数据说明

data文件夹中式所有的数据

raw\_data是爬虫爬下来的数据，mozilla.py, eclipse.py,gcc.py分别爬mozilla.json,eclipse.json,gcc.json。爬数据的时候就已经将数字去掉了

so\_data是stack\_overflow的数据，我已经转成了json格式。json格式就是对应python中的一个字典，用json.loads和json.dumps来加载文件和存文件

nohttp\_data文件夹中是删了http的数据。使用delete\_http.py删除Http

preprocessed\_data文件夹中是执行了所有预处理的文本（不包括去数字和http）。preprocess.py使用nohttp\_data中的数据进行初始化，分词，去停用词，词干化

full\_data是填充stack\_overflow后的数据。在根目录的knn\_filter.py将so的文本按照相似度过滤，然后存入full\_data中，然后手动吧preprocessed\_data拷贝到full\_data中。

wordEmbedding为词向量

默认使用word2Vec方法，对应word2Vec.py

测试词向量为lsi方法，亦即svd方法,对应svd.py

knn文件夹中存knn的结果

根目录的文件

knn.py knn算法

knn\_filter.py knn选择so评论存入data/full\_data/

lstm.py 长短时记忆网络

NaiveBayes.py 朴素贝叶斯

percent.py 计算每个标签的百分比

utils.py 一些辅助函数

朴素贝叶斯的做法：选取所有在文本中出现频率超过5的词作为特征，生成每个位置都是0或者1的特征向量。然后使用朴素贝叶斯方法。