

# 项目 2A 说明

## 目标

这个项目的目标是是通过体验提高你以下的能力：

1. 做技术选择。
2. 综合各种观念来解决一个问题。
3. 了解如何把一个问题当作先例来解决另一个问题。
4. 处理一些比典型问题所提供的说明少的问题。
5. 如果你与人合作，则组织并开展工作。
6. 写出你的工作。

## 邮件分类问题

每天都有洪水般的邮件寄给我们。其中有我们想看的和不想看的，还有垃圾邮件——我们不想收到的广告邮件。虽然在 MIT，垃圾邮件相对而言被屏蔽了一些，然而，仍有大量的传进来。我们希望你能开发可以帮我们解决这个问题的软件。

## 基于内容的分类

已经有好多处理垃圾邮件的途径。一些系统具有黑名单，可以对其中的垃圾域名进行封闭。其它的通过查看邮件头部来拒绝许多垃圾邮件使用的假地址。还有一些寻找相关证据的方法，比如太多的惊叹点或许多大写字母。

不幸的是，当垃圾邮件过滤器正变得更加成熟时，反过滤技术也正在逐步提高。当过滤器变得越来越好时，垃圾邮件却变得越来越像真正的邮件。那么，我们如何来探测它呢？

幸运的是，垃圾邮件有一个永远不会改变的特点：它基本上是一个销售广告。基于内容的过滤器意味着我们设法寻找销售广告语言，而不是表面的一些特征，比如邮件使用“hot chixxx!” 多少次。

关于垃圾邮件过滤器的另一个需要考虑的重要事项是其失误带来的损失。如果过滤器拒收了一个真正的重要邮件，这比允许接收一个垃圾邮件更糟糕。另外一方面，如果接收太多的垃圾邮件，则这个过滤器又没有用处。

想得到关于垃圾邮件和基于内容的过滤器的一些好参考材料，请查看 [paulgraham.com](http://paulgraham.com)。

## 你的任务

你的任务是构建基于内容（题目行和邮件内容）的一个系统，可以确定一个邮件是否是垃圾邮件。你的系统要用垃圾邮件和非垃圾邮件来训练。

为了更精确，你应该按照以下来做：

基本要求：

1. 为确定垃圾邮件选择一个方法。
2. 实现基于这个方法的一个程序。
3. 测试并评估你的程序。

可选择的要求：

1. 允许用户反馈来对垃圾邮件的定义做进一步提炼。
2. 从成组的非垃圾邮件发现共性，将它们归类。

## 你以什么开始

已经为你提供了两个文件，包括从过去一周的一份邮件中选出的一部分。邮件已被去掉 HTML 和垃圾字符，只剩下题目和本体。另外，一些非常长的邮件已经被截短了。

1. spam.txt (60 个垃圾邮件)
2. notspam.txt (99 个非垃圾邮件)

注意你将从用来训练的数据中分离出检测数据。

## 检查要点

只有一个检查要点：在 MIT 允许的期限内，你应该提供给我们工作代码证据和你的最终报告。

## 报告长度

一片论文的恰当长度应该将你想说的内容做最简短的叙述。作为大概的指导，我们希望你写的不超过 5 页，除了图表、代码、打印输出等。如果你能在非常少的页面上叙述需要的内容，我们将会很吃惊。