**CISB5123 Text Analytics**

**Lab 1**

**Working with Text Data**

**Exercise**

1. Extract text from all pages of Business\_Proposal.pdf and save it in business\_proposal\_all.txt.

import PyPDF2

# Extract text from all pages

def extract\_all\_text(pdf\_path, output\_path):

with open(pdf\_path, 'rb') as file:

reader = PyPDF2.PdfReader(file)

text = "\n".join(page.extract\_text() for page in reader.pages if page.extract\_text())

with open(output\_path, 'w', encoding='utf-8') as output:

output.write(text)

print(f"Extracted text from all pages saved to {output\_path}")

# File paths

pdf\_file = 'Business\_Proposal.pdf'

all\_text\_output = 'business\_proposal\_all.txt'

# Run extraction

extract\_all\_text(pdf\_file, all\_text\_output)

2. Extract text from only page 2 of Business\_Proposal.pdf and save it in business\_proposal\_page\_2.txt.

import PyPDF2

# Extract text from a specific page (Page 2)

def extract\_page\_text(pdf\_path, output\_path, page\_number):

with open(pdf\_path, 'rb') as file:

reader = PyPDF2.PdfReader(file)

if 0 <= page\_number - 1 < len(reader.pages):

text = reader.pages[page\_number - 1].extract\_text()

with open(output\_path, 'w', encoding='utf-8') as output:

output.write(text)

print(f"Extracted text from page {page\_number} saved to {output\_path}")

else:

print("Invalid page number.")

# File paths

pdf\_file = 'Business\_Proposal.pdf'

page\_2\_output = 'business\_proposal\_page\_2.txt'

# Run extraction

extract\_page\_text(pdf\_file, page\_2\_output, 2)