Project Deliverable 1

Sufficient data should have been collected to perform a preliminary analysis of the data and attempt to answer one question relevant to your project proposal which you will submit as a pull request. If data has already been collected for your project you must answer two questions.

Checklist

- Collect and pre-process a preliminary batch of data
- Perform a preliminary analysis of the data
- ~~Answer one key question~~
- Refine project scope and list of limitations with data and potential risks of achieving project goal
- Submit a PR with the above report and modifications to original proposal

Deliverable 1

For this deliverable, we completed step one of our suggested steps (refer back to Deliver 0) and created a web scraper for the 10 year database of Boston police disciplinary action from the Boston Globe. The features of this dataset includes: type of misconduct, rank, race/ethnicity, name, year, unique case ID, outcome of the investigation.

For the web scraper, which is found in our repository under "webscraper.ipynb", we used the Selenium webdriver package in Python, as well as the webdriver_manager.chrome package. These worked together to create an automated web browser that is able to get the URL of the Boston Globe database from which we want to collect the data, and extract the information necessary based on the features we are interested in analyzing. The data collected by the webscaper can be found as a comma-separated value file in our repository titled BostonPoliceInternalAffairs.csv.

After discussion with our PM (Gowtham Asokan), we agreed that at this point, we are unable to answer the key questions yet as the project is data collection heavy. For our next steps, we will complete step 2 and 3 of the suggested steps (from deliverable 0), which includes merging datasets and overlaying lead police blacklist officers into the merged dataset.

(email from PM)

Gowtham Asokan <gashokan@bu.edu>   Feb 28, 2021, 8:54 PM (17 hours ago)
to Tina, Shelli, me, Daniel,, Rachel,

Hey Tina,

Submit the scraper as a pull request to your GitHub project folder! I've also attached a pasted checklist for Deliverable 1:

Sufficient data should have been collected to perform a preliminary analysis of the data and attempt to answer one question relevant to your project proposal which you will submit as a pull request. If data has already been collected for your project you must answer two questions.

Checklist

1. Collect and pre-process a preliminary batch of data
2. Perform a preliminary analysis of the data
3. ~~Answer one key question~~
4. Refine project scope and list of limitations with data and potential risks of achieving project goal
5. Submit a PR with the above report and modifications to original proposal

I've crossed out answering one key question as that is not possible for this project for the first step. Additionally, how did you bypass the Globe's security and bot measures if any? Have you also started collecting the database?

Best regards,

...