Applications of Regular Expressions

1. **Regular Expression in UNIX:** Basically Regular Expression are used in search tools. In UNIX we use tool **egrep** to search a text file. This command searches for a text pattern in the file and list the file names containing the pattern.
2. **Regular Expressions in String processing:** Regular expressions are useful in a wide variety of text string processing, common applications include data validation, data scraping, data wrangling, simple parsing.
3. **Regular Expressions in String replacement:** Regular expressions are useful in replacing a particular matching pattern with our pattern.
4. **Regular Expressions in Lexical Analysis:** The Lexical Analyzer reads source code and generates a stream of tokens. Tokens can be described using regular expressions.
5. **Regular Expressions on Search Engines:** While regexps would be useful on Internet search engines, processing them across the entire database could consume excessive computer resources depending on the complexity and design of the regex.

Python Packages for accessing PDF documents

1. **PyPDF2:** A Pure-Python library built as a PDF toolkit. It is capable of:

Function	Description	Example

The PdfFileReader Class	Initializes a PdfFileReader object. This operation can take some time, as the PDF stream's cross-reference tables are read into memory.	# importing required modules
		import PyPDF2
		
		# creating a pdf file object
		pdfFileObj = open('example.pdf', 'rb')
		
		# creating a pdf reader object
		pdfReader=PyPDF2.PdfFileReader(pdfFileObj)

The PdfFileMerger Class	Initializes a PdfFileMerger object. PdfFileMerger merges multiple PDFs into a single PDF. It can concatenate, slice, insert, or any combination of the	import PyPDF2
		def PDFmerge(pdfs, output):
		    pdfMerger = PyPDF2.PdfFileMerger()
		
		    # appending pdfs one by one
		    for pdf in pdfs:
		        with open(pdf, 'rb') as f:
		            pdfMerger.append(f)

	above.	# writing combined pdf to output pdf file with open(output, 'wb') as f: pdfMerger.write(f)

The PdfFileWriter Class	This class supports writing PDF files out, given pages produced by another class (typically PdfFileReader).	# creating a pdf writer object for new pdf
		pdfWriter = PyPDF2.PdfFileWriter()
		
		#writing modified pages to new file
		pdfWriter.write(newFile)

2. **Textract:** This package provides a single interface for extracting content from any type of file, without any irrelevant markup

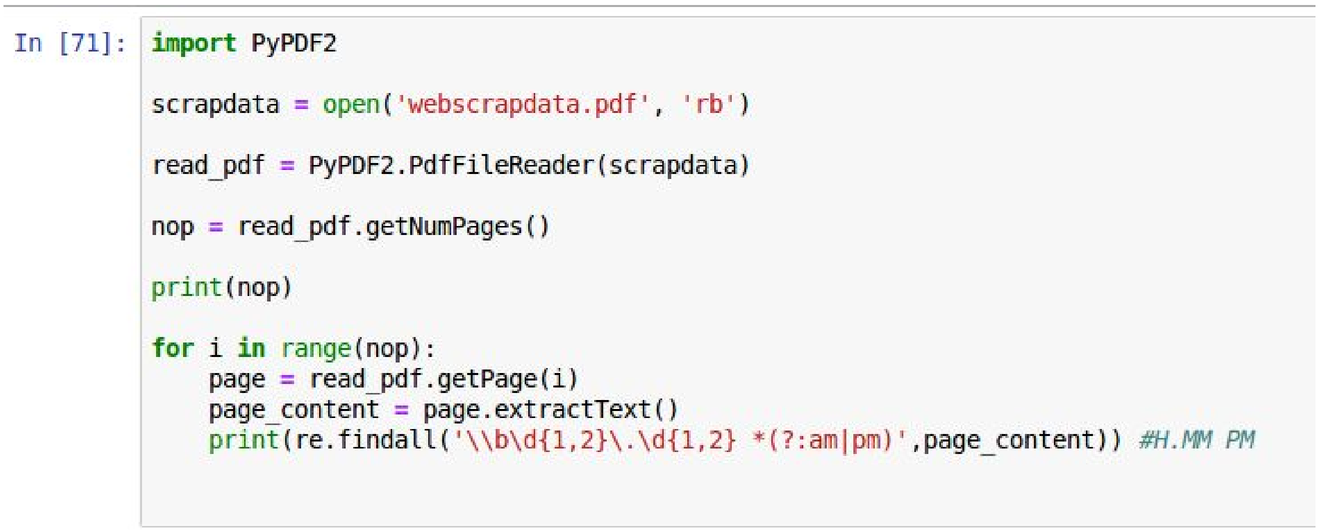
Function	Description	Example

textract.process()	to obtain text from a document. You can also pass keyword arguments to textract.process, for example, to use a particular method for parsing a pdf	import textract
		text = textract.process('path/to/a.pdf', method='pdfminer')

