 40.73
  },
  "weather": [
    {
      "id": 800,
      "main": "Clear",
      "description": "clear sky",
      "icon": "01d"
    }
  ],
```

```
"base": "stations",
 "main": {
  "temp": 15.45,
  "feels_like": 12.74,
  "temp_min": 14.44,
  "temp_max": 16.11,
  "pressure": 1017,
  "humidity": 64
 },
 "visibility": 10000,
 "wind": {
  "speed": 4.63,
  "deg": 180
 },
 "clouds": {
  "all": 1
 },
```

```
"dt": 1617979985,
  "sys": {
    "type": 1,
    "id": 5141,
    "country": "US",
    "sunrise": 1617951158,
    "sunset": 1618000213
  },
  "timezone": -14400,
  "id": 5128581,
  "name": "New York",
  "cod": 200
}
```