WEBSCRAPING USING PYTHON

Scraping “coldwellbanker.com” for agent’s clients details.

Language : Python

Package : BeautifulSoup

CODE

import requests

from bs4 import BeautifulSoup

import os,urllib,json,urllib2,MySQLdb

r=requests.get(“https://www.coldwellbanker.com/agents/Melissa-Hjeltness-83814”)

soup=BeautifulSoup(r.content,’lxml’)

db=MySQLdb.connect("localhost","root","","laravel") //DB connection

cursor=db.cursor()

Name=soup.find(“h1”,{“class”:”notranslate”})

cursor.execute("insert into agents(Name) values(%s)",(str(Name.text)))

db.commit()

addr2=soup.find\_all("a",{"data-ga-action":"viewed-similar-property"})

url="https://www.coldwellbanker.com"

url\_list=[]

i=0

for x in addr2:

i=i+1

url\_list.insert(i,url+x['href'])

CODE TO VISIT EACH PROPERTY IN AGENT’S PROFILE AND SAVING THE DETAILS IN JSON FILE:

bed\_bath={}

url\_dict={}

address\_dict={}

length=len(url\_list)

i=0

k=0

with open("name.json","w") as outfile:

outfile.write("{")

outfile.write('"property"')

outfile.write(":")

outfile.write("{")

for x in url\_list:

response=urllib2.urlopen(x)

soup=BeautifulSoup(response,'lxml')

dt=soup.find\_all("dt")

dd=soup.find\_all("dd")

name=os.path.basename(x)

img=soup.find\_all("img",{"class":"imgfill"})

for x in img:

k=k+1

value=x['src']

urllib.urlretrieve(value,"%d.jpg" %(k))

address=soup.find("address")

address\_text=address.get\_text(strip=True)

email=soup.find("a",{"class":"email email-lead"})

phone=soup.find("a",{"class":"phone-lead"})

price=soup.find("strong",{"class":"price"})

json.dump(name,outfile)

outfile.write(":")

for x,y in zip(dt,dd):

bed\_bath.update({str(x.text):str(y.get\_text(strip=True))})

bed\_bath.update({"address":address\_text})

bed\_bath.update({"email":email.text})

bed\_bath.update({“price”:str(price)})

json.dump(bed\_bath,outfile)

bed\_bath.clear()

i=i+1

if i<>length:

outfile.write(",")

else:

outfile.write("}}")

**JSON FILE TO DB:**

cursor.execute(“select Id from agents where Name=%s”,str(Name.text))

id=”%x” %cursor.fetchone()

import re

with open("name.json") as outfile:

values=json.load(outfile)

for x in url\_list:

prop=values.get("property")

name=os.path.basename(x)

each= prop.get(name)

if 'Beds:' in each:

beds=int(re.search(r'\d+',each['Beds:']).group())

else:

beds=0

if 'Baths:' in each:

baths=int(re.search(r'\d+',each['Baths:']).group())

else:

baths=0

if 'Year Built:' in each:

year=int(re.search(r'\d+',each['Year Built:']).group())

else:

year=0000

if 'Property Type:' in each:

type=each['Property Type:']

else:

type='no type'

sql="insert into scraping(Id,Year,Beds,Baths) values(%s,%d,%d,%d)"%(id,year,beds,baths)

cursor.execute(sql)

db.commit()

cursor.execute("update scraping set Type=%s where Id=%s and Beds=%s",(type,id,beds))

db.commit()