Yin Chenxi

19089951D

I’m the leader of this project.

First, I conducted an analysis of tasks and a reasonable distribution of work.

I'm responsible for testing with multiple base models first and then choosing the ones that work. I used Decision Tree, Random Forest, SVC, K-NN, and Naive Bayesian were used for choosing the model

After that, I was mainly responsible for the KNN model.

I have completed reports and video recordings related to my part.

The first problem our entire team had was choosing a model, and our opinions were divided. I want to choose a model that is a big improvement over the base model. This is how we can show our results. My teammates wanted to choose a model that would perform well initially.

So, I made demos of several models to calculate accuracy and silhouette coefficients. The final solution is that we have to implement both models and see the final result. In the end we did it and we worked well together.

There was also a problem with our team's video recording. I am in mainland China and my teammates are in Pakistan. We tried to use MSteam to record video, always disconnected and couldn't get to the next page of PPT.

So we chose to let one person operate the PPT, and record the PPT in the morning when few people use the Internet.

The problem I have encountered is that MapReduce cannot be implemented. I want to use this method to help KNN filter out some features and assist in assigning labels. But I fail repeatedly. The final solution is to combine "word cloud" and "bag of word" using TF-IDF model

I also ran into problems when I was doing data visualization, I'm not very familiar with the new package. It took me quite a while to make a data visualization.

Once again I realized the importance of cleaning the data first. The process of observing the data was too rough. I learned from other people's methods to know how many ways to deal with invalid characters when doing text analysis.

I learned to use multiple models to complement each other. I try MapReduce and hierarchical clustering to assist KNN. Although the final effect is not good, but let me experience the actual application