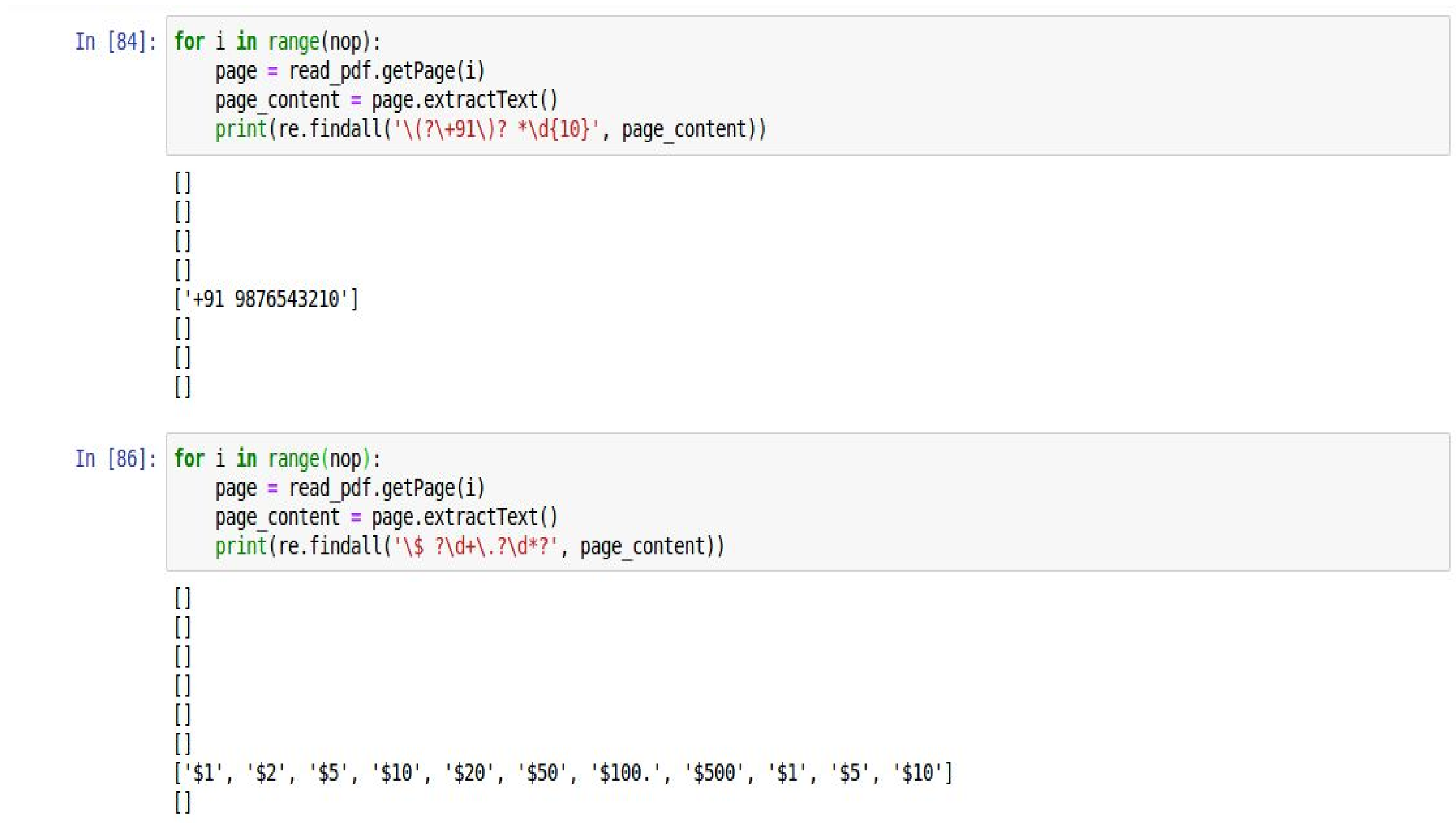
3. **Slate:** Slate is a Python package that simplifies the process of extracting text from PDF files. It depends on the PDFMiner package.

