Homework 2 NBC 实验报告

实验内容：

实现朴素贝叶斯分类器，测试其在 20 Newsgroups 数据集上的效果

实验步骤：

1. 读取文章并预处理

与上一个实验一样，我需要将文章进行分词，提取词干，去停用词等操作来使得选取的单词能更好的表示文章内容并且压缩单词空间。 将文章按 topic 分类并按比例划分数据集。同学有直接读取上个实验的结果的，然而我感觉写起来并没有简单多少就每次处理了。然后我运行一次全部数据就要花 5 分钟。get_articles 函数返回一个以 topic 为关键字，以文章的 list 为值的词典。

2. 统计词频

用朴素贝叶斯分类器我需要统计的东西有：每个 topic 中每个词出现的频率，每个 topic 中的单词数量，还有用于平滑的所有不同单词的数量。

3. 朴素贝叶斯分类

贝叶斯定理和朴素贝叶斯：

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j | x_1, x_2, \ldots, x_n)$$

$$= \arg\max_{v_j \in V} \frac{P(x_1, x_2, \ldots, x_n | v_j) P(v_j)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \arg\max_{v_j \in V} P(x_1, x_2, \ldots, x_n | v_j) P(v_j)$$

Using the Naïve Bayes assumption:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(x_i | v_j)$$

为了避免有单词的词频是 0 导致结果变成零，加入平滑：

$$\hat{P}(x_i | v_j) \leftarrow \frac{n_c + mp}{n + m}$$

我这里假设每个单词在所有文章出现了 0.1 次。
为避免精度问题对公式取 log，最终写成 python：

```
for word in article:
    if not word in count[test_topic].keys():
        count[test_topic][word] = 0
    p += math.log(count[test_topic][word] + 0.1) - math.log(words_in_topics[test_topic] + 0.1*len(all_words))
p += math.log(len(train_articles[test_topic]))
```

4. 最终结果:

```
alt.atheism : 141 correct in 159 ,correct rate : 88.67924528301887%
comp.graphics : 139 correct in 194 ,correct rate : 71.64948453608247%
comp.os.ms-windows.misc : 102 correct in 197 ,correct rate : 51.776649746192895%
comp.sys.ibm.pc.hardware : 166 correct in 196 ,correct rate : 84.6938775510204%
comp.sys.mac.hardware : 177 correct in 192 ,correct rate : 92.1875%
comp.windows.x : 155 correct in 196 ,correct rate : 79.08163265306123%
misc.forsale : 164 correct in 194 ,correct rate : 84.5360824742268%
rec.autos : 193 correct in 198 ,correct rate : 97.47474747474747%
rec.motorcycles : 190 correct in 198 ,correct rate : 95.95959595959596%
rec.sport.baseball : 191 correct in 198 ,correct rate : 96.46464646464646%
rec.sport.hockey : 190 correct in 199 ,correct rate : 95.47738693467338%
sci.crypt : 186 correct in 198 ,correct rate : 93.93939393939394%
sci.electronics : 158 correct in 196 ,correct rate : 80.61224489795919%
sci.med : 178 correct in 198 ,correct rate : 89.8989898989899%
sci.space : 178 correct in 197 ,correct rate : 90.35532994923858%
soc.religion.christian : 186 correct in 199 ,correct rate : 93.46733668341709%
talk.politics.guns : 174 correct in 182 ,correct rate : 95.6043956043956%
talk.politics.mideast : 173 correct in 188 ,correct rate : 92.02127659574468%
talk.politics.misc : 125 correct in 155 ,correct rate : 80.64516129032258%
talk.religion.misc : 81 correct in 125 ,correct rate : 64.8%
all : 3247 correct in 3759 ,correct rate : 86.37935621175845%
[Finished in 368.0s]
```

用 80%做训练数据，20%做测试数据，最终在 3759 篇文章中有 3247 篇分类正确，正确率 86.4%。