实验报告:

文件夹目录说明:
（1）关于数据（开始全部的数据是放在 data 目录下,但是一直没法将其传到 gitup 上（原因没找到，明明是按照网上的步骤一步步操作的可还是交不上去），所以项目原本是在 DataMining 下有个 data 文件夹存放数据
（2）vector:导入数据->数据预处理（大小写转化，去除特殊字符，去除停用词，词干提取等操作)->创建词典->创建 0-1 型向量
（3）knn:将数据利用 sklearn.model_selection 中的 train_test_split 函数划分为测试训练集和测试训练集->计算余弦相似度->>>计算准确率
（4）out 文件夹下存放数据处理过程中的结果（其中 Input_data,labels,dict,vsm 是预处理阶段产生的，dictionary,text_x,text_y 是 knn 阶段产生的）

心得体会:
由于之前没使用过 gitup，没学过机器学习和 python 语言，此次作业虽然过程很艰难但是收获也很大。

knn 准确率截图:
由于程序跑起来太慢，所以只截取了两个 k 不同时的取值，准确率是在上升的

```
18827
18828
15062 15062 3766 3766
计算余弦距离中
3766
When k is 2 ,the succession of knn is :0.014604354753053638
```

```
18828
15062 15062 3766 3766
计算余弦距离中
3766
When k is 3 ,the succession of knn is :0.016994158258098777
```