# 贝叶斯垃圾短信分类

————姚攀

## 程序功能

使用朴素贝叶斯分类识别垃圾短信,垃圾短信的特征抽取由余文艳完成,数据格式：

0 0 0 0 0 no

0 0 0 0 1 yes

……

0 0 0 0 1 yes

0 0 0 0 0 no

每条数据代表一条垃圾短信的特征和标注结果，前五列代表特征的取值(0 0 0 0 1)，第六列(yes/no)代表标注结果，no代表不是垃圾短信，yes代表是垃圾短信.用于训练的短信共80万条.

## 二、程序说明

训练数据存放着工程根目录的data/8w\_cut\_train.txt\_feature.txt 文本文件中，共有80万条.测试数据存放在test\_feature\_2W.txt文本文件中，共2万条,测试数据只包含短信特征，没有标注分类结果。使用贝叶斯分类后的结果存放在bayes\_2w\_feature.txt中。

输入：

test\_feature\_2W.txt

输出：

bayes\_2w\_feature.txt

对应文件内部每条数据：

输入：

1 1 0 0 0

输出：

1 1 0 0 0 no

## 三、运行方式

我的开发环境：

jdk版本：jdk 1.8

IDE:Eclipse

运行nlp.project.bayes包下Main.java即可。

工程目录：

