😊 Working with PDF Files

1. In this lab, we are going to work with PDF files in Jupyter Notebook.
2. To work with a PDF file, we are going to install PyPDF2 and then import this library.

```
[2]: !pip install PyPDF2
```

```
Collecting PyPDF2
  Downloading pypdf2-3.0.1-py3-none-any.whl (232 kB)
     ---------------------------------- 232.6/232.6 kB 490.8 kB/s eta 0:00:00
Requirement already satisfied: typing_extensions>=3.10.0.0 in c:\users\jmpor\anaconda3\lib\site-packages (from PyPDF2) (4.3.0)
Installing collected packages: PyPDF2
Successfully installed PyPDF2-3.0.1
```

```
[3]: # note the capitalization
     import PyPDF2
```

3. The code opens a PDF file named Working_Business_Proposal.pdf in binary read mode ('rb'). It then initializes a PyPDF2.PdfReader object to read the PDF and determines the number of pages using len(pdf_reader.pages). After that, it selects the first page (page_one) from the pages list, which can later be used for extracting text or other operations.

```
[4]: # Notice we read it as a binary with 'rb'
     f = open('Working_Business_Proposal.pdf','rb')
```

```
[6]: pdf_reader = PyPDF2.PdfReader(f)
```

```
[8]: len(pdf_reader.pages)
```

```
[8]: 5
```

```
[10]: page_number = 0
      page_one = pdf_reader.pages[0]
```

4. The code extracts text from the first page of the PDF using extract_text() and stores it in page_one_text. Finally, it closes the file to free system resources. This process allows reading and processing text data from PDFs programmatically.

```
[12]: page_one_text = page_one.extract_text()
```

```
[13]: page_one_text
```

```
[13]: 'Business Proposal The Revolution is Coming Leverage agile frameworks to provide a robust synopsis for high level overviews. Iterative approaches to corp
orate strategy foster collaborative thinking to further the overall value proposition. Organically grow the holistic world view of disruptive innovation
via workplace diversity and empowerment. Bring to the table win-win survival strategies to ensure proactive domination. At the end of the day, going forw
ard, a new normal that has evolved from generation X is on the runway heading towards a streamlined cloud solution. User generated content in real-time w
ill have multiple touchpoints for offshoring. Capitalize on low hanging fruit to identify a ballpark value added activity to beta test. Override the digi
tal divide with additional clickthroughs from DevOps. Nanotechnology immersion along the information highway will close the loop on focusing solely on th
e bottom line. Podcasting operational change management inside of workflows to establish a framework. Taking seamless key performance indicators offline to
maximise the long tail. Keeping your eye on the ball while performing a deep dive on the start-up mentality to derive convergence on cross-platform integ
ration. Collaboratively administrate empowered markets via plug-and-play networks. Dynamically procrastinate B2C users after installed base benefits. Dram
atically visualize customer directed convergence without revolutionary ROI. Efficiently unleash cross-media information without cross-media value. Quickly
maximize timely deliverables for real-time schemas. Dramatically maintain clicks-and-mortar solutions without functional solutions. BUSINESS PROPOSAL!1'
```

```
[14]: f.close()
```

5. The code reads a PDF file, extracts the first page, and writes it to a new PDF file. It uses PyPDF2.PdfReader to read the original document, PdfWriter to create a new PDF,

and add_page() to add the extracted page before writing it to a new file named "Some_New_Doc.pdf." Finally, the original file is closed to free resources.

```
[23]: f = open('Working_Business_Proposal.pdf','rb')
      pdf_reader = PyPDF2.PdfReader(f)

[24]: page_number = 0
      page_one = pdf_reader.pages[0]

[25]: pdf_writer = PyPDF2.PdfWriter()

[26]: pdf_writer.add_page(page_one);

[27]: pdf_output = open("Some_New_Doc.pdf","wb")

[28]: pdf_writer.write(pdf_output)

[28]: (False, <_io.BufferedWriter name='Some_New_Doc.pdf'>)

[29]: f.close()
```

6. The code reads a PDF file and extracts text from each page, storing it in a list where each index corresponds to a page number. It initializes an empty list, opens the PDF in binary mode, and iterates through all pages using PyPDF2.PdfReader. However, it mistakenly extracts only the first page repeatedly instead of iterating over all pages dynamically.

```
[31]: f = open('Working_Business_Proposal.pdf','rb')

      # List of every page's text.
      # The index will correspond to the page number.
      pdf_text = []

      pdf_reader = PyPDF2.PdfReader(f)

      for p in range(len(pdf_reader.pages)):

          page = pdf_reader.pages[0]

          pdf_text.append(page.extract_text())
```

7. This line attempts to print the text from the fourth page (index 3) of the pdf_text list. However, since the loop in the previous code mistakenly extracts only the first page repeatedly, pdf_text may contain duplicate content instead of unique text from each page. If the PDF has fewer than four pages, this line will raise an IndexError.

