CSC 332 Programming Assignment #1

Due Wednesday, February 7 at 5pm

- You are welcome (and encouraged) to work in pairs
- Mind the academic integrity policy described in the syllabus
- Late assignments are not accepted (but you can earn partial credit for an incomplete assignment)

We talked in class about how modern webpages contain a multitude of embedded objects, including images, scripts, and compressed binary data. This programming assignment will make that abundantly clear to you!

Instructions:

You will need the Python libraries `requests` and `BeautifulSoup4` for this assignment. Before you begin, install them by typing this command into your Terminal/Command Prompt:

```
pip3 install requests bs4
```

Here are links to the official documentation for `requests` and `BeautifulSoup4`.

1. Using the libraries you just installed, write a Python program that requests a base HTML file and then requests each of the referenced objects on that page. (By "referenced" or embedded objects, I mean objects with URLs referred to in the `src` attribute of an HTML tag. URLs in the `href` attribute of an HTML tag need not be requested, since they are not embedded in the page.)
2. For at least three distinct pages on three distinct domains, calculate the following values:
    a. The total size, in bytes, of the base HTML page and all its embedded objects
    b. The percentage of objects on the page that fall into each Content-Type category
    c. The percentage of objects on the page that are hosted on a different domain than the base HTML page. (You may consider objects hosted on subdomains, for example, `static.stereogum.com`, to be hosted on the same domain.)
    d. The time, in seconds, that it takes to load the base HTML page and all its embedded objects sequentially
3. Use Wireshark or the Network tab in Google Chrome to find the page load time for each of the three pages in your web browser.
4. Compile a document containing:
    a. The URLs of the three pages you tested
    b. The four values described in step 2 for each of the three pages
    c. Browser page load times for each of the three pages

    d. A comparison between the load time for each page as calculated in step 2d and the load time from your browser, as shown in the Chrome Network tab or Wireshark. Are they similar or different? Why?
5. On Moodle, submit a zip file containing your Python source code and the document described in step 4. If you are working in pairs, only one partner needs to submit. (Make sure both of your names are on the document.)