STA242 Project Proposal

Shuxin Li
Mingyu Tian
Yilun Yang

1. In this project, we want to write an R package on some statistical methods from recently published papers. Specifically, we are interested in some topics, they are:

- Text mining and its application on pattern recognition, gene recognition and so on.
- Data visualization for a specific statistical model.

2.

Methods:

As we said, we want to write a package on methods from some papers, so we can refer to these papers for more details.

Techniques:

Basically, the first topic is about machine learning, we may refer to some great references like *Pattern Recognition and Machine Learning* by *Christopher M. Bishop* and so on to find more techniques that may help us.

The second topic is about plotting thus we may refer to how ggplot2 works and we could also perform our work based on ggplot just like a very famous map drawing package --- ggmap.

Software:

We would like to use python for web data acquisition and R for further data analysis and write the package.

Computational approach:

We should be familiar with how to use R to write package as efficient and robust as possible. We also need to know some machine learning algorithms and their implementation.

3.

We would like to get the data we need from three main resources: Web, Papers, and the data we simulated (or made up just for validation) in our package.

4.

The first problem here is that it is not easy to find a good question. Writing a package of several functions is not much difficult, but finding a meaningful question is a big issue. Another problem is that we should think about our question thoroughly in order to avoid missing something. For example: Is this problem too hard or too impractical to be realized? Is the robustness and efficiency promised in our package? Have we considered all the situations and functions we should implement in the package and so on?