2018年南京信息工程大学第十二届数学建模竞赛赛题

（请先阅读“南京信息工程大学校内数学建模竞赛论文格式规范说明”）

**B题：垃圾短信的识别**

世界步入信息时代，通信业务的快速发展对于人们的日常生活产生重要影响。近年来，部分非法运营商运用垃圾短信传播不良信息，给人们造成了负面的影响。因此，识别和过滤垃圾短信具有重要的实际意义和应用价值。我们截取了某运营商部分数据，其中附件spamdata\_T1.csv包含571条短信；附件spamdata\_T2包含已标注类别的4973条短信；附件testdata.csv包含30条短信。要求你的研究团队根据附件中的短信信息，解决下列问题：

1. 通过分析附件spamdata\_T1.csv中文本数据，确定特征，给出短信类别。
2. 进一步分析附件spamdata\_T2.csv中垃圾短信(spam)及正常短信(ham)的特征，研究spam和ham的差异性。
3. 建立识别spam的数学模型，并用于识别附件testdata.csv中的spam。

附件1：spamdata\_T1.csv

附件2：spamdata\_T2.csv

附件3：testdata.csv