VirtuOx CIGNA Database

Kinetix Solutions

This report contains the information regarding a web scrapping project

Kelvin Njeri. (Part Time Intern)

10/16/2016

VirtuOx CIGNA Database

Kinetix Solutions

Consumer Request

VirtuOx request a way to get information from the webpage "http://cigna.benefitnation.net/cigna/SearchResult.aspx?txtAddress=&zip=33706&radius=150&searchType=phys&searchName=&city=&state=&ccn=Y&lang=&gender=&hplan=&hca=&hden=&hphm=&hvis=&productplan=OA&netid=&NPOid=&ResidentZipCode=&Pspec=&role=S&Sspec=PL " which contains physicians contact information within the CIGNA data base. The consumer wanted the physicians' names, their practice names, address and phone numbers from all 43k plus zip codes in the United States of America

Solution Approach

The solution to this problem is software based. I used a high level language Python 2.7.12 using PyCharm Edu 3.0. The program first gets information of all USA zip codes from a file and stores in a data structure called list. It then uses a loop to go through all the stored zip codes replacing the zip section "&zip=33706" using "'&zip=%d' %x", with **x** being the variable that contains the zip code. Each loop changes the url, which changes the generated list of physicians within 150miles of their zip code.

The list is then collecting using a web scrapping technique. This involves getting information from respective HTML tags such as <table></table>, <tr></tr> and <td></td> that has the information required.

The information was stored in the data structure in accordance to the consumer's requests. This was then stored in an excel/csv file.

Processing information from over 43k zip codes takes a long time and any internet disruptions will cause a halt in the processing of the required information. To solve this problem, I broke down the process to get the information 1000 zip codes at a time. This cuts down the time greatly in getting the information. The data was split in over 43 files and the next task was to combine all the contacts in one file, edit it, remove duplicate files and rearrange the data alphabetically.

