Announcements

- Readings due today!

- Assignment 02:
  - due February 05

UCSD DSC96

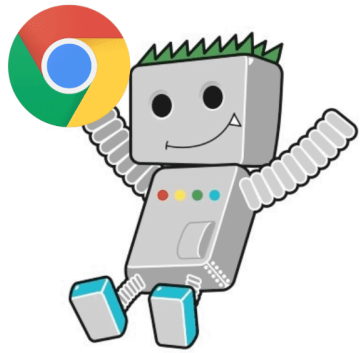
Web Scraping

Access large amount of data in an automated way

WHY? To analyze the large amount of data in the web, without manually reading (or copying) it.
E.g., Google is looking at all website to update Google search engine!

1. API: large websites, like Facebook, provide APIs to access their data in a structured way

2. Web scraping: if there is no API, you may need to scrape the website using your own program (we will see how today!)



The rules

1. Don't break anything. Many rapid requests to smaller sites can overload the host server.

2. Use a published API if possible - it is more robust and usually much easier!

3. Respect the policy published at robots.txt

4. Don't spoof your UserAgent (or try to trick the server into thinking you are a person)

5. Read the Terms of Service for the site and follow it.

1. Open the website of the class

2. Click on 04Text to download the latest ipynb files we are using today!

3. Complete the processing_text.ipynb notebook

4. Let’s work with web_scaping.ipynb !