[32]:
```
pdf_text
```

[32]:
```
['Business Proposal The Revolution is Coming Leverage agile frameworks to provide a robust synopsis for high level overviews. Iterative approaches to cor
porate strategy foster collaborative thinking to further the overall value proposition. Organically grow the holistic world view of disruptive innovation
via workplace diversity and empowerment. Bring to the table win-win survival strategies to ensure proactive domination. At the end of the day, going forw
ard, a new normal that has evolved from generation X is on the runway heading towards a streamlined cloud solution. User generated content in real-time w
ill have multiple touchpoints for offshoring. Capitalize on low hanging fruit to identify a ballpark value added activity to beta test. Override the digi
tal divide with additional clickthroughs from DevOps. Nanotechnology immersion along the information highway will close the loop on focusing solely on th
e bottom line. Podcasting operational change management inside of workflows to establish a framework. Taking seamless key performance indicators offline to
maximise the long tail. Keeping your eye on the ball while performing a deep dive on the start-up mentality to derive convergence on cross-platform integ
ration. Collaboratively administrate empowered markets via plug-and-play networks. Dynamically procrastinate B2C users after installed base benefits. Dram
atically visualize customer directed convergence without revolutionary ROI. Efficiently unleash cross-media information without cross-media value. Quickly
maximize timely deliverables for real-time schemas. Dramatically maintain clicks-and-mortar solutions without functional solutions. BUSINESS PROPOSAL!1',
 'Business Proposal The Revolution is Coming Leverage agile frameworks to provide a robust synopsis for high level overviews. Iterative approaches to cor
porate strategy foster collaborative thinking to further the overall value proposition. Organically grow the holistic world view of disruptive innovation
via workplace diversity and empowerment. Bring to the table win-win survival strategies to ensure proactive domination. At the end of the day, going forw
ard, a new normal that has evolved from generation X is on the runway heading towards a streamlined cloud solution. User generated content in real-time w
ill have multiple touchpoints for offshoring. Capitalize on low hanging fruit to identify a ballpark value added activity to beta test. Override the digi
tal divide with additional clickthroughs from DevOps. Nanotechnology immersion along the information highway will close the loop on focusing solely on th
e bottom line. Podcasting operational change management inside of workflows to establish a framework. Taking seamless key performance indicators offline to
maximise the long tail. Keeping your eye on the ball while performing a deep dive on the start-up mentality to derive convergence on cross-platform integ
ration. Collaboratively administrate empowered markets via plug-and-play networks. Dynamically procrastinate B2C users after installed base benefits. Dram
atically visualize customer directed convergence without revolutionary ROI. Efficiently unleash cross-media information without cross-media value. Quickly
maximize timely deliverables for real-time schemas. Dramatically maintain clicks-and-mortar solutions without functional solutions. BUSINESS PROPOSAL!1',
```

[33]:
```
print(pdf_text[3])
```

```
Business Proposal The Revolution is Coming Leverage agile frameworks to provide a robust synopsis for high level overviews. Iterative approaches to corpo
rate strategy foster collaborative thinking to further the overall value proposition. Organically grow the holistic world view of disruptive innovation v
ia workplace diversity and empowerment. Bring to the table win-win survival strategies to ensure proactive domination. At the end of the day, going forwa
rd, a new normal that has evolved from generation X is on the runway heading towards a streamlined cloud solution. User generated content in real-time wi
ll have multiple touchpoints for offshoring. Capitalize on low hanging fruit to identify a ballpark value added activity to beta test. Override the digit
al divide with additional clickthroughs from DevOps. Nanotechnology immersion along the information highway will close the loop on focusing solely on the
bottom line. Podcasting operational change management inside of workflows to establish a framework. Taking seamless key performance indicators offline to m
aximise the long tail. Keeping your eye on the ball while performing a deep dive on the start-up mentality to derive convergence on cross-platform integr
ation. Collaboratively administrate empowered markets via plug-and-play networks. Dynamically procrastinate B2C users after installed base benefits. Drama
tically visualize customer directed convergence without revolutionary ROI. Efficiently unleash cross-media information without cross-media value. Quickly
maximize timely deliverables for real-time schemas. Dramatically maintain clicks-and-mortar solutions without functional solutions. BUSINESS PROPOSAL!1
```