Function	Description	Example
Slate provides one class, PDF	PDF takes a file-like object and will extract all text from the document, presenting each page as a string of text	with open('example.pdf') as f: doc = slate.PDF(f)

Applying RE on PDF Document

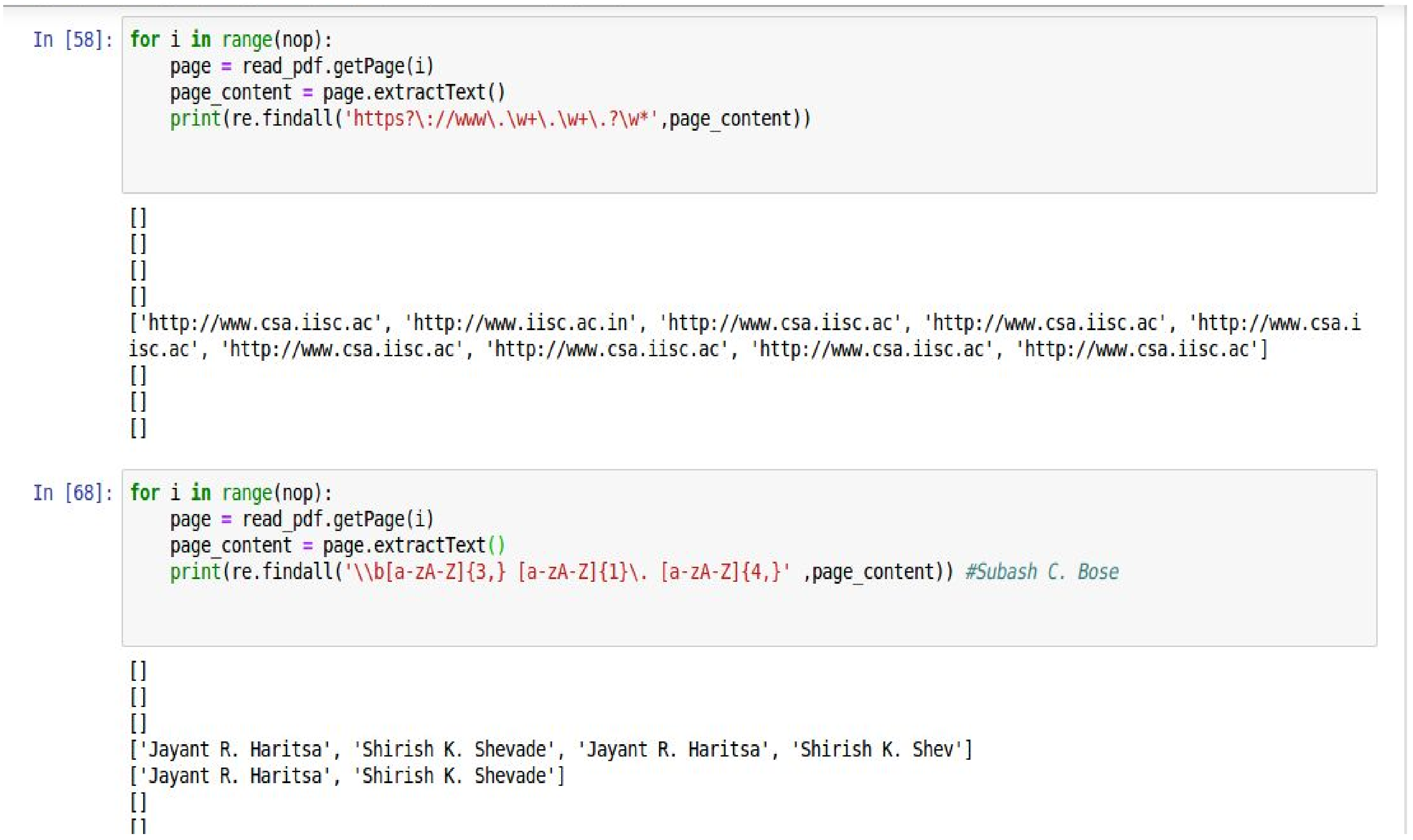


```
In [71]: import PyPDF2

         scrapdata = open('webscrapdata.pdf', 'rb')

         read_pdf = PyPDF2.PdfFileReader(scrapdata)

         nop = read_pdf.getNumPages()

         print(nop)

         for i in range(nop):
             page = read_pdf.getPage(i)
             page_content = page.extractText()
             print(re.findall('\\b\d{1,2}\.\d{1,2} *(?:am|pm)',page_content)) #H.MM PM
```



```
In [84]: for i in range(nop):
             page = read_pdf.getPage(i)
             page_content = page.extractText()
             print(re.findall('\(?\+91\)? *\d{10}', page_content))

         []
         []
         []
         []
         ['+91 9876543210']
         []
         []
         []
```

```
In [86]: for i in range(nop):
             page = read_pdf.getPage(i)
             page_content = page.extractText()
             print(re.findall('\$ ?\d+\.?\d*?', page_content))

         []
         []
         []
         []
         []
         []
         ['$1', '$2', '$5', '$10', '$20', '$50', '$100.', '$500', '$1', '$5', '$10']
         []
```



```
In [58]: for i in range(nop):
             page = read_pdf.getPage(i)
             page_content = page.extractText()
             print(re.findall('https?\://www\.\w+\.\w+\.?\w*',page_content))

         []
         []
         []
         []
         ['http://www.csa.iisc.ac', 'http://www.iisc.ac.in', 'http://www.csa.iisc.ac', 'http://www.csa.iisc.ac', 'http://www.csa.i
