# Homework2 报告

霍瑜 201605130119

## 问题一：分词？

直接使用第一次作业的分词结果，即 articles3.txt

## 问题二：开始实验

### 1. 拉普拉斯平滑因子的选区？

尝试了几个参数，发现效果都是在 0.84 左右，最好效果不超过 0.85。

```
#测试结果
#alpha = 1   正确：4705 总的：5642 正确率：0.8339241403757532
#alpha = 1.5 正确：4667 总的：5642 正确率：0.8271889400921659
#alpha = 0.5 正确：4740 总的：5642 正确率：0.8401276143211627
#alpha = 0.1 正确：4771 总的：5642 正确率：0.8456221198156681
```

### 2. 尝试优化

在得知别人的效果是 0.95 时，我决定优化。

#### 1.训练/测试集比例选区的问题

初始我选取的是训练：测试=7：3，换为训练：测试=19：1，后：

```
787 932 0.8444206008583691
[Finished in 3.7s]
```

事实证明，毫无效果。

#### 2.分词的效果不好？

因为我在实验一为了控制训练词典的大小，把出现频率<=10 的单词全部去除，这就可能会导致删掉了每一类新闻的核心关键词语，所以尝试把阈值设置为<=3
训练：测试 = 7：3 时的效果如下：

```
4756 5642 0.8429634881247784
[Finished in 22.6s]
```

训练：测试 = 19：1 时的结果如下：

```
797 932 0.8551502145922747
[Finished in 6.5s]
```

阈值为 5 且训练：测试 = 19：1 时的结果如下：

```
792 932 0.8497854077253219
[Finished in 5.6s]
```

事实证明，效果不明显

## 3. 误差分析

分词：分词的千差万别很容易影响效果，自己在做分词的过程中做了很多取舍，导致没能最大化的保留信息

文章本身：每篇文章都很短，且文章数量不是很大，这就导致每一类文章的信息不是很明显。