Deliverable 1 Preliminary Analysis

WGBH - Reliability of Informant Cases
Project Members: Chen Xin, Xiang Liu, Qitong Wang, Changhao Liang

1.Question: Miscellaneous data source

We are provided with 2 data files: cases.json and cases_appeal.json
They are collected by the previous team on this “Informant” project, which is posted in the following repository:
https://github.com/BU-Spark/Data-Science-Fall-2019/tree/master/benchmarks_team_1
We are not sure how the previous team divided the two files. We do not have an explicit explanation of the data source, and it is unclear how exactly the data extracted.

2.Question: Work we did and potential improvement:

The cases in our data source were attributed into 2 classes as “Criminal”’ and “Civil”. Our project extracts all criminal cases into a filtered data source. Then we search for keywords “CI” and “informant” in the criminal cases. 97 cases satisfied our conditions. Our result is the same as the other team. Next step, we plan to expand our data in 2019: our data is from 2008-2018, we can expand our data to 2019.

3.Question: We need reliable data resources:

Our first analysis is cases from MA, but the client asked us to also collect the data in NH and Rhode Island. We think this could be difficult to accomplish for the following reasons:

1) The data from NH and Rhode on the website is in different format from that of MA, MA’s data are html format, which can be easily grabbed. But, the NH and Rhode’s data are pdf format, which are really hard to be extracted text to do project of next step.

2) The cases in MA have three keys: ‘cases’, ‘headnote’, ‘text’. Data from NH and Rhode Island have different headnote structure from that of MA. In addition, there is no clear definition to separate the content into civil or criminal cases.