Web scraping assignment 4:

**Question 1:**

```python
import requests
from bs4 import BeautifulSoup

url = 'https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos'

response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

table = soup.find('table', {'class':'wikitable plainrowheaders'})

for row in table.find_all('tr')[1:]:
    columns = row.find_all('td')

    rank = columns[0].text.strip()
    name = columns[1].text.strip()
    artist = columns[2].text.strip()
    upload_date = columns[3].text.strip()
    views = columns[4].text.strip()

    print(rank, name, artist, upload_date, views)
```

**Question 2:**

```python
import requests
```

```
from bs4 import BeautifulSoup

url = 'https://www.bcci.tv/'

response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

fixtures = soup.find('div', {'class':'col-md-9'}):

for fixture in fixtures.find_all('div', {'class':'list-group'}):
    series = fixture.find('h4').text.strip()

    for match in fixture.find_all('a'):
        place = match.find('span', {'class':'text-muted'}).text.strip()
        date = match.find('span', {'class':'text-sm'}).text.strip()
        time = match.find('span', {'class':'text-sm'}).text.split('-')[1].strip()

        print(series, place, date, time)
```

**Question 3:**

```python
import requests
from bs4 import BeautifulSoup

url = 'http://statisticstimes.com/'

response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')
```

```
table = soup.find('table', {'class':'table table-striped'})

for row in table.find_all('tr')[1:]:
    columns = row.find_all('td')

    rank = columns[0].text.strip()
    state = columns[1].text.strip()
    gsdp1819 = columns[2].text.strip()
    gsdp1920 = columns[3].text.strip()
    share1819 = columns[4].text.strip()
    gdp = columns[5].text.strip()

    print(rank, state, gsdp1819, gsdp1920, share1819, gdp)
```

**Question 4:**

```python
import requests
from bs4 import BeautifulSoup

url = 'https://github.com/trending'

response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

repos = soup.find('div', {'class':'application-main'}).find('div', {'class':'repo-list'}).find_all('li',
{'class':'col-12 d-block width-full py-4 border-bottom'})

for repo in repos:
    title = repo.find('h3').text.strip()
```

```
    desc = repo.find('p', {'class':'col-9 color-text-secondary my-1 pr-4'}).text.strip()
    contributors = repo.find('span', {'class': 'd-inline-block float-sm-right'}).text.strip()
    language = repo.find('span', {'itemprop': 'programmingLanguage'}).text.strip()

    print(title, desc, contributors, language)
```

**Question 5:**

```python
import requests
from bs4 import BeautifulSoup

url = 'https://www.billboard.com/charts/hot-100'

response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

songs = soup.find('div', {'class':'chart-list'}).find_all('div', {'class':'o-chart-results-list-row-container'})

for song in songs:
    name = song.find('h3', {'class':'c-title'}).text.strip()
    artist = song.find('span', {'class':'c-label'}).text.strip()
    last_week = song.find('span', {'class':'c-week-current'}).text.strip()
    peak_rank = song.find('td', {'class':'c-ells'}).text.strip()
    weeks_on_board = song.find_all('td', {'class':'c-ells'})[1].text.strip()

    print(name, artist, last_week, peak_rank, weeks_on_board)
```

**Question 6:**

```python
import requests
from bs4 import BeautifulSoup

url = 'https://www.theguardian.com/news/datablog/2012/aug/09/best-selling-books-all-time-fifty-shades-grey-compare'

response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

table = soup.find('table', {'class':'in-article sortable'}).find('tbody')

for row in table.find_all('tr'):
    book = row.find('td').text.strip()
    author = row.find_all('td')[1].text.strip()
    volumes = row.find_all('td')[2].text.strip()
    publisher = row.find_all('td')[3].text.strip()
    genre = row.find_all('td')[4].text.strip()

    print(book, author, volumes, publisher, genre)
```

**Question 7:**

```python
import requests
from bs4 import BeautifulSoup

url = 'https://www.imdb.com/list/ls095964455/'
```

```
response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

shows = soup.find('div', {'class':'lister-list'}).find_all('div', {'class':'lister-item mode-advanced'})

for show in shows:
    name = show.find('h3', {'class':'lister-item-header'}).text.strip()
    year_span = show.find('span', {'class':'lister-item-year'}).text.strip()
    genre = show.find('span', {'class':'genre'}).text.strip()
    runtime = show.find('span', {'class':'runtime'}).text.strip()
    rating = show.find('div', {'class':'inline-block ratings-imdb-rating'}).text.strip()
    votes = show.find('span', {'name':'nv'}).text.strip()

    print(name, year_span, genre, runtime, rating, votes)
```

**Question 8:**

```python
import requests
from bs4 import BeautifulSoup

url = 'https://archive.ics.uci.edu/ml/datasets.php'

response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

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

for row in table.find_all('tr')[1:]:
    columns = row.find_all('td')
```

```
    name = columns[0].text.strip()
    data_type = columns[1].text.strip()
    task = columns[2].text.strip()
    attributes = columns[3].text.strip()
    instances = columns[4].text.strip()
    attributes_num = columns[5].text.strip()
    year = columns[6].text.strip()

    print(name, data_type, task, attributes, instances, attributes_num, year)
```