Executive Summary – SToPA Group 3, 2022 Datathon4Justice

Group members:

- Manuchehr Aminian
- Clare Jones
- Kath Landgren

Our group's work was split between developing additional tools for Williamstown dispatch data, and preliminary analysis for Rochester files (provided as a collection of pdf files for 2022). With Rochester data, Manuch built our understanding of the provided data and narrowed focus to establishing an OCR pipeline to automatically handle "Genl101A" forms, which seem to be given in all cases. With the Williamstown data set, Clare and Kath built tools to categorize records by "shift number" and geographic region and saw some differences along some dimensions with a preliminary analysis on these factors.

Motivated by questions from local activists in Williamstown, we have considered comparing the data from the historically black neighborhood White Oaks to the rest of the logs, looking to see if any patterns emerge. There seem to be some differences in call reason and action taken. Further investigation is needed. Notably, looking at the rolling 7-day averages seems to suggest that the calls in White Oaks did not have as steep a decline once the COVID-19 pandemic began, but the rest of Williamstown did have a steep decline. This might be due to a number of factors, and further investigation is needed. We have created a Python module with the functions we have developed over the weekend, ready to be pulled in to the SToPA repository and made available for use in further analyses. Here are some of our general findings:

- We found that there were over 100 entries that had the dispatch time reported as being before the call time. This could be a result of many things (at least in one instance, an OCR issue, 19-1028, page 145 of Logs2019.pdf), and is something important to keep in mind when looking at the results of the data.
- Also, several call times and many dispatch times were not recorded in the data that we had, so that needs to be kept in mind while analyzing the data.
- The logs from the White Oaks neighborhood constitute about 11% of total logs.
- Differences in the rolling averages:
    - The total number of logs declined sharply at the start of the COVID-19 pandemic. The number of logs corresponding to White Oaks does not seem to show as steep a decline.
- Differences in call reason:
    - The logs corresponding to White Oaks are more likely to list building checks (40% of all WO logs, 32% of all other logs)
    - The logs corresponding to White Oaks are more likely to list motor vehicle stops (22% of all WO logs, 12% of all other logs)
- Differences in call action:
    - The logs corresponding to White Oaks are more likely to list building checks (39% of all WO logs, 32% of all other logs)
    - The logs corresponding to White Oaks are less likely to list "services rendered" (22% of all WO logs, 31% of all other logs)

See our Jupyter notebook/modules/pdf for detailed information and figures.

The Rochester data provided during the Datathon came in several pieces; one, a short document describing the big-picture goals of Empire Justice, an Excel file summarizing some traffic stops in 2020 (not provided), and a zip file of approximately 89 pdf files. Cases are distinguished by a UTT ("Universal Traffic Ticket") number, and all cases have a so-called "Genl101A" form; amounting to 49 unique cases in this data set. We found that, while the pdf files here have text embedded, the so-called "read order" of the pdf files makes tools which automatically extract text virtually worthless; hence, we needed to revert to using OCR (optical character recognition) followed by ad-hoc parsing rules to extract information. Given the complexity of some of the other forms, we only built up a prototype for the Genl101A forms. While we do not have answers to the original questions, here are some basic facts and lessons learned from the process:

- On the Rochester data set itself…
    - There are 90 pdf files overall;
    - The data set consists of 49 cases, once grouped by UTT number, for 2022;
    - Often officer names are abbreviated, and on inspection it isn't clear if the "first initial, last name" included in the form matches the signature. But this is probably just an artifact of the few forms inspected by eye and probably not a serious issue.
    - No conclusions on the original question prompts. But we'll make the data scientist excuse of "90% cleaning the data, 10% analysis" and claim the analysis should be straightforward with a few more hours of intensive work after the Datathon.
    - The forms themselves include the city name, and suggest the forms are uniform across New York State, which suggests follow-up analyses within that state can be done easily once tools for Rochester are polished.
- The code used to scan the Genl101A forms will be uploaded as a branch to the SToPA Github repository, for the research lab members to discuss the best way to adapt the code base.
- The existing OCR pipeline in the SToPA repository is highly tuned to processing those logs – and those logs have some nuances to them that makes code hard to generalize. However, trimming down the code shows the existing pipeline.
- The *scanning* aspect of the pipeline is fairly straightforward (converting the pdf to a dataframe which describes bounding boxes for pieces of text. What is likely highly case-specific is the follow-up parsing of this dataframe; often requiring staring side-by-side at the dataframe output and the original pdf file, to engineer hard rules to extract the information. The UTT files in particular – while not attempted – look intimidating to parse using the approach used in Williamstown or the Rochester Genl101A forms.