**README**

**Minzhe Zhang mxz163730**

**Yongqi Yue yxy175330**

1. Code is written in python 3. Following package is need to do the analysis:

* argparse
* os
* nltk
* re
* json
* math
* collections

1. The program has two python script: ExtractFeatures.py and NaiveBayesClassifier.py

* ExtractFeatures.py: Process the raw text files to get features. The temporary feature result is store in the json files which are the input of second python script.

**$ python ExtractFeatures.py --train data/20news-bydate-train/ --test data/20news-bydate-test/**

* NaiveBayesClassifier.py: Train naïve Bayes model to classify text.

**$ python NaiveBayesClassifier.py --train train\_data\_json.txt --test test\_data\_json.txt**

1. To simplify the model, I use the most frequent 3000 words in all texts to do construct feature matrix and train naïve Bayes model. The final prediction accuracy in test dataset is 73%.

