Homework2 -- Naïve Bayesian (2018.11.23)

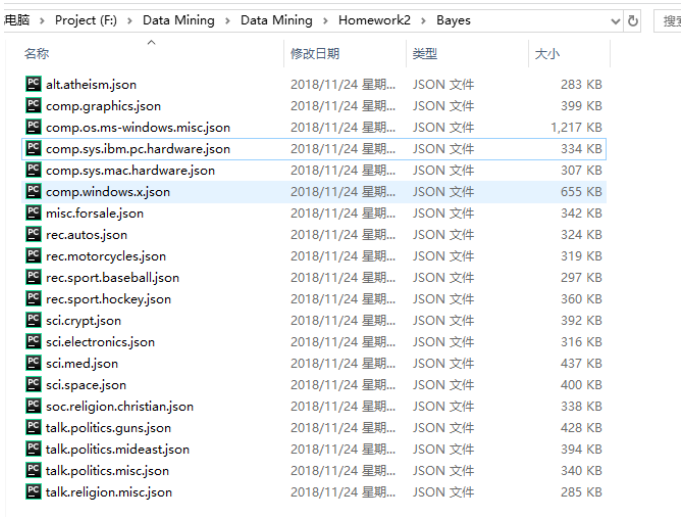
姓名：陈潇琳 学号：201834857

任务

实现贝叶斯分类器，测试其在 20Newsgroups 数据集上的效果

流程

1. 将数据划分为 80% training_data, 20%为 testing_data；
2. 同 Homework1，先对数据进行预处理：分词、分句、寻找词根、统计词频；
3. 根据 P_i = 此类的样本数/所有的样本数，计算每个类别的概率，此处应用拉布拉斯平滑；
4. 根据 **Naïve Bayesian** 公式，计算每个类别中单词的概率；
5. 为每个类别创建一个以类别为名的 json 文件，将 3 与 4，以字典的形式写入 json 文件，下图为文件列表；



6. 在 testing_data 中每个类别随机选取 10 个文件，进行测试，得到实验结果：0.425。

result.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
9905 predict: comp.sys.ibm.pc.hardware Groundtruth: comp.os.ms-windows.mis
21710 predict: soc.religion.christian Groundtruth: soc.religion.christian
179022 predict: rec.autos Groundtruth: talk.politics.misc
179026 predict: rec.autos Groundtruth: talk.politics.misc
59541 predict: rec.motorcycles Groundtruth: sci.med
103697 predict: rec.autos Groundtruth: rec.autos
68207 predict: comp.sys.ibm.pc.hardware Groundtruth: comp.windows.x
55095 predict: comp.sys.ibm.pc.hardware Groundtruth: talk.politics.guns
16048 predict: comp.sys.ibm.pc.hardware Groundtruth: sci.crypt
16050 predict: sci.crypt Groundtruth: sci.crypt
59543 predict: comp.sys.ibm.pc.hardware Groundtruth: sci.med
76798 predict: misc.forsale Groundtruth: misc.forsale
105144 predict: rec.motorcycles Groundtruth: rec.motorcycles
54284 predict: sci.electronics Groundtruth: sci.electronics
54178 predict: soc.religion.christian Groundtruth: alt.atheism
61087 predict: rec.autos Groundtruth: comp.sys.ibm.pc.hardware
55094 predict: comp.sys.mac.hardware Groundtruth: talk.politics.guns
179029 predict: talk.politics.guns Groundtruth: talk.politics.misc
61558 predict: sci.space Groundtruth: sci.space
21711 predict: soc.religion.christian Groundtruth: soc.religion.christian
54542 predict: rec.sport.baseball Groundtruth: rec.sport.hockey
54548 predict: rec.sport.hockey Groundtruth: rec.sport.hockey
```

总结

这次实验比第一次实验进行的要顺利，但是实验结果还需要改进，目前估计实验结果的不理想是因为 testing_data 数据的选取的造成。