Homework 5

This homework is designed to get you thinking about how to implement a web crawler.

Pick your favorite (small) blog. Using the code from class as a starting point, write a web crawler that starts at the root url of the blog and collects information about all of its pages. If the blog that you are crawling is too big, then come up with some reasonable constraint, for example only pages created in the past 2 years.

Your crawler should create a results file that stores the following information about each page in CSV format (BONUS: sort his file chronologically):

- `is_post`: a boolean value that is 1 if your crawler thinks that the page is a post.
- `publish_date`: time the article was created
- `author`: author name if available
- `url`
- `post_title`: The title of the post
- `comment_count`: Number of comments on for the post (this may be difficult)

The blog that you are crawling may have a sitemap, but do NOT use it for this exercise. While we will ignore the robots.txt file for this exercise only, make sure you abide by the other good citizenship rules (e.g. insert a small delay between requests).

A few other questions to think about as you get started:

- What if two pages link to each other, how can you resolve this cycle?
- What are features of the HTML that indicate that this is a post?
- How can I limit my crawl to only the domain of interest?