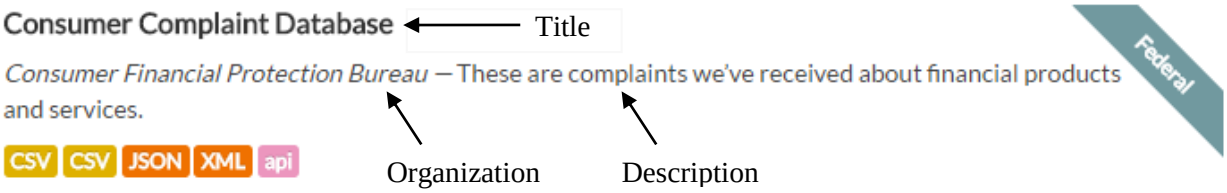
PIC 16 Final Exam Part I - Track B - Option 3

In this exam option, you will use Scrapy to help you download the titles, descriptions, provider organizations, and contact information for all the U.S. Government's open data available at `data.gov`. *Since this option requires so little original Python code, you are required to complete* one (your choice) *of the Part II options if you choose this for Part I. If you submit only Part I Track B Option 3 without any Part II options, the maximum score is 70%.*

Here are some steps to help you divide the work into smaller chunks. For all steps but the first, you should generate a file `datasets.csv` containing the scraped data with appropriate column titles like "title", "description", "organization", and "contact". I do not care about the order of rows or columns of scraped data. **To submit your solution online, simply compress (`.zip`) the entire Scrapy project folder including the most complete version of `datasets.csv` inside.**

1. First, create a Scrapy project and a new spider with filename `datagov.py`, class name `DataGovSpider`, and name field `"datagov"`. Your spider will begin at the URL `http://catalog.data.gov/dataset`, which contains "entries" that look like:



2. Using features of your browser, figure out a candidate CSS expression for extracting the titles of the datasets (e.g. "Consumer Complaint Database") from the first page of entries. Test it using the Scrapy shell, and finally modify your spider as needed to extract all these titles. `crawl` your spider to produce `datagov.csv`, which should contain a column labeled "title" followed by 20 dataset titles.
3. Modify your spider to extract both the titles and descriptions from the first page of entries when it crawls. *Note: by modify, I mean you don't need to submit a copy of your spider from the previous step to get credit for it; completing this step will get you credit for previous steps. However, you might wish to save a working copy of your spider before modifying it for your own sake!*
4. Modify your spider to extract the titles, descriptions, and organizations from the first page.
5. When you open up your `.csv`, you may notice that the organization names contain something strange at the end. Fix this. (Don't worry about other strange things like "These are complaints weâ€™ve received…". I trust you could fix all these with sufficient time.)
6. When you open up your `.csv`, you may notice that some descriptions include unnecessary whitespace at the beginning and/or end. Remove it.
7. Modify your spider to extract the same information from the first *N* pages of entries. Let *N* be two for the purpose of creating your `.csv`, but your code should work for any positive integer *N* (up to the total number of pages) by changing the value of a single integer variable (named `N`).
8. Use the "meta trick" to also collect contact information from *within* an entry. That is, clicking on the title of an entry will lead you to a page with more information. Find the contact information at the left of the website (which looks like the image at the left of this page), and extract *whatever text* is shown there, which may be a name or email address. Include it in the same row as the other data for that entry in your spreadsheet.