Module 10: Web Scraping using Python and Project Work

Assignment Solution

1. Write down a program to find urls from any website that you pass as command line argument. You may pass one or multiple web pages at a time.

Example: You may want to execute it by:

Python **your_script.py** http://reddit.com http://yellowpages.com

Note: It may not work for all the websites as we cannot help if they keep them under different tags, but should work for few.

Feel free to use any package that you may need to.

Solution

```
# generic_url_finder.py

import sys

import urllib

import urlparse

from bs4 import BeautifulSoup

class urlOpen(urllib.FancyURLopener):

   version = 'Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.15)
Gecko/20110303 Firefox/3.6.15'


def use_url(url):

   openurl = urlOpen()

   read_page = openurl.open(url)

   raw_text = read_page.read()

   read_page.close()


   soup = BeautifulSoup(raw_text)


   for tag in soup.find_all('a', href=True,limit=20):
```

```
        tag['href'] = urlparse.urljoin(url, tag['href'])
        print tag['href']


def main():
    if len(sys.argv) == 1:
        print "Usage: %s URL [URL]..." % sys.argv[0]
        sys.exit(-1)
    for url in sys.argv[1:]:
        use_url(url)


if __name__ == "__main__":
    main()
```

2. Write a web scraping program to display top 250 movies rated in IMDB.com. Output should be in the below format:

Seral num . Movie name (Release_year)

Ex 242. Gravity (2013)

You are free to use any package of your choice.

Solution

```
#import requests
import urllib2
from bs4 import BeautifulSoup
url = 'http://www.imdb.com/chart/top?ref_=nv_ch_250_4'
test_url = urllib2.urlopen(url)
readHtml = test_url.read()
test_url.close()
soup = BeautifulSoup(readHtml)
bs = BeautifulSoup(readHtml)
```

```
for movie in bs.find_all('td', 'titleColumn'):
    title = movie.get_text().strip().split('\n')
    print title[0],title[1],title[2]
```