数据说明：train_set.csv/test_set.csv

市长电话数据集是某北方城市政府部门冬季收到的“12345”服务热线的来电信息记录。数据集中的变量包括:

(1) 单位名称：即每一条来电信息记录最 终受理的政府部门；

(2) 词汇出现的频数：已有词汇表中各个词汇在投诉建议的文本内容中出现的频数。

数据总共有 2000 行，每一行代表一条投诉建议中各个词汇的出现频数，第 2-6236 列的列名为某个词汇。

分析任务：

1. 读入市长电话训练集 trainset.csv 和测试集 testset.csv，统计训练集中各个政府单位接到的市民投诉量，并绘制柱状图（按照投诉量降序），简单陈述你观察到的现象；

2. 统计每条投诉用词数并绘制分布直方图，简要分析你观察到的结果；

3. 并将各单位收集到投诉信息分词后总词数的差异用箱线图表示出来，尝试解读这个箱线图呈现的现象；

4. 以训练集中的政府单位为因变量类别，电话文本为自变量，尝试用朴素贝叶斯方法对市政电话文本进行分类，并使用该模型对测试集电话文本进行预测，计算混淆矩阵，简要分析模型的效果（提示：由于投诉建议文本整体长度较短，将每个词汇的被使用频数用是（1）-否（0）被使用替代，更为简便；利用 “e1071”包提供的 naiveBayes()函数进行朴素贝叶斯分类模型的建模；使用 R 自带的 image()函数进行混淆矩阵的绘制）