Assignment-4

B kanaka durga

Step-1

pip install requests

pip install html5lib

pip install bs4

Step 2: Accessing the HTML content from webpage

```
import requests
URL = "https://www.geeksforgeeks.org/data-structures/"
r = requests.get(URL)
print(r.content)
headers = {'User-Agent': "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/42.0.2311.135 Safari/537.36 Edge/12.246"}
# Here the user agent is for Edge browser on windows 10. You can find your browser user
agent from the above given link.
r = requests.get(url=URL, headers=headers)
print(r.content)
```

Step 3: Parsing the HTML content

```
This will not run on online IDE
import requests
from bs4 import BeautifulSoup

URL = "http://www.values.com/inspirational-quotes"
r = requests.get(URL)

soup = BeautifulSoup(r.content, 'html5lib') # If this line causes an error, run 'pip install
html5lib' or install html5lib
print(soup.prettify())
 soup = BeautifulSoup(r.content, 'html5lib')
```

Step 4: Searching and navigating through the parse tree

```
#Python program to scrape website
#and save quotes from website
import requests
from bs4 import BeautifulSoup
import csv

URL = "http://www.values.com/inspirational-quotes"
r = requests.get(URL)
```

```
soup = BeautifulSoup(r.content, 'html5lib')

quotes=[]  # a list to store quotes

table = soup.find('div', attrs = {'id':'all_quotes'})

for row in table.findAll('div',
                         attrs = {'class':'col-6 col-lg-3 text-center margin-30px-bottom sm-margin-30px-
top'}):
    quote = {}
    quote['theme'] = row.h5.text
    quote['url'] = row.a['href']
    quote['img'] = row.img['src']
    quote['lines'] = row.img['alt'].split(" #")[0]
    quote['author'] = row.img['alt'].split(" #")[1]
    quotes.append(quote)
filename = 'inspirational_quotes.csv'
with open(filename, 'w', newline='') as f:
    w = csv.DictWriter(f,['theme','url','img','lines','author'])
    w.writeheader()
    for quote in quotes:
        w.writerow(quote)

table = soup.find('div', attrs = {'id':'all_quotes'})

for row in table.find_all_next('div', attrs = {'class': 'col-6 col-lg-3 text-center margin-30px-
bottom sm-margin-30px-top'}):

    quote = {}

    quote['theme'] = row.h5.text

    quote['url'] = row.a['href']

    quote['img'] = row.img['src']

    quote['lines'] = row.img['alt'].split(" #")[0]

    quote['author'] = row.img['alt'].split(" #")[1]

    quotes.append(quote)

quote['theme'] = row.h5.text

quote['url'] = row.a['href']

filename = 'inspirational_quotes.csv'

with open(filename, 'w', newline='') as f:
```

```
w = csv.DictWriter(f,['theme','url','img','lines','author'])
w.writeheader()
for quote in quotes:
    w.writerow(quote)
```