# Problem

Write a Python script that:

* fetches Caltrans offices page: <http://www.dot.ca.gov/mail.htm>
* extracts offices information
* and returns office info in JSON form as shown below:

[

  {

    "office\_name": "Headquarters",

    "office\_link": "http://www.dot.ca.gov/",

    "office\_address": "1120 N Street",

    "office\_city": "Sacramento",

    "office\_state": null,

    "office\_zip": null,

    "office\_phone": "916-654-5266",

    "mail\_address": null,

    "mail\_pobox": "P.O. Box 942873",

    "mail\_city": "Sacramento",

    "mail\_state": "CA",

    "mail\_zip": "94273-0001",

    "mail\_phone": null

  },

  ...

  {

    "office\_name": "District 12",

    "office\_link": "http://www.dot.ca.gov/dist12/",

    "office\_address": "1750 East 4th Street",

    "office\_city": "Santa Ana",

    "office\_state": "CA",

    "office\_zip": "92705",

    "office\_phone": "657-328-6000",

    "mail\_address": "1750 East 4th Street",

    "mail\_pobox": null,

    "mail\_city": "Santa Ana",

    "mail\_state": "CA",

    "mail\_zip": "92705",

    "mail\_phone": null

  }

]

import requests

from bs4 import BeautifulSoup

def func():

result = [] #list that will be returned

response = requests.get("http://www.dot.ca.gov/mail.htm") #request to get the webpage

soup = BeautifulSoup(response.text,"html.parser") #using Beautiful Soup to parse the respoonse text

rows = soup.find\_all("tr") #finding all table rows

for i in range(1,len(rows)-1): # we have to iterate from 1st table row because 0th row is header and we excelude last row since

#it is irrelevant

mega\_elements = rows[i].find\_all("td") #we find all the table data

newdict = {} #craeting an empty dictionary so that we can add respective office information

for j in range(0,2): #iterate from office data to mailing data

if j == 0: #iterating over office data

xaxa = mega\_elements[j].get\_text() #get office data

xaxalist = xaxa.splitlines() #get a list of this data by splitting on lines

for line in range(0,len(xaxalist)):

xaxalist[line] = xaxalist[line].replace('\xa0',' ') # replace \xa0 chracter with space

xaxalist = [x.strip(' ') for x in xaxalist] #strip the extra space

if len(xaxalist) == 5: #some information for names of offices comes on new line so we combine the two lines

xaxalist[0:2] = [' '.join(xaxalist[0:2])]

newdict["office\_name"] = xaxalist[0] #get office name

if xaxalist[0] == "Headquarters":

newdict["office\_link"] = "http://www.dot.ca.gov/" #get the link if headquarter

else:

name\_list = xaxalist[0].split()

number = name\_list[1]

newdict["office\_link"] = "http://www.dot.ca.gov/dist" + number + '/' #get the link if not headquarter

newdict["office\_address"] = xaxalist[1] #get office address

if ',' not in str(xaxalist[2]): #case when state not present

newdict["office\_city"] = xaxalist[2]

newdict["office\_state"] = None

newdict["office\_zip"] = None

newdict["office\_phone"] = xaxalist[3]

else: #case when state is present

city\_state\_zip = xaxalist[2].split(',')

newdict["office\_city"] = xaxalist[2].split(',')[0]

newdict["office\_state"] = xaxalist[2].split(',')[1].split()[0]

newdict["office\_zip"] = xaxalist[2].split(',')[1].split()[1]

newdict["office\_phone"] = xaxalist[3]

else: #iterate over mailing address data

xaxa = mega\_elements[j].get\_text()

xaxalist = xaxa.splitlines()

for line in range(0,len(xaxalist)):

xaxalist[line] = xaxalist[line].replace('\xa0',' ')

xaxalist = [x.strip(' ') for x in xaxalist]

if len(xaxalist) == 2: #case when there are only 2 lines in the data

if 'Box' not in xaxalist[0].split(): #case when PO BOX is not there

newdict["mail\_address"] = xaxalist[0]

newdict["mail\_pobox"] = None

newdict["mail\_city"] = xaxalist[1].split(',')[0]

newdict["mail\_state"] = xaxalist[1].split(',')[1].split()[0]

newdict["mail\_zip"] = xaxalist[1].split(',')[1].split()[1]

newdict["mail\_phone"] = None

else: #case when PO Box is there

newdict["mail\_address"] = None

newdict["mail\_pobox"] = xaxalist[0]

newdict["mail\_city"] = xaxalist[1].split(',')[0]

newdict["mail\_state"] = xaxalist[1].split(',')[1].split()[0]

newdict["mail\_zip"] = xaxalist[1].split(',')[1].split()[1]

newdict["mail\_phone"] = None

else: #case when there are more than 2 lines in the data

newdict["mail\_address"] = xaxalist[0]

if 'Box' not in xaxalist[1].split():

newdict["mail\_pobox"] = None

newdict["mail\_city"] = xaxalist[1].split(',')[0]

newdict["mail\_state"] = xaxalist[1].split(',')[1].split()[0]

newdict["mail\_zip"] = xaxalist[1].split(',')[1].split()[1]

if len(xaxalist) > 2:

newdict["mail\_phone"] = xaxalist[2]

else:

newdict["mail\_phone"] = None

else:

newdict["mail\_pobox"] = xaxalist[1]

newdict["mail\_city"] = xaxalist[2].split(',')[0]

newdict["mail\_state"] = xaxalist[2].split(',')[1].split()[0]

newdict["mail\_zip"] = xaxalist[2].split(',')[1].split()[1]

if len(xaxalist) > 3:

newdict["mail\_phone"] = xaxalist[3]

else:

newdict["mail\_phone"] = None

result.append(newdict) #appned the dictionary to the result

return result

mli = func()

for i in mli:

print(i)

print('\n\n')