Newegg Scraper

Tate Haga

Commands

The scraper is run in any terminal with python3 support. To run, use:

python3 scraper.py *url* *output_filename*

Make sure for your output_filename you don’t include a file extension, since it is automatically included with an epoch timestamp

URL formatting

Newegg is very particular about how the URL to display data, so it is important to have a correct link:

https://www.newegg.com/Computer-Cases/SubCategory/ID-7?name=Computer%2DCases&PageSize=96&Order=REVIEWS is NOT CORRECT

https://www.newegg.com/Computer-Cases/SubCategory/ID-7 is NOT CORRECT

https://www.newegg.com/Computer-Cases/SubCategory/ID-7/Page-1?name=Computer%2DCases&PageSize=96&Order=REVIEWS is CORRECT

Make sure the page number is at one and both page and Order are in the URL

Functions

- getPagesfromCategory
  - takes a Newegg subcategory link, and generates list of all pages in that subcategory until a page with some "n" number of reviews is reached, in this case n=2
- getCategoryAndUrls
  - takes an individual page with multiple items on it (the output of getPagesfromCategory) and returns the subcategory of items as well as a list of links to each item on the page
- newReviews
  - takes the url to any given item and returns the item id and a dictionary object with the information about the object, including all its reviews Note: During development, Newegg changed the way you viewed reviews from being a part of the static html to a dynamic javascript loading, so for most pages, you only are able to access the first 8 or so reviews on a page, sorry. Also, for some reason there are lots of items on Newegg that have reviews but these reviews aren't available.

Future Improvements

Most of the code is well documented, but for future reference if something breaks catastrophically, Newegg has most likely completely changed the formatted something on their end of the project, which happened to me in the middle of my work. The codebase should hopefully be compartmentalized enough to minimize the number of changes necessary. Most likely the page with review data has been altered, in which case you will have to change the tags that BeautifulSoup is searching for, most of the other formatting should still work.

In the future, the most important functionality to increase review count would be to use something like Selenium to dynamically access the reviews with javascript

Statistics

Number of items: 4222

Number of items with reviews: 3148

Number of reviews: 21765

Number of reviewers (non-anonymous): 12954

Statistics

Most common Reviewers by ID

AuthorID	Number of Reviews
Anonymous	5424
bE9qdmdFQ1ZzMjBnYlhTTFhJQ0lRWWhMSU1tVXp1YlE=	28
YXlyMG5lL2pGSzhONlA3Umt0MExBaVp2OEZvdG1qSko=	28
MlVlY2ozVHZEbmk1ZTV3RDZkdFlteFF6UFlHL3JvQ2s=	28
Zm5BcndKMUhhVFQ3V2U4NG9nS01ZSHBpdU1oV3FpOGM=	27
aDdZZVpneE1SNzhnRi9xNTJQNkpiNXFOa0JPZjdoYng=	27
cFVGZ0d3U1RjWFZacld2SWJlbTZ1VzBzOWoxbmdNOFl=	27
RW1WeTFZV25vMFQ3eGhiM0JvTWdKcEVwaFQzbkhEOGY=	27
MGJmbVkrdnQ0ZTJNN1ROM2kyRjdYanBrb0tvUzNiOEw=	27
WGp4UDlFQ3BCVXhTbk5NY1RzbHBKcnZBakFVQzBmVUw=	16
aDJFbGJPc1ltaDVuK0Mra25WNVJDZjlSRk1KdVN6eHc=	15
cjJjWmNsZytoVVlnbFI5RHhSanJQa3lpS214N3RmcFc=	15
YjNwaFgzeVY5aWhtdElxaU1uVVBhdz09	15
MWw3VkZ3RXlOVmF0N09zZE1JdUFZakd3UnB4eTR4MFc=	15
dTlrM2lxQjdHdk5XWGJhZWxLT0d3OVp2ODNDekR0VWg=	15