Web scraping using Python

Michelle

February 1, 2018

The following documentation assumes user has installed Google Chrome.

1. Download and install Anaconda from www.anaconda.com/download/.
2. To install Selenium library using Anaconda, open Anaconda Prompt and run this command:

   ```
   conda install -c conda-forge selenium
   ```

3. In this example, the keyword used is "tesla". News headlines and abstracts on Tesla are saved as text files for further analysis.

   Copy and paste the code in Spyder. **Please fix the indentations because all the indentations necessary for the loop to work will disappear when copying and pasting.** Save the code as a .py file. This step can be performed using an IDE or a text editor of your choice.

   ```
   import codecs
   import time
   from selenium import webdriver
   from selenium.webdriver.common.keys import Keys
   import urllib
   import requests
   from urllib.request import urlopen
   from urllib.parse import urljoin
   from urllib.request import pathname2url
   from bs4 import BeautifulSoup
   import os

   KEYWORDS = "tesla"
   n = 2 # Specify pages required

   # Optional argument, if not specified will search path.
   driver = webdriver.Chrome("./chromedriver")
   driver.get("https://www.google.com.sg")

   # Programmatically control the browser such as clicking, entering text, etc
   # Enter keywords
   search_input = driver.find_element_by_id("lst-ib")
   search_input.send_keys(KEYWORDS)

   btn_input = driver.find_element_by_name("btnK")
   btn_input.send_keys(Keys.RETURN)

   time.sleep(3)
   ```

```
# Click on News
news_btn = driver.find_element_by_link_text("News")
news_btn.click()

for i in range(0, n): # Example loop
    # Save search
    html = driver.page_source
    f_out = codecs.open(KEYWORDS + str(i+1) + ".html", "w", "utf-8")
    f_out.write(html)

    # Read the saved html file
    dir = os.path.dirname(os.path.abspath(__file__))

    url = (urljoin("file:", pathname2url(dir) + "/%s" % KEYWORDS + str(i+1) + ".html"))

    html = urllib.request.urlopen(url).read()
    soup = BeautifulSoup(html, "lxml")

    # Remove all script and style elements
    for script in soup(["script", "style"]):
        script.extract()

    # Get text
    text = soup.get_text()

    # Break into lines and remove leading and trailing space
    lines = (line.strip() for line in text.splitlines())
    # Break multi-headlines into a line
    chunks = (phrase.strip() for line in lines for phrase in line.split("  "))
    # Drop blank lines
    text = '\n'.join(chunk for chunk in chunks if chunk)

    f_out = codecs.open(KEYWORDS + "_test_" + str(i+1) + ".txt", "w", "utf-8")
    f_out.write(text)

    time.sleep(10)
    next_page = driver.find_element_by_link_text(str(i+2))
    next_page.click()
    time.sleep(10)

time.sleep(10)
print(str(n) + " pages saved. Done!")

time.sleep(10)
driver.quit()
```

4. Get the latest ChromeDriver from sites.google.com/a/chromium.org/chromedriver/. Unzip the file in the same folder as the python code.
5. Run the code. Text files generated will be saved in the same folder. Depending on the IDE used, the final text file generated may only appear when the IDE is closed.