**Task report**

**Task**: to extract information from multiple pdfs in a folder

**Steps:**

1. Go to command prompt> run as administrator> type “pip install pdfminer.six”
2. Now open your command prompt and type the following code:

import re # Import regex library

import csv # For csv writing

import os # For filesystem operations

from pdfminer.high\_level import extract\_text # Pdf parsing library

# Compile regex patterns

phone\_pattern=re.compile(r"\b\d{10}\b") # 10 digit phone numbers

email\_pattern=re.compile(r"[\w\.-]+@[\w\.-]+") # Email addresses

college\_pattern=re.compile(r".\*?\b(?:college|university|institute|school)\s+(?:of\s+)?[\w\s-]+.\*",re.IGNORECASE) # College names

degree\_pattern=re.compile(r".\*\b(?:bachelor|bachelors|master|masters|b.tech|m.tech)\b.\*",re.IGNORECASE) # Degree names

skill\_pattern = re.compile(r"(?i)\b(python|java|c\+\+|javascript|aws|power bi|sql|data analytics|statistics|machine learning|ms excel|eda|data visualisation)\b") # Skills

exp\_pattern = re.compile(r'^.\*?\bintern\b.\*?$', re.IGNORECASE | re.MULTILINE) # Internship experience

# Set resume folder path

resume\_folder = 'C:/Users/arora/Desktop/onelogica/OneLogica/OneLogica'

# Iterate through pdf files

for resume\_file in os.listdir(resume\_folder):

# Check for pdf extension

if resume\_file.endswith('.pdf'):

# Extract text using pdfminer

pdf\_path = os.path.join(resume\_folder,resume\_file)

text = extract\_text(pdf\_path)

# Extract details from text using patterns

name = text.strip().split('\n')[0]

phone = phone\_pattern.findall(text)

email = email\_pattern.findall(text)

college = college\_pattern.findall(text)

degree = degree\_pattern.findall(text)

skills = skill\_pattern.findall(text)

experience = exp\_pattern.findall(text)

# Write to csv row wise

with open('C:/Users/arora/Desktop/onelogica/pdfminer/output.xlsx', 'a', newline='') as file:

writer = csv.writer(file)

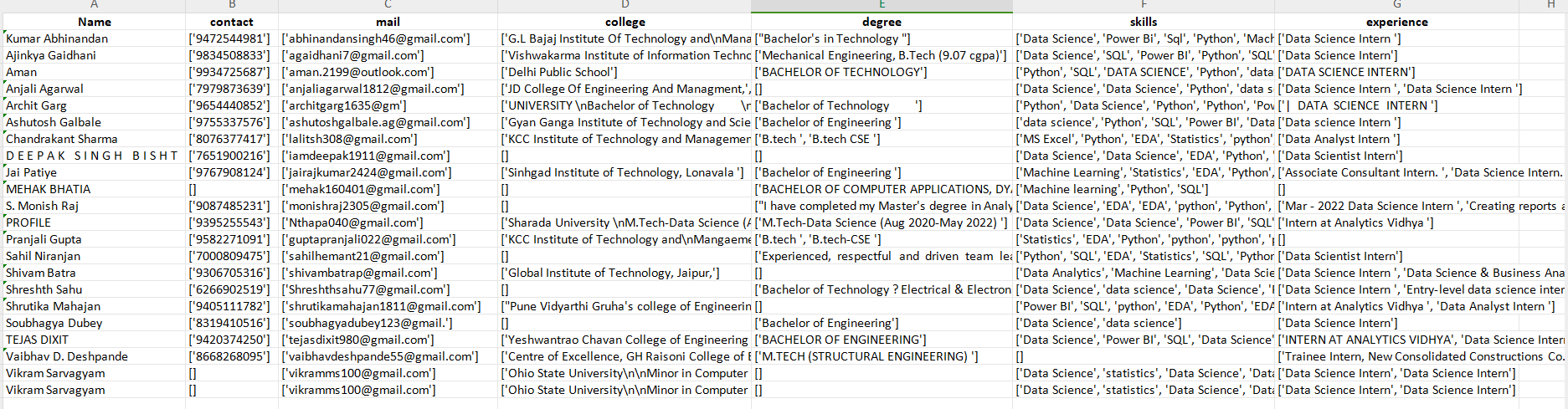
writer.writerow([name,phone,email,college,degree,skills,experience])

print('Resume extraction complete!')

**Output of code:**

****

**Output in file:**



Note: you will have to format the excel sheet a little bit as the data will be saved in csv format. Also You can easily split the columns by setting the delimiter as “,”.