Variable construction

Besides topics, we found that agency, location, residency requirement, length of description of job related requirements and preference as well as reading score can actually affect the salary.

Agency

In the original dataset we have 39 agencies, which indicates the 39 departments posting the hiring posts. In order to give our final regression more degrees of freedom and provide us with more specific results regarding the relationship between industry and salary, we bin the agency into 6 categories: Finance, Infrastructure, Social Service, Law, Security and IT.

Location

We matched the incomplete location data with NYC address database to get the complete address for each posts and categorize them as Manhattan, Bronx, Brooklyn, Queens and State Island. With the 200 unmatched address which due to the unstructured dataset, we manually categorize them into these 5 bureau.

Residency Requirement

For residency requirement we have got 35 different phrases in our original dataset describing whether a New York City Residency is needed or not. We used regular expression to convert those 35 phrases in to 2 categories, residency required or not.

Length of Text

In each posts we’ve got different length of sentences describing the jobs, requirements as well as preferences. We think the length of the text somehow affects the salary of the job, thus we make the length of this part as a numeric variable, which indicates the length (number of words used) of the “description” for each post.