!pip install PyPDF2 openpyxl

from google.colab import files

import re

import PyPDF2

import openpyxl

# Upload PDF file

uploaded = files.upload()

pdf\_path = list(uploaded.keys())[0]

excel\_path = 'invoice\_Breakdown.xlsx'

# Create Excel workbook

wb = openpyxl.Workbook()

ws = wb.active

ws.title = 'Invoices'

ws.append(['Invoice Date', 'BPAY Biller Code', 'BPAY Reference', 'Invoice #', 'HAWB / Entry No', 'Total Outstanding'])

# Function to extract the largest $ amount from a page

def extract\_max\_amount(text):

    amounts = re.findall(r'\$([0-9,]+\.\d{2})', text)

    if amounts:

        max\_amount = max(float(a.replace(',', '')) for a in amounts)

        return f"${max\_amount:,.2f}"

    return 'NOT FOUND'

# Process each page independently

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

    reader = PyPDF2.PdfReader(file)

    for page in reader.pages:

        text = page.extract\_text() or ""

        # Extract data per page

        invoice\_date = re.search(r'Issue Date[:\s]\*([0-9]{1,2}[/-][A-Za-z]{3}[/-][0-9]{2,4})', text)

        biller\_code = re.search(r'Biller Code[:\s]\*(\d+)', text)

        bpay\_ref = re.search(r'Reference[:\s]\*(\d+)', text)

        invoice\_no = re.search(r'(100\d{7,})', text)  # Finds any 10+ digit number starting with 100

        tracking = re.search(r'(?:Flight:\s\*)?((?:1Z|W)[0-9A-Z]{10,})', text)

        entry = re.search(r'\b(Q[0-9A-Z]+)\b', text)

        total\_outstanding = extract\_max\_amount(text)

        # Use tracking number or entry number

        hawb\_or\_entry = tracking.group(1) if tracking else (entry.group(1) if entry else 'NOT FOUND')

        # Append row

        ws.append([

            invoice\_date.group(1) if invoice\_date else 'NOT FOUND',

            biller\_code.group(1) if biller\_code else 'NOT FOUND',

            bpay\_ref.group(1) if bpay\_ref else 'NOT FOUND',

            invoice\_no.group(1) if invoice\_no else 'NOT FOUND',

            hawb\_or\_entry,

            total\_outstanding

        ])

# Save and download

wb.save(excel\_path)

files.download(excel\_path)