基于朴素贝叶斯的垃圾邮件分类

1. 问题描述
2. 算法原理
3. 算法实现

根据上述说明，朴素贝叶斯算法进行邮件分类需要以下几步：

机器学习阶段：

* + 1. 设定非法字符集。如果文章中的某些文字属于该非法字符集，则进行过滤。
    2. 获取正常邮件中的词频。对训练集中的每一封正常邮件进行分词，使用设定的非法字符集进行过滤，并将邮件中出现的词保存到正常邮件单词表（normal-word-list）中。
    3. 统计正常邮件单词表中每个词在正常邮件中出现的次数，存入正常邮件词典中。
    4. 同理，获取垃圾邮件中的词频，（同上）。
    5. 同理，统计垃圾邮件词频表中每个词在垃圾邮件中出现的次数，存入垃圾邮件词典中。

测试阶段：

* + 1. 获取一封测试邮件中的词频。经过非法字符集过滤后，将邮件中的内容进行分词。
    2. 统计该邮件中每个词出现的次数。
    3. 计算该邮件中对分类影响最大的n个词，并对这些词求贝叶斯概率。
       1. 如果词语未出现在所有邮件中，则记为0.5
       2. 如果词语出现在正常邮件/垃圾邮件中，求该词在垃圾邮件中的频率、该词在正常邮件中的频率，然后计算该词的贝叶斯概率。
    4. 计算总贝叶斯概率，通过判断总概率与0.5的大小关系判断是否为垃圾邮件。

算法使用python实现。使用“jieba”python第三方库进行中文分词。

1. 训练及测试

训练集的大小是正常邮件与垃圾邮件各5000封左右。

测试集大小为300封左右。

我统计了

1. 结果分析