WE1S Chomp

Chomp is a set of Python tools designed to find and collect text from webpages on specified sites that contain search terms of interest. Its function is to build a useful corpus from born-digital sources focused on particular areas of research. Unlike other web scraping tools, Chomp is designed first and foremost to take a12 Sept. wide sweep--working at scale and across a variety of different platforms to gather material for topic modelling and other forms of broad statistical analysis and “distant reading.”

There are four key tools in the Chomp package: the Selenium browser interface, the Google Custom Search Engine (CSE) interface, the Web content scraper, and the Wordpress API interface. All of these are easily customizable and run through a set of Python notebooks in concert with other WE1S collection tools.

The Selenium browser interface is the means by which Chomp accesses the Web. Using purely programmatic Web tools like the Requests module can be faster, but, for some purposes, unreliable. Selenium processes the Web through an actual browser window, allowing JavaScript to run, ads to display, redirects to occur, etc. Since many of the pages that interest WE1S rely on these features to publish content, Selenium has proven to be a necessary addition to Chomp.

The Google CSE interface is how Chomp finds Web pages to scrape. Chomp can also function by way of the Google search page itself, but this requires periodic human intervention to solve CAPTCHAs. Using the CSE API instead limits us to a certain number of results per day, as per Google’s pricing structure, but has the advantage of making the process fully automated.

The Web content scraper gathers content from Web pages by making both basic and customizable assumptions about how content is typically arranged within a given page’s DOM, which it navigates by way of the BeautifulSoup module. By default, the scraper looks for anything within a `<p>` tag over 75 characters long:

```
for tag in [t for t in
            soup.find_all(tag_type)
            if len(t.text) > length]:
       content += " " + str(tag.text)
```

This setup has proven successful for our purposes, but users are encouraged to take a sample of their results and adjust the tag as necessary for their own projects.

The Wordpress API interface acts in place of the Web content scraper for sites that run the Wordpress CMS, and that allow back-end API access. This is easier and generally more thorough than finding pages through Google and does not contribute to the CSE query limit. Chomp pings each Website in its search list first to see whether or not it has an accessible Wordpress API in order to maximize this advantage.

Further Information:

* Selenium API

* Google CSE API

* Wordpress API

WE1S resources:

* Chomp on GitHub:
https://github.com/seangilleran/we1s_chomp

7 July 2020; rev. 12 Sept. 2020 (Sean Gilleran)