Report

1. Dataset:

The data size is 32736 and there are 15 features for each data. We are going to take four features, name, number, rating, add_time to implement multi-scale event detection.

2. Implementation:

To implement multi-scale event detection, we take the name label of the data as tweet, the number and rating of the data as the location of the data happen, and the add_time as the time of the data happen.
I use the Chinese word segmentation module on python called Jieba to cut the name of the data. With this, I can build the dictionary which contain all the words occur in the dataset.
Based on the model of multi-scale event detection, we can get the similarity of every pair tweet with the similarity of the pattern of the common words shared by the pair tweet. Last, the output is clusters that correspond to events of different spatial and temporal scales.

3. Result

Because the frequent words are limited to occurring larger than or equal to 10, the output cluster should contain at least 10 tweets. As the result, there are 594 clusters that represent different events. It shows that every cluster has different size which is computed by the difference of the maximum and minimum value of the label “number” and “rating”. From the label “add_time”, we can see the different span of the time. Moreover, the tweet which contain more frequent words in one cluster can be taken as the overall representation of the cluster(event).