```
# Introduction to Web Scraping with Python
# NaLette Brodnax nbrodnax@indiana.edu
# February 5, 2016

# ACCESS #
# Import all the libraries that you need
import requests
import bs4
import csv

webpage = 'http://www.amstat.org/publications/jse/jse_data_archive.htm'
server_response = requests.get(webpage)


# PARSE #
soup = bs4.BeautifulSoup(server_response.text)

# create a list of dictionaries (one dict for each link)
link_info_list = []
for tag in soup.find_all('a'):
    link = tag['href']
    name = tag.text
    # print(name)
    if name[-3:] == 'txt':
        link_info_list.append({'link': link, 'name': name})


# TRANSFORM #
# add a new category to each dictionary to categorize the file as either
# data or documentation
host = 'http://www.amstat.org/publications/jse/'
for dataset in link_info_list[:3]:
    url = host + dataset['link']
    data_response = requests.get(url)
    if data_response.text[:5] == 'NAME:':
        dataset['type'] = 'doc'
    else:
        dataset['type'] = 'dat'


# STORE #
def download_to_txt(file_name, data):
    with open(file_name, 'w') as txtfile:
        txtfile.writelines(data)


def strip_extension(file_name):
    i = -1  # start at the end of the filename
```

```
49      while i > -len(file_name):
50          if file_name[i] == '.':
51              break  # stop when you get to a period
52          else:
53              i -= 1  # this is the same as i = i - 1
54      return file_name[:i]
55
56  # store individual data files as text files
57  for dataset in link_info_list[:3]:
58      url = host + dataset['link']
59      data_response = requests.get(url)
60      description = strip_extension(dataset['name'])
61      filename = description + '_' + dataset['type'] + '.txt'
62      download_to_txt(filename, data_response.text)
63
64  # store list of links as csv file
65  with open('data_links.csv', 'w') as csvfile:
66      fieldnames = ['link', 'name', 'type']
67      writer = csv.DictWriter(csvfile, fieldnames)
68      writer.writeheader()
69      for link in link_info_list:
70          writer.writerow(link)
71      print('Links added to file: ' + str(len(link_info_list)
    ))
72
```