URL Availability

Web Science

10 points

Work in teams of two to create a Python script called createGraph.py. The script should produce a graph with the matplotlib library that plots the percent of URLs accessible on the web and from archive.org at each date.

Discover URLs

Find at least 20 URLs to probe. The URLs should be from:

1. The same collection. Example: You could gather 20 URLs that were cited in the latest issue of a magazine, URLs that are trending in reddit, or URLs from Twitter using the hashtag #HurricaneDorian.
2. Randomly sampled from the web. Example: Issue randomly selected search terms to Google, jump to a random result page, and grab one of the URLs at random.

Use the script we created in class that samples the URLs on the live web and archive.org to produce a data file to use as input for createGraph.py.

Input

The script should read input from STDIN that is formatted like so:

URL
Date (YYYY-MM-DD)
Status (200 or 404)
Archived (yes or no)

Example:

```
http://www.harding.edu/fmccown/
2019-09-01
200
yes
https://en.wikipedia.org/wiki/Archive.today
2019-09-01
200
yes
https://example.org/blah
2019-09-01
404
no
```

If the above data existed in a file called input.txt, the script could read the input using the < operator:

```
$ createGraph.py < input.txt
```

Ideally, the input should contain the same URLs repeatedly accessed over multiple weeks. But for now you may put some test data into the file in order to verify your createGraph.py script is working correctly.

Graph

Your script should produce a line graph from the input using the matplotlib library.

- The y-axis should range from 0 to 100 percent.
- The x-axis should range from the earliest to the latest dates found in the input.
- Two lines should be plotted:
    - The percent of URLs on that day that are accessible (200)
    - The percent of URLs on that day that are archived (yes)

For help on using matplotlib, see Graph Plotting in Python.
https://www.geeksforgeeks.org/graph-plotting-in-python-set-1/

Submit

Only one person from the team should submit createGraph.py to Canvas. Put in comments at the top of the program:

1. Both teammates' names
2. Source or methodology used to collect the URLs