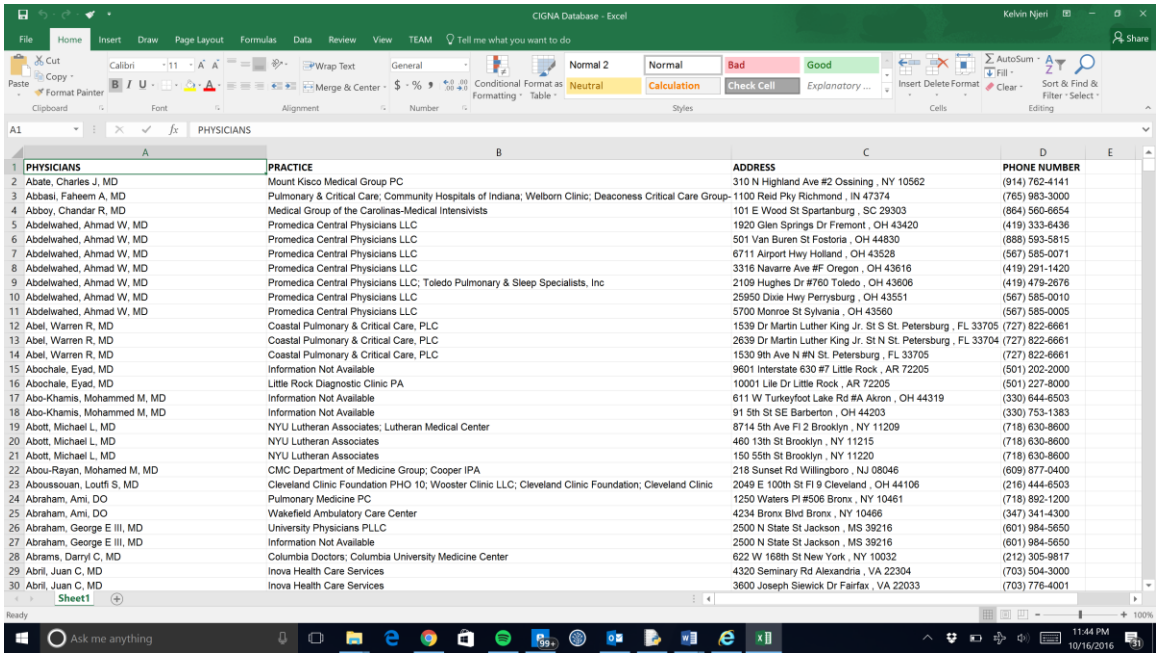
Program Code

```
# Kinetix Solutions, Inc
# VIRTOUX Screen Scrapper request
# author: Kelvin Njeri
# kelvin@kinetixsolutions.com
# IDLE: PyCharm 3.0 (python 2.7.12)
# The purpose of this script is to extract data from a
# generated list of contact information in a web page

# necessary imports
import csv
import urllib2
from xlwt import *
import xlrd

try:
   from bs4 import BeautifulSoup
except ImportError:
   from BeautifulSoup import BeautifulSoup

# storing information on a file of type xlxs
book = Workbook()

# storing 1000 USA zip codes
with open('codes.txt') as f:
   lines = f.readlines()

zips = []
for x in range(1000):
   zips.append(lines[x])

# extracting and storing data in a list of list
def extractData(line, contact):
   try:
      for row in line:
         list_of_cells = []
         for cell in row.find_all('td')[1:]:
            data = cell.text.replace(u'\xa0', '')
            data.replace(u'\xae', '')
            data.replace('\n', '')
            data.replace('\t', '')
            data.lstrip()
            data.rstrip()
            data.encode('utf-8')
            list_of_cells.append(data)
         contact.append(list_of_cells)  # created lists of lists
         #print "number of contacts collected = %d" % len(contact)
   except Exception as p:
      print p
      pass
```

```
# loops through every zip code in the search option
for x in zips:
    contact = []  # stores primary contact information
    physician = []
    practice = []
    address = []
    sheet = book.add_sheet(x)
    z = int(x)
    print "Accessing information in zipcode = %s" % x
    link =
'http://cigna.benefitnation.net/cigna/SearchResult.aspx?txtAddress=&zip=%s&radius=150&s
earchType=phys&searchName=&city=&state=&ccn=Y&lang=&gender=&hplan=&hca=&hde
n=&hphm=&hvis=&productplan=OA&netid=&NPOid=&ResidentZipCode=&Pspec=&role=S
&Sspec=PL' % z

    try:  # use of beautiful soup feature
        http = urllib2.urlopen(link)
        soup = BeautifulSoup(http, "html5lib")
    except urllib2.HTTPError:
        pass #clears out invalid URLs
    except Exception as he:
        print (he)
        exit()
    except AttributeError as se:
        print (se)
        exit()
    oddLine = soup.find_all('tr', {'class': 'resultslistitem'})
    evenLine = soup.find_all('tr', {'class': 'gridAlter'})
    extractData(oddLine, contact)
    extractData(evenLine, contact)

    for x in contact:
        p = x[0]
        j = x[3]
        a = x[4]
        print (p, j, a)

    for x in contact:
        physician.append(x[0])
        practice.append(x[3])
        address.append(x[4])

    for n in range(len(contact)):
        sheet.write(n, 0, physician[n])
        sheet.write(n, 1, practice[n])
        sheet.write(n, 2, address[n])

    book.save("cigna 1-1000.xlxs") #sample file name
```

