一．实验内容

利用 KNN 实验中得到的词典和统计词频等数据，完成文档的 Naive Bayes 分类，在 20news-18828 数据集上测试效果。

二．实验过程

1.数据预处理

为了完成 naive bayes 分类

$$C_{map} = argmax_{c \in C} \hat{P}(c|d) = argmax_{c \in C} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c)$$

需要根据训练数据计算 $\hat{P}(c) = \frac{N_c}{N}$ ，$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$

根据生成的词典需要统计 20 个文档类中的单词词频以及每个类中的单词数和单词的出现次数。

2.平滑

为了避免未出现在训练集中的 term 出现导致零概率问题，使用加 1 平滑技术，同时分母要加上类内不同单词个数。

$$\hat{P}(t_k|c) = \frac{T_{c,t_k} + 1}{\sum_{t' \in V} (T_{c,t'} + 1)} = \frac{T_{c,t_k} + 1}{(\sum_{t' \in V} T_{c,t'}) + B}$$

3.取 log 求和

由于 log 函数的单调性，将求得的后验概率取 log 值，将乘积问题转变为求和问题。

三．实验结果

Acc = 0.7943

四．存在问题

1.关键数据的保存：在做前一个实验时，并没有存储词典和词频的信息，在此次实验中重新调整代码进行了计算，之后要注意保存实验过程中的关键数据。