unit test - 05

① web scraping is a term used to describe the use of a program or algorithm to extract & process large amounts of data from the web.

Downloading a web page

```
>>> import requests
>>> res = requests.get('http:// www.gutenberg.org/cache/
              epub/1112/pg1112.txt')
>>> type(res)
 <class 'requests.models.Response'>
>>> res.status_code == requests.codes.ok
  True
>>> len(res.text)
  178981
>>> print(res.text[:250])
```

Saving Downloaded files to the Hard Drive

```
>>> import requests
>>> res = requests.get('http://www.gutenberg.org/cache/
              Epub/1112/pg1112.txt')
>>> res.raise_for_status()
>>> playFile = open('RomeoAndJuliet.txt','wb')
>>> for chunk in res.iter_content(100000):
          playFile.write(chunk)
  100000
  78981
>>> playFile.close()
```

Scanned with CamScanner

(2) Creating PDF's

(1) open one or more Existing PDF's into pdffileReader objects.

(2) Create a new pdf.file writer object.

(3) Copy pages from the pdf fileReader objects into the

(3) PdfFileWriter object.

(4) Finally, use the pdf.fileWriter object to write the output PDF.

Copying Pages

```
import PyPDF2
pdf1File = open('meeting minutes.pdf', 'rb')
pdf2File = open('meeting minutes.pdf', 'rb')
pdf1Reader = PyPDF2.PdfFileReader(pdf1File)
pdf2Reader = PyPDF2.PdfFile~~Writer~~Reader(pdf2File)
pdfWriter = PyPDF2.PdfFileWriter()

for pageNum in range(pdf1Reader.numPages):
        pageObj = pdf1Reader.getPage(pageNum)
        pdfWriter.addPage(pageObj)

for pageNum in range(pdf2Reader.numPages):
        pageObj = pdf2Reader.getPage(pageNum)
        pdfWriter.addPage(pageObj)

pdfOutputFile = open('combinedminutes.pdf', 'wb')
pdfWriter.write(pdfOutputFile)
pdfOutputFile.close()
pdf1File.close()
pdf2File.close()
```

Rotating pages:-

```
import PyPDF2
minutesFile = open('meetingminutes.pdf', 'rb')
  pdfReader = PyPDF2. PdfFileReader(minutesFile)
    page = pdfReader.getpage(0)
    page. rotateClockwise(90)

  pdfWriter = pyPDF2. PdfFileWriter()
  pdfWriter.addPage(page)
  resultpdffile = open ('rotatedPage.pdf', 'wb')
  pdfWriter.write(resultPdfFile)
  resultPdfFile.close()
   minutesFile.close()
```

[illegible]

[illegible]

[illegible]

Scanned with CamScanner