Yang Zhang 11529139

Beiyu Lin 11387703

Yang Zheng 11478827

**Document Classifier**

**Overview**

For the class project, our group want to develop a document classifier that can classify its topic from its content and rank labels by specified priority criterion.

**Motivation**

Nowadays, the amount of information in the internet is enormous. However, not all information is useful in any time. For example, people read news that those topics they are interested in and will often skip those content that they are not familiar with. Then a question come to us, why couldn’t we just filter out those content we are not interested in. One solution to the above question is document classifier.

**Methodology**

There are generally two ways for document classifying. One is to apply supervised learning method (such as Naïve Bayes), while the other one is to apply unsupervised learning (clustering is one way to do it). Before make the decision, we will look into both method and compare pros and cons of them.

**Expected Result**

Training a document classifier that can more strictly classify out the test with higher topic priority. And we expect that the document classifier can classify most of file based on the content, for example, applying the classifier to a homework cluster, it can classify computer science homework from other homework.