Application Programming Interface (API) for Web-Scraping

Garrett Morrow

Northeastern University
NULab for Texts, Maps, and Networks

Learning Objectives

- Understand the definition and purpose of an API and web-scraping.
- Understand the importance of API documentation.

What is an API and what is web-scraping?

An API, or application programming interface, is a set of subroutine definitions, communication protocols, and tools for building software that ultimately allows applications to communicate with one another. An API may be for a web-based system, operating system, database system, computer hardware, or software library.

Web scraping is the process of extracting large amounts of data from an internet source and downloading the data to a local repository. The scraping process can be done manually, but is usually automated by using software because of the large amount of data typically involved.

API Documentation

- When using APIs for web-scraping, it is necessary to refer to the API documentation and a link is usually found on the API homepage.
- Why?
  - While the concepts remain roughly the same, APIs differ and the syntax for accessing data can be very different.
  - You will likely need an API key, and the links for registering for the key will be found in the documentation.
  - There may be other unaccounted for differences and API specifics that require a close understanding of API structure.

Popular APIs

- New York Times: https://developer.nytimes.com/
- Reddit: https://www.reddit.com/dev/api/
- IMDB: http://www.omdbapi.com/
- FBI: https://crime-data-explorer.fr.cloud.gov/api
  - Other Federal government APIs: https://api.data.gov/docs/
- Twitter: https://developer.twitter.com/en/docs.html

Contact and Resources

If you have any questions, contact me at:

Garrett Morrow
Digital Teaching Integration Research Fellow
Morrow.g@husky.neu.edu

Insert link for APIs on GitHub