Professionalism & Reproducibility

Homework #1

The goal of this assignment is to construct, analyze, and publish a dataset of monthly article traffic for a select set of pages from English Wikipedia from July 1, 2015 through September 30, 2024. Your notebook(s) and your data files will be uploaded to a repository of your choosing. You will submit a link to your repository to enable grading of this assignment. The purpose of the assignment is to develop and follow best practices for open scientific research as exemplified by your repository.

Step 0: Read About Reproducibility

Review the chapters "Assessing Reproducibility" and "The Basic Reproducible Workflow Template" from The Practice of Reproducible Research.

Rokem, Marwick, and Staneva. Assessing Reproducibility in Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Oakland, CA: University of California Press.

Kitzes. The Basic Reproducible Workflow Template in Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Oakland, CA: University of California Press.

Step 1: Data Acquisition

In order to measure article traffic from 2015-2024, you will need to collect data from the Wikimedia Analytics API. The Pageviews API call (documentation) provides access to desktop, mobile web, and mobile app traffic data starting from July 2015 through the previous complete month.

To get you started, you can refer to this example notebook that contains sample code for the API call. This sample code is licensed CC-BY. You can reuse any of the sample code, but make sure you follow the license requirements.

You will be collecting counts of pageviews using a specified subset of Wikipedia article pages. This is a subset of the English Wikipedia that represents a large number of articles related to rare diseases. This list of pages was collected by using a database of rare diseases maintained by the National Organization for Rare Diseases (NORD) and matching them to Wikipedia articles that are either about a rare disease or have a section that mentions a rare disease.

You will use the same article data to generate several related data sets. All of the resulting data sets should be time series of monthly activity. For all of the data sets we are only

interested in actual **`user`** pageview requests. The three resulting datasets should be saved as JSON files ordered using article titles as a key for the resulting time series data for that article. You should store each monthly record data as returned from the API, with the exception of removing the 'access' field as it is misleading for mobile and cumulative files.

You will produce three files as follows:

1. **Monthly mobile access** - The API separates mobile access types into two separate requests, you will need to sum these to make one count for all mobile pageviews. You should store the mobile access data in a file called:

   rare-disease_monthly_mobile_<startYYYYMM>-<endYYYYMM>.json

2. **Monthly desktop access** - Monthly desktop page traffic is based on one single request. You should store the desktop access data in a file called:

   rare-disease_monthly_desktop_<startYYYYMM>-<endYYYYMM>.json

3. **Monthly cumulative** - Monthly cumulative data is the sum of all mobile, and all desktop traffic per article. You should store the monthly cumulative data in a file called:

   rare-disease_monthly_cumulative_<startYYYYMM>-<endYYYYMM>.json

For all of the files the <startYYYYMM> and <endYYYYMM> represent the starting and ending year and month as integer text strings.

Step 2: Analysis

You will conduct a very basic visual analysis. The analysis for this homework is to graph specific subsets of the data as a timeseries. You will produce three different graphs.

Maximum Average and Minimum Average - The first graph should contain time series for the articles that have the highest average page requests and the lowest average page requests for desktop access and mobile access over the entire time series. Your graph should have four lines (max desktop, min desktop, max mobile, min mobile).

Top 10 Peak Page Views - The second graph should contain time series for the top 10 article pages by largest (peak) page views over the entire time series by access type. You first find the month for each article that contains the highest (peak) page views, and then order the articles by these peak values. Your graph should contain the top 10 for desktop and top 10 for mobile access (20 lines).

Fewest Months of Data - The third graph should show pages that have the fewest months of available data. These will likely be relatively short time series, some may only have one

month of data. Your graph should show the 10 articles with the fewest months of data for desktop access and the 10 articles with the fewest months of data for mobile access.

In order to complete the analysis correctly and receive full credit, your graph will need to be the right scale to view the data; all units, axes, and values should be clearly labeled. Your graph should possess a legend and a title. You must generate a .png or .jpeg formatted image of your final graph.

You should visualize the data using tools or libraries within your notebook, rather than using an external application to facilitate reproducibility.

Step 3: Prepare Documentation

Follow best practices for documenting your project, as outlined in "Assessing Reproducibility" and "The Basic Reproducible Workflow Template" from The Practice of Reproducible Research (e.g., the readings from Step 0)

Your documentation should be in your notebook(s), a README file, and a LICENSE file.

At minimum, your Notebook(s) should:

- Provide a short, clear description of every step in the acquisition, processing, and analysis of your data in full Markdown sentences (not just inline comments or docstrings)

At minimum, your README file, in .txt or .md format, should:

- Describe the goal of the project.
- List the license of the source data and a link to the Wikimedia Foundation terms of use. You should probably provide a very brief summary as to how the ToU applies to the dataset you have created.
- Link to all relevant API documentation
- Clearly name any intermediary data files and any final output files that your code creates.
- For any files that your code creates you should provide a data schema and brief description
- List any known issues or special considerations with the data that would be useful for another researcher to know.

Lastly, you should create a LICENSE file that contains an MIT LICENSE for your code.

Step 4: Prepare and Submit your Repository

You will complete this homework by submitting a link to a repository. For the purposes of this course a repository is simply a collection of files that can be accessed by the teaching staff (TAs and Instructor). You can use GitHub if you like, or you can create a shared folder in Google Drive, or even a Dropbox folder. The main requirement is that you have all of the required files in the repository and that you **SHARE IT WITH THE INSTRUCTIONAL STAFF**. Sorry about shouting, but you would be surprised how many students do all the work for the homework and then forget to share it with us. If we can't see it, we can't give you credit for your work.

1. Create a repository folder named `data-512-homework_1`
2. Copy your notebook(s) into the folder.
3. Copy the JSON data files into the folder.
4. Make screen shots, PNG or JPEG, of your visualizations, from your notebook(s), and copy those into the folder
5. Complete and add your README file in .txt or .md format and LICENSE file.
6. Set permissions on your repository folder to share it with both TAs and the Instructor. You should share it with the email addresses they listed on the syllabus.
7. Submit the link to your repository through the Homework 1 submission link on Canvas

Required Deliverables

A repository folder called `data-512-homework_1` that contains the following files:

1. The three JSON data files from Step 1 following the specified naming convention and format specifications.
2. Your notebook(s) clearly named to indicate what they do, containing code as well as information necessary to understand each processing step.
3. A README file in .txt or .md format that contains information to reproduce the analysis, including data descriptions, attributions and provenance information, and descriptions of all relevant resources and documentation (inside and outside the repo) and hyperlinks to those resources.
4. A LICENSE file that contains an MIT LICENSE for your code.
5. Image files of your analysis graphs in either PNG or JPEG format.