知识分析与处理——朴素贝叶斯

实验数据

- 本次使用的数据集为讲解人提供spam_train和()的垃圾邮件数据集
- 需要将数据分词（jieba）,再用stopsword去除停用词
- 最后将读到的数据处理为词向量（词袋模型）
- 输出结果在run.txt中，前面的行为每一个训练数据的分类结果。最后一行为正确率

实现功能

- 读取数据,去除停用词
- 将处理后的数据转化为词向量
- 通过朴素贝叶斯算法，计算测试级词向量的分类概率
- 得到分类标签

算法分析

首先读入垃圾邮件数据，再分词后去除停用词
将去除分分词后的数据作为输入数据
对所以训练数据，将所有的词生成一个大的词集合
对于每一条输入数据：
    计算其词向量表示
计算训练数据的正负条件概率
对每一个测试样本：
    计算其为一般邮件的概率和垃圾邮件的概率
    取大的概率为其标签

实验结果

- 将测试结果与标注的结果做对比: 错误率 = 错误数量/测试数据总数
- 实验结果：

检验一

取500垃圾邮件信息为测试数据其余为训练数据
错误率为 0.02

检验二

res.txt保存了对于无标签测试数据的分类结果