INFO370 Problem Set 3: Webscraping

July 4, 2021

Instructions

This problem set is about collecting real-world data, and learning about the legal and ethical aspects, and the messiness of the results.

Your task is to scrape recipes, in particular ingredients of the recipies. Your final product should be a dataframe with all the ingredients, and also an explanation about how well did it work.

You will use Beautifulsoup library, keep its documentation https://www.crummy.com/software/BeautifulSoup/bs4/doc/ nearby.

1. Explain what your code is doing. Here is an example of my code:

   ```
   ## store the original text before we strip links
   texts.append(LIclone.text.strip())
   ## Normally the ingredient line contains three things:
   ## 1. quantity (not in A), 2. unit (in A), 3. ingredient (in A)
   ## But sometimes there are no A links, so in that case we
   ## put everything to ingredient as we cannot split quantity and unit
   As = LIclone.find_all("a")
   if len(As) > 0:
       ingredient = As[-1].text.strip()
   else:
       ingredient = LIclone.text.strip()
   ```

   Note the extensive explanations in comments!

2. Your results will only count if accompanied with sufficiently clear explanatory text. Just plain output, with no explanation, will not count.

3. As the final submission, you should submit a) code (notebook); b) output (html or pdf); c) the scraped ingredients as a csv file.

1 Legal and Ethical Issues (30pt)

Before you can start scraping recipes, you have to find a webpage where webscraping is legal. In recent years it is getting more and more common that the sites explicitly ban it. For instance, allrecipies.com states in Terms of Use that:

> (e) you shall not use any manual or automated software, devices or other processes (including but not limited to spiders, robots, scrapers, crawlers, avatars, data mining tools or the like) to "scrape" or download data from the Services…

Some websites permit downloading for “personal non-commercial use”. Some websites stay silent about scraping. I think what you will do will be “personal non-commercial use” but I am not a lawyer...

Here are some recipe websites that do not ban scraping:

- What’s in the pan? `https://whatsinthepan.com/` seems not to mention scraping at all
- Feasting at home `https://www.feastingathome.com/` tells not to re-publish recepes, not even on social media, but does not mention scraping.
- You may find recipes on github too. Github “Acceptable Use Policies” state that

  > You may scrape the website for the following reasons:
  > - Researchers may scrape public, non-personal information from the Service for research purposes, only if any publications resulting from that research are open access.
  > - Archivists may scrape the Service for public data for archival purposes.

  My reading is the education is OK too but I am not a lawyer...
- Wiki Cookbook: `https://en.wikibooks.org/wiki/Cookbook:Recipes`

1. (8pt) Find a website that you consider scraping.
2. (11pt) Discuss the legality of scraping this site: does it mention data collection in its terms-of-service? Does `robots.txt` file allow/disallow what you want to do?
3. (11pt) Discuss what kind of steps do you take to lessen the burden to the provider. Consider:
   (a) Only downloading minimum amount necessary for developing and debugging your code, and dowloading the full amount just once. Note that this PS asks you to scrape *50* recipies, no more, no less. So you may stay with downloading only 60, just in case some of those are too complex to handle.
   (b) Adding a wait time (e.g. 1s) between successive page downloads
   (c) In case of wikimedia, I donated them money as a “thank you” for the open access.

2 Scrape the recipies (60pt)

Your coding task is to scrape the website and collect all the ingredients for at least 50 recipes.

1. Write the code to scrape (at least 50) recipes from the webpage you chose.
2. Pull out the list of ingredients, and split those into quantities, units, and ingredients as well as you can.

   Hint: sometimes you find unstructured ingredient lines. In that case there is little you can do with it. It is extremely hard to tell what is what in “a handful of sugar peas” or how much exactly is “a bunch” of finely chopped cilantro.
3. Store the ingredients from each recipe in a data frame. The data frame should contain the following variables:

   id recipe id (just a number)

   name recipe name

text the ingredient line as text as printed on the webpage, stripped of all html attributes

quantity quantity of the ingredient

unit unit of the quantity

ingredient the name/explanation of the ingredient

The data frame should be in long form, i.e. one line for each ingredient, and several lines for each recipe.

It should look something like

```
id      name                           text  quantity      unit       ingredient
...
23  Kal Kals  1/2 cup (120g) powdered sugar       1/2       cup   powdered sugar
23  Kal Kals   1/2 teaspoon vanilla extract       1/2  teaspoon  vanilla extract
...
31    S'more                 4 marshmallows         4                marshmallows
31    S'more              4 graham crackers         4             graham crackers
...
```

So the recipe #23, "Kal Kals", contains quantity "1/2" of unit "cup" of "powdered sugar" (and other things), and recipe #31 "S'more" contains quantity 4 of marshmallows (and other items).

Hint: You will quickly discover that it may be impossible to do this well. If this is the case then:

(a) Do your best

(b) Explain what are the issues you cannot easily solve.

4. Store your dataframe in a file so you don't have to repeat scraping.

3 What is more common: sweet or salty food? (10pt)

Now let's do some analysis based on the data you got. Are there more salty foods or sweet foods?

1. For each recipe, find if it contains salt, and if it contains sugar. Are there more recipes with salt or more with sugar?

   Hint: check out functions: `pd.Series.str.contains` and `np.any`.

2. But what about the absolute quantities? Do all these recipes contain more salt or more sugar? Add the salt and sugar quantities over all recipes and see which one wins.

   If you cannot do this, explain what are the problems!

Remember to upload:

- Your code (notebook)
- HTML
- The resulting ingredients file (this is specific for this problem set).

4 Finally...

Enough of coding. Now pick your favorite recipe and treat yourself :-) Feel free to add a picture of the result! Below is mine. I have to admit though that I was using a printed book, not scraping...

Sausage and jumbo shrimp pilaf

How much time did you spend on this PS?