         isc.ac', 'http://www.csa.iisc.ac', 'http://www.csa.iisc.ac', 'http://www.csa.iisc.ac', 'http://www.csa.iisc.ac']
         []
         []
         []
```

```
In [68]: for i in range(nop):
             page = read_pdf.getPage(i)
             page_content = page.extractText()
             print(re.findall('\\b[a-zA-Z]{3,} [a-zA-Z]{1}\. [a-zA-Z]{4,}' ,page_content)) #Subash C. Bose

         []
         []
         []
         ['Jayant R. Haritsa', 'Shirish K. Shevade', 'Jayant R. Haritsa', 'Shirish K. Shev']
         ['Jayant R. Haritsa', 'Shirish K. Shevade']
         []
         []
```

```
In [95]: for i in range(nop):
             page = read_pdf.getPage(i)
             page_content = page.extractText()
             print(re.findall('(?:Jan|Feb|Mar|Apr|May|June|July|Aug|Sept|Oct|Nov|Dec) \d{1,2} *\, *\d{4}\\b', page_content))
```

```
[]
[]
[]
[]
[]
[]
['Apr 16, 2018']
['Oct 4, 2017', 'Jan 17, 2018', 'Nov 6, 2017']
```

```
In [96]: for i in range(nop):
             page = read_pdf.getPage(i)
             page_content = page.extractText()
             print(re.findall('\w+@\w+\.\w+', page_content))
```

```
[]
[]
[]
[]
[]
[]
['ziaur47779@Hotmail.com', 'Zrahman@YourBusiness.com']
['Zrahman@Yahoo.com', 'Zrahman@Gmail.com', 'JDoe@Gmail.com', 'JDoe4855@Gmail.com', 'JohnDoe@YourDomain.com', 'John@YourDo
main.com', 'JohnD@YourDomain.com', 'JDoe@YourDomain.com', 'JohnDoe@YourDomain.com', 'Info@YourDomain.com', 'JohnDoe@YourD
omain.com', 'JohnD@YourDomain.com', 'ter@YourDomain.com', 'Info@YourDomain.com', 'Sales@YourDomain.com', 'Press@YourDomai
n.com', 'Info@YourDomain.com', 'Info@moonyguitars.com', 'Jeremy@moonyguitars.com', 'Holway@outlook.com', 'last@yourdmain.
com']
```