Web Scraping in Python for Data Extraction

Web scrapping means scrapping the information from web pages for example info about real estate to know the price trend over time and place .the data will be not arranged in excel sheets so we cant see the clear data by download those however they be spread over HTML web pages

Of real estates in the websites so we scrap them and put them in well structure format into excel sheets using python and beautifulsoup.

Beautiful Soup will help us read the Html document. It picks the text from the response and parses the information in a way that makes it easier for us to navigate in its structure and get its contents. The requests module allows you to send HTTP requests using Python.

First we have to install the required libraries such as Requests,bs4(BeautifulSoup)

Lets import our libraries,

```
In [1]: import requests
        from bs4 import BeautifulSoup
```

We create a variable to load the source data using get method which gets the URL required so tht create request object and stores it in another variable

```
In [1]: import requests
        from bs4 import BeautifulSoup

In [3]: r=requests.get("http://pythonhow.com/example.html")
        c=r.content
```

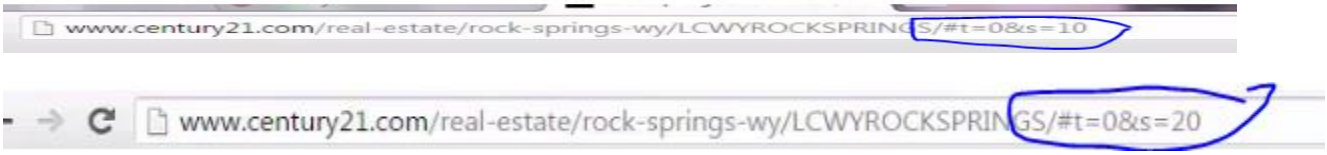
Now the beautifulsoup does the parsing of our source code giving the required info with specific html tags and divisions

```
In [6]: soup=BeautifulSoup(c,"html.parser")

In [7]: print(soup.prettify())
<!DOCTYPE html>
<html>
 <head>
  <style>
   div.cities {
    background-color:black;
    color:white;
    margin:20px;
    padding:20px;
}
```

We can also extract the specific tags such as header tags using Findall methods in our HTML script.

We further add something that can be extract the data from two or more pages ,we use the technique that URL changes as we go to the next page for example,



url gets to change every single and we can observe that It will change ,so we can grab the first url and we can iterate through the loop by increasing 10,10 each time and so on

```
In [39]: base_url="http://www.century21.com/real-estate/rock-springs-wy/LCWYROCKSPRINGS/#t=0&s="
for page in range(0,30,10):
    print(base_url+str(page))
    r=requests.get(base_url+str(page))
    c=r.content
    soup=BeautifulSoup(c,"html.parser")
    print(soup.prettify())
```

And we code on the necessisities we will actually get the information that we need as follows

In [13]:
```
df
```

Out[13]:

	Address	Area	Beds	Full Baths	Half Baths	Locality	Lot Size	Price
0	0 Gateway	None	None	None	None	Rock Springs, WY 82901	NaN	$725,000
1	1003 Winchester Blvd.	None	4	4	None	Rock Springs, WY 82901	0.21 Acres	$452,900
2	3239 Spearhead Way	3,076	4	3	1	Rock Springs, WY 82901	Under 1/2 Acre,	$379,900
3	600 Talladega	3,154	5	3	None	Rock Springs, WY 82901	NaN	$379,000
4	3457 Brisol Avenue	3,236	5	3	None	Rock Springs, WY 82901	0.34 Acres	$349,900
5	234 Via Spoleto	2,688	4	3	None	Rock Springs, WY 82901	Under 1/2 Acre,	$330,000
6	2425 Cripple Creek	8,263	4	35	None	Rock Springs, WY 82901	NaN	$279,900
7	522 Emerald Street	1,172	3	3	None	Rock Springs, WY 82901	Under 1/2 Acre,	$254,000
8	1302 Veteran's Drive	1,932	4	2	None	Rock Springs, WY 82901	0.27 Acres	$252,900
9	343 Via Rucce	None	3	2	1	Rock Springs, WY 82901	0.16 Acres	$219,900
10	913 Madison Dr	1,344	3	2	None	Rock Springs, WY 82901	Under 1/2 Acre,	$209,000
11	3107 White Mountain Blvd	1,540	3	2	1	Rock Springs, WY 82901	Under 1/2 Acre,	$204,900
12	1021 Cypress Cir	1,676	4	3	None	Rock Springs, WY 82901	Under 1/2 Acre,	$199,900
13	4 Minnies Lane	1,664	3	2	None	Rock Springs, WY 82901	2.02 Acres	$196,900
14	910 Eisenhower DR	1,858	3	1	None	Rock Springs, WY 82901	Under 1/2 Acre,	$194,900
15	826 Bushnell Ave	1,038	3	2	None	Rock Springs, WY 82901	0.11 Acres	$189,000
16	845 Ridge Ave 839 and 843 Ridge	1,584	5	4	None	Rock Springs, WY 82901	Under 1/2 Acre,	$185,000
17	505 Ridge Avenue	2,316	5	2	None	Rock Springs, WY 82901	0.22 Acres	$169,900
18	2600 Pueblo Trl	1,820	4	3	None	Rock Springs, WY 82901	0.11 Acres	$144,000