COMP 370 Homework 7 – Web Scraping

Assigned Oct 10, 2024
Due Oct 21, 2024 @ 11:59 PM

Conceptual Exercises

1. Why is the difference between found and designed data?
2. What are the two primary challenges that necessitate sampling when collecting data?
3. Why are website owners more likely to be upset about collecting data using scrapers than using APIs?

Technical Exercise

In this homework, we'll be using web scraping to collect the 5 trending stories from the Montreal Gazette. Our objective is to collect the title, publication date, author, and opening "blurb" (as seen in the screen capture here).

Jake Allen holds the fort again, Canadiens record second straight win

Canadiens 3, Sabres 1. "I'm just trying to get a little bit better every time right now," Allen says.

Herb Zurkowsky • Montreal Gazette

Published Oct 23, 2023 • Last updated 6 hours ago • 4 minute read

We want a script "collect_trending.py" that does all the work (i.e., we don't have to specify the trending articles one by one or in a list. The script goes, figures out which they are, and then grabs them off the website).

So to do this, you'll have to write a scraper for two different page templates.

- To get the trending stories (and links to them), you'll need to first scrape the homepage of Montreal Gazette (https://montrealgazette.com/category/news/)
- Then once you have links to the trending stories, you'll need to scrape the key information off the article page itself.

collect_trending.py is run as follows:

```
python collect_trending.py -o trending.json
```

Such that trending.json has the format:

```
[
    {
        "title": "article title",
        "publication_date": "date",
        "author": "author",
        "blurb": "blurb"
    },
    {
        … article info
    },
    …
]
```

For both page templates, use cache-ing to avoid overly taxing the Montreal Gazette website.

Submission Instructions

- Submit your conceptual answers in a file called conceptual_answers.md
- Submit all your code in a zip file code.zip (internal structure is up to you)