Data Engineer

Let us do a quick end-to-end project which represents the work of a Data Engineer here at qure.ai. You will be assessed on

- Quality and Approach – 50
- Architecture - 25
- Presentation in the report – 25

Please submit a report with your approach. You are free to google and find as many relevant resources as possible. Just make sure to cite them in your report. This project is open-ended. Also feel free to mention any assumptions you made and rationale behind them in the report.

Problem

Medical Data Management 101

When working with medical data encountering dicoms is inevitable. In this problem we want to extract and store the data present inside this widely used data modality in a structured manner.

You are provided with some dicom files containing chest x-rays examination records. These records contain imaging and accompanying metadata.
Your task is to extract data from these dicoms using python scripts to the below mentioned formats.

a) Extract the images from the dicoms into a folder with a suitable name for each extracted image. **Note** – The extracted image should look visually similar to the dicom image. You can use a suitable dicom viewer to view the images (Home Page - Horos Project, MicroDicom - Free DICOM viewer and software etc.)
b) Design a suitable class(es) and store the image and other metadata into the class objects. Explain the design choices in the report
c) Export the extracted metadata into suitably structured Json.

Note: Data privacy is a critical aspect while working with real world medical data. Therefore, it is of utmost importance to anonymize any patient/doctor information that might lead to the identification of the person concerned. Make sure any such information is not present in the data you export.

Download Link for data - assignment_dcms

Problem

Load the energy data from the file Energy Indicators.xls, which is a list of indicators of energy supply and renewable electricity production from the United Nations for the year 2013, and should be put into a DataFrame with the variable name of **energy**.

Keep in mind that this is an Excel file, and not a comma separated values file. Also, make sure to exclude the footer and header information from the datafile. The first two columns are unnecessary, so you should get rid of them, and you should change the column labels so that the columns are:

['Country', 'Energy Supply', 'Energy Supply per Capita', '% Renewable's]

Convert the energy supply and the energy supply per capita to gigajoules (there are 1,000,000 gigajoules in a petajoule). For all countries which have missing data (e.g. data with "...") make sure this is reflected as np.NaN values.

Rename the following list of countries (for use in later questions):

"Republic of Korea": "South Korea", "United States of America": "United States", "United Kingdom of Great Britain and Northern Ireland": "United Kingdom", "China, Hong Kong Special Administrative Region": "Hong Kong"

There are also several countries with parenthesis in their name. Be sure to remove these, e.g. 'Bolivia (Plurinational State of)' should be 'Bolivia'.

Next, load the GDP data from the file world_bank.csv, which is a csv containing countries' GDP from 1960 to 2015 from World Bank. Call this DataFrame **GDP**.

Make sure to skip the header, and rename the following list of countries:

"Korea, Rep.": "South Korea", "Iran, Islamic Rep.": "Iran", "Hong Kong SAR, China": "Hong Kong"

Finally, load the Sciamgo Journal and Country Rank data for Energy Engineering and Power Technology, which ranks countries based on their journal contributions in the aforementioned area. Call this DataFrame **ScimEn**.

Join the three datasets: GDP, Energy, and ScimEn into a new dataset (using the intersection of country names). Use only the last 10 years (2006-2015) of GDP data and only the top 15 countries by Scimagojr 'Rank' (Rank 1 through 15).

The index of this DataFrame should be the name of the country.

This function should return a DataFrame with 20 columns and 15 entries.

Problem

Extract data from the image below which is inside the box into the excel sheet :

Your Company Name

INVOICE

Street Address
City, ST ZIP Code
Phone Number,Web Address, etc.

InvoicingTemplae.com

DATE: March 18, 2015
INVOICE #: INV1001

BILL TO

Name	Test Customer 1
Address	
ity, State ZIP	
Country	
Phone	
Email	
Client #	C1000

SHIP TO

Name	Test Customer 1
Address	
City, State ZIP	
Country	
Contact	

P.O. #	Sales Rep. Name	Ship Date	Ship Via	Terms	Due Date
		3/18/2015			

# / Taxable		Description	Quantity	Unit Price	Line Total
P1002	□	Test Product 3 (Non-taxable)	1	300.00	300.00
P1001	□	Test Product 2 (Service)	1	200.00	200.00
P1000	□	Test Product 1	1	100.00	100.00
	□				
	□				
	□				
	□				
	□				
	□				
	□				
	□				
	□				

SUBTOTAL		600.00
PST	8.000%	24.00
GST	6.000%	18.00
SHIPPING & HANDLING		-
TOTAL		642.00
PAID		-
TOTAL DUE		642.00

PayPal

NOTES:

THANK YOU FOR YOUR BUSINESS!

For references on the task:

Pydicom - Pydicom |

SITK - SimpleITK - Home

Excel Data - https://pandas.pydata.org/pandas-docs/stable/

OCR - https://nanonets.com/blog/ocr-with-tesseract/#:~:text=Tesseract%20OCR%20engine-,OCR%20with%20Pytesseract%20and%20OpenCV,image%20to%20text%20use%20cases.