Preparing for Homework 4

Create an empty repository on Github. Put your work (catalog-downloader script and helper functions) in `downloadcatalog.py` and your extracted catalog data in `catalog.csv`, and a department table in `department.csv`. Check your files in to git, push to github, and submit your repository to Gradescope.

The goals of this assignment are to give you practice with webscraping, as well as potential practice with regular expressions. You can use pandas for this assignment.

Problem 1: Webscraping

Write a Python script that will, starting from http://collegecatalog.uchicago.edu/, download the descriptions of all the classes for the entire college and parse the entire college catalog into a CSV file with at least six columns of relevant data. This whole collection operation should entail fewer than a hundred HTTP GET requests each time the script is run (but you can probably write most it in test mode where you only make two GET requests and produce the final output only for a single department's listings).

Your script should do the following:

- Not download anything that is not relevant to the data collection (images, external links, check out our channel on social media, send email to the webmaster),
- Not download anything outside of http://collegecatalog.uchicago.edu/,
- Not download anything more than once each time the script is run,
- Wait at least 3 seconds between queries,
- Construct the datastructure (to be saved) while the pages are being crawled, and
- Not contain more than a handful of hard-coded urls. Do not put "data" into your code, but put code that determines what links to pursue and which to ignore.

Your fields should include course number, description, terms offered, equivalent courses, prerequisites, instructors. Some of the classes won't have all the fields you are looking for, or might have extra fields that you hadn't planned on retaining.

Once you have your database of all the 2023-2024 class offerings, answer the following questions:

1. How many classes are there overall?
2. How many classes do you get if you put a fair attempt into de-duplicating them?
3. Which department offers the most (different) classes? Make a table, put it in a separate file.
4. Is there an apparent difference in the number of classes offered between Autumn, Winter, and Spring quarters?

Submission

This assignment is not well-suited for autograding. You should be uploading three things:

1. Your data in csv format,
2. Your python script, and

3. The table of classes offered by department
4. A pdf (or text file) containing the answers to the analysis questions.