Python File: Screen Scrapper.py

Database



PHYSICIANS	PRACTICE	ADDRESS	PHONE NUMBER
Abate, Charles J, MD	Mount Kisco Medical Group PC	310 N Highland Ave #2 Ossining , NY 10562	(914) 762-4141
Abbasi, Faheem A, MD	Pulmonary & Critical Care; Community Hospitals of Indiana; Welborn Clinic; Deaconess Critical Care Group-	1100 Reid Pky Richmond , IN 47374	(765) 983-3000
Abboy, Chandar R, MD	Medical Group of the Carolinas-Medical Intensivists	101 E Wood St Spartanburg , SC 29303	(864) 560-6654
Abdelwahed, Ahmad W, MD	Promedica Central Physicians LLC	1920 Glen Springs Dr Fremont , OH 43420	(419) 333-6436
Abdelwahed, Ahmad W, MD	Promedica Central Physicians LLC	501 Van Buren St Fostoria , OH 44830	(888) 593-5815
Abdelwahed, Ahmad W, MD	Promedica Central Physicians LLC	6711 Airport Hwy Holland , OH 43528	(567) 585-0071
Abdelwahed, Ahmad W, MD	Promedica Central Physicians LLC	3316 Navarre Ave #F Oregon , OH 43616	(419) 291-1420
Abdelwahed, Ahmad W, MD	Promedica Central Physicians LLC; Toledo Pulmonary & Sleep Specialists, Inc	2109 Hughes Dr #760 Toledo , OH 43606	(419) 479-2676
Abdelwahed, Ahmad W, MD	Promedica Central Physicians LLC	25950 Dixie Hwy Perrysburg , OH 43551	(567) 585-0010
Abdelwahed, Ahmad W, MD	Promedica Central Physicians LLC	5700 Monroe St Sylvania , OH 43560	(567) 585-0005
Abel, Warren R, MD	Coastal Pulmonary & Critical Care, PLC	1539 Dr Martin Luther King Jr. St S St. Petersburg , FL 33705	(727) 822-6661
Abel, Warren R, MD	Coastal Pulmonary & Critical Care, PLC	2639 Dr Martin Luther King Jr. St N St. Petersburg , FL 33704	(727) 822-6661
Abel, Warren R, MD	Coastal Pulmonary & Critical Care, PLC	1530 9th Ave N #N St. Petersburg , FL 33705	(727) 822-6661
Abochale, Eyad, MD	Information Not Available	9601 Interstate 630 #7 Little Rock , AR 72205	(501) 202-2000
Abochale, Eyad, MD	Little Rock Diagnostic Clinic PA	10001 Lile Dr Little Rock , AR 72205	(501) 227-8000
Abo-Khamis, Mohammed M, MD	Information Not Available	611 W Turkeyfoot Lake Rd #A Akron , OH 44319	(330) 644-6503
Abo-Khamis, Mohammed M, MD	Information Not Available	91 5th St SE Barberton , OH 44203	(330) 753-1383
Abott, Michael L, MD	NYU Lutheran Associates; Lutheran Medical Center	8714 5th Ave Fl 2 Brooklyn , NY 11209	(718) 630-8600
Abott, Michael L, MD	NYU Lutheran Associates	460 13th St Brooklyn , NY 11215	(718) 630-8600
Abott, Michael L, MD	NYU Lutheran Associates	150 55th St Brooklyn , NY 11220	(718) 630-8600
Abou-Rayan, Mohamed M, MD	CMC Department of Medicine Group; Cooper IPA	218 Sunset Rd Willingboro , NJ 08046	(609) 877-0400
Aboussouan, Loutfi S, MD	Cleveland Clinic Foundation PHO 10; Wooster Clinic LLC; Cleveland Clinic Foundation; Cleveland Clinic	2049 E 100th St Fl 9 Cleveland , OH 44106	(216) 444-6503
Abraham, Ami, DO	Pulmonary Medicine PC	1250 Waters Pl #506 Bronx , NY 10461	(718) 892-1200
Abraham, Ami, DO	Wakefield Ambulatory Care Center	4234 Bronx Blvd Bronx , NY 10466	(347) 341-4300
Abraham, George E III, MD	University Physicians PLLC	2500 N State St Jackson , MS 39216	(601) 984-5650
Abraham, George E III, MD	Information Not Available	2500 N State St Jackson , MS 39216	(601) 984-5650
Abrams, Darryl C, MD	Columbia Doctors; Columbia University Medicine Center	622 W 168th St New York , NY 10032	(212) 305-9817
Abril, Juan C, MD	Inova Health Care Services	4320 Seminary Rd Alexandria , VA 22304	(703) 504-3000
Abril, Juan C, MD	Inova Health Care Services	3600 Joseph Siewick Dr Fairfax , VA 22033	(703) 776-4001

Figure 1: Sample Output

Excel file: CIGNA Database