ASSIGNMENT Solution

**Question:** Parse the data from the below websites to the excel/csv file using any language i.e. Python, R, Java, etc. Also, attach the script/code with the output file.

**Website:** [**https://www.agtta.co.in/individuals.php**](https://www.agtta.co.in/individuals.php)

****

***NOTE: Scrap data for all the members available***

# Sample Output File:



**Solution:**

#!/usr/bin/env python3

# -\*- coding: utf-8 -\*-

"""

Created on Tue Dec 8 23:25:36 2020

@author: vinayaklinux

"""

from bs4 import BeautifulSoup as BS

import urllib3

import json

import re

import requests

import xml.etree.ElementTree as ET

from lxml import html

import pandas as pd

keys = ['Firm','Address','Email','PhoneNo','Route','B.Date','J.Date','Fax']

colnames = ["Contact Person Name","Mobile Number\_1","Mobile Number\_2","Company Name","Address","Phone Number 1","Phone Number 2","Fax","Route","B Date","J Date"]

url = input("Enter url: ")

print(url)

page = requests.get(url)

http = urllib3.PoolManager()

r = http.request('GET', url)

soup = BS(r.data.decode('utf-8'))

sections = soup.findAll("section",{"class":"b-branches"})

names = [section.h3.text for section in sections]

print(len(names))

#print(sections[0].h3.text)

numbers = []

mobile1 = []

mobile2 = []

for section in sections:

n = section.p.text[11::].split(',')

if len(n) == 2:

numbers.append(n)

else:

n.append("NA")

numbers.append(n)

print(len(numbers))

for num in numbers:

mobile1.append(num[0])

mobile2.append(num[1])

print(len(mobile1),len(mobile2))

details = []

for section in sections:

dictionary = {}

data = section.findAll("div",{"class":"b-branches\_\_item"})

for d in data:

s = d.text.split(" ")

if s[0] == "Phone" or s[0] == "B." or s[0] == "J.":

ss = s[0]+s[1]

dictionary[ss] = (" ").join(s[2::])

else:

dictionary[s[0]] = (" ").join(s[1::])

details.append(dictionary)

print(len(details))

for dictt in details:

value = dictt["J.Date"]

value = re.sub(r"LIFETIME MEMBER", "", value)

dictt["J.Date"] = value

for dictt in details:

for key in keys:

val = dictt.setdefault(key)

firm = [dictt['Firm'] for dictt in details]

address = [dictt['Address'] for dictt in details]

email = [dictt['Email'] for dictt in details]

phoneNo = [dictt['PhoneNo'] for dictt in details]

route = [dictt['Route'] for dictt in details]

bDate = [dictt['B.Date'] for dictt in details]

jDate = [dictt['J.Date'] for dictt in details]

fax = [dictt['Fax'] for dictt in details]

print(len(names),len(mobile1),len(mobile2),len(firm),len(address),len(email),len(phoneNo),len(route),len(bDate),len(jDate),len(fax))

phone1 = []

phone2 = []

for p in phoneNo:

if p != None:

ps = p.split(",")

if len(ps) == 1:

phone1.append(ps[0])

phone2.append("NA")

else:

phone1.append(ps[0])

phone2.append(ps[1])

else:

phone1.append("NA")

phone2.append("NA")

print(len(phone1),len(phone2))

names = pd.Series(names)

mobile1 = pd.Series(mobile1)

mobile2 = pd.Series(mobile2)

firm = pd.Series(firm)

address = pd.Series(address)

email = pd.Series(email)

phone1 = pd.Series(phone1)

phone2 = pd.Series(phone2)

route = pd.Series(route)

bDate = pd.Series(bDate)

jDate = pd.Series(jDate)

fax = pd.Series(fax)

colnames = ["Contact Person Name","Mobile Number\_1","Mobile Number\_2","Company Name","Address","Email","Phone Number 1","Phone Number 2","Fax","Route","B Date","J Date"]

dataseries = [names,mobile1,mobile2,firm,address,email,phone1,phone2,fax,route,bDate,jDate]

for vals in dataseries:

for index,val in enumerate(vals):

if val == "NA":

vals[index] = None

result = pd.DataFrame()

for cols,data in zip(colnames,dataseries):

result[cols] = data

print(result)

result.to\_csv("agtta.csv")

**Link to csv file:**

https://drive.google.com/file/d/1QjY2zZFKmavs7xDiB2QyQ-XxNeBTClov/view?usp=sharing

