Violence in Brazil Dataset

Expanding Capacity for Newspaper Scraping

Research Objective

- To scrape online news articles from different Brazilian media websites and determine whether these articles contained violence/police-related incidents.
- Geocode and time-stamp these incidents into a dataset.

My main focus: attempting to expand the number of articles we could scrape at once

- 20,000+ articles

Research Process

- 1st time scraping a website - learning how to scrape a dynamic site
  - requests-HTML
- How to get a list of articles
- Accessing article content and cleaning the output
- How to display the output
  - Using Google Sheets API
  - Result: output file and csv file (import to Google Sheets/Excel)
    - Run 14 (100) and Run 17 (30)

Research Process

- Biggest issue I ran into: runtime and memory use/accessing old articles - running the script overnight
  - Using a server
  - Run 12 (10,000) and Run 13 (500) -> 1 month of data

Next Steps

- Reducing runtime and improving the script
  - Cleaning the script to more easily determine file names (more readable)
  - Making the script more efficient and reducing looping
  - Using a server
- Adapting the script
  - Test cases - limitation of “Most Recent” articles
  - Using the script with other websites
  - Initializing the starting page