Technology Day: Beautiful Soup

10/28/2020

Andrew & Karthik

What is Beautiful Soup?

- Beautiful Soup is a web scraping library for Python
- It allows for extracting information from static and some dynamic websites with little overhead or knowledge of the sites infrastructure
- Allows data collection where an API or other data plugin is not available

https://pypi.org/project/beautifulsoup4/

Application of Beautiful Soup in Our Project

- NYTimes provides ground truth data for college COVID cases however (at the time) it was only available as a webpage
- We built a scraper to extract the case data from the webpage and exported it into a CSV format for analysis
- The NYTimes college data is collected via a semi-manual process so using it for our project saves us time

Pros and Cons

- Pros:
  - Simple to use
  - Collects data without needing API access
- Cons:
  - Only works for static data, dynamic scripts need a rendering library to collect data
  - Can break if a webpage changes structure/format

Demo & Q&A

Demo Repo: github.com/andrewjohnston99/CS2803-TechDay-BS4