04

垃圾短信识别实验

Research results and applications

技术路线图

1	样本预处理	1.语料库	2.划分数据集	3.使用jiaba分词库进行分词
2	特征选择	1.Count vectorizer、2.TF-IDF		
3	模型构建	1.朴素贝叶斯、2.支持向量机		
4	评价	1.查准率、2.查全率、3.调和平均数		

样本预处理

样本包含带有0-1标记的文本信息80万条标签为1表示为垃圾短信，反之，标签0代表非垃圾短信

样本当中垃圾短信和非垃圾短信的数量分别为80000和720000条

样本缺失数量为0，样本重复数量为0，样本总字符数20513652。

样本内容

序号	标签	内容
1	0	商业秘密的秘密性那是维系其商业价值和垄断地位的前提条件之一
2	1	南口阿玛施新春第一批限量春装到店啦　　春暖花开淑女裙、冰蓝色公主衫　气质粉小西装、冰丝女王长半裙、　皇
3	0	带给我们大常州一场壮观的视觉盛宴
4	0	有原因不明的泌尿系统结石等
5	0	23年从盐城拉回来的麻麻的嫁妆
…	…	…
799999	0	费了半天劲各种找关系终于联系上心仪公司的内部人
800000	0	是汉奸还是被强奸自己对号入座吧

样本预处理

（b）垃圾短信频率最高词汇

（b）正常短信频率最高词汇

特征、模型选取

Count Vectorizer

TF-IDF

a. 特征

朴素贝叶斯

支持向量机

b. 模型

从表2中可以看出，支持向量机与CountVectorizer的模型对预测结果比较好。

表1 实验结果

特征向量 \ 模型		朴素贝叶斯			支持向量机		
		查准率	查全率	调和平均率	查准率	查全率	调和平均率
		precision	recall	f1-score	precision	recall	f1-score
Count Vectorizer	0	0.92	0.98	0.95	**0.99**	**0.95**	**0.97**
	1	0.98	0.91	0.94	**0.95**	**0.99**	**0.97**
	avg	0.95	0.94	0.94	**0.97**	**0.97**	**0.97**
TF-IDF	0	0.91	0.98	0.94	0.97	0.96	0.96
	1	0.98	0.9	0.94	0.96	0.97	0.97
	avg	0.94	0.94	0.94	0.97	0.97	0.97