COMP 598 Homework 7 – Data Scraping

30 pts
Assigned Nov 5, 2020
Due Nov 13, 2020 @ 11:59 PM

This is an INDIVIDUAL Assignment – each student's work must be their own, each student completes this assignment, there are no teams for homework 7.

Non-standard (i.e., built-in) python libraries you can use:

- pandas
- requests
- BeautifulSoup

Task 1: Scraping relationships (10 pts)

In lecture, we began work on a system for scraping the whosdatedwho website. Here, you need to finish that system.

Write a script collect_relationships.py that collects the relationships for a set of celebrities provided in a JSON configuration file as follows:

```
python scripts/collect_relationships.py -c <config-file.json> -o <output_file.json>
```

where config-file.json contains a single JSON dictionary with the following structure (the exact path and list of celebrities can, obviously, change):

```
{
        "cache_dir": ".data/wdw_cache",
        "target_people": [ "robert-downey-jr", "justin-bieber" ]
}
```

Your script will then go and fetch the relationships for the target individuals. Note that the target people are indicated using the identifier that follows "/dating/". All pages visited MUST be cached in the cache directory specified – as described in the lecture. This means that, if run twice on the same config file, it will use data exclusively from the cache the second time.

The output format for the file is:

```
{
        "robert-downey-jr": [ "person-1", "person-2", "person-3" ],
        "justin-bieber": []
}
```

Where the identifiers in the list are the people the person had a relationship with. If the person has had no relationships, then they will have an empty list.

Task 2: Getting course information (20 pts)

Write a scraper to pull the courses off pages with URLs of the form: https://www.mcgill.ca/study/2020-2021/courses/search?page=X where X is a number. Your script ONLY has to read the courses off the page specified (it doesn't have to traverse to other pages). It will be run as follows:

```
python scripts/scrape_courses.py -c <caching_dir> <page#>
```

Your script must cache to the directory specified. The page# indiciates which URL will be loaded. The courses should be printed in CSV format to stdout with the following columns (header included):

CourseID, Course Name, # of credits

You should assume that all courses will be delivered with structure like this:

ACCT 626 Data Analytics in Accounting (1.5 credits)

Desautels Faculty of Management | Management | Graduate | Not Offered

Where “ACCT 626” is the CourseID, “Data Analytics in Accounting” is the course name, and “1.5” is the # of credits. If the course encountered does NOT have this structure, ignore it. (Note that the course # if the course ID can have letters in it as well, e.g., “ACCT 645D1”).

Submission Instructions

Your MyCourses submission must be a single zip file entiled HW7_<studentid>.zip. It should contain the following items:

- scripts/
  - collect_relationships.py – script for Task 1
  - scrape_courses.py – script for Task 2