2015 年“认证杯”数学中国数学建模网络挑战赛

第二阶段

B 题 替换式密码

历史上有许多密码的编制方法。较为简单的是替换式密码，也就是将文中出现的字符一对一地替换成其它的符号。对拼音文字而言，最简单的形式是单字母替换加密，也就是以每个字母为一个单位，将每个字母替换成另外的字母或者另外的符号。较为复杂的形式是以多个字母为一个单元，整体替换成其它的字符。这个映射方法被称为密码表，拿到密码表的人就能够将密文破译成明文。单字母替换加密是在古代就使用过的一种加密方法，但由于其容易被破解，所以在现代需要加密的场合已经很少使用。

单字母替换加密的破译方法有频率分析等。这种密码和破译方法在小说中也经常提到，例如爱伦·坡的《金甲虫》和柯南·道尔在福尔摩斯系列故事《归来记》中的“跳舞小人”。但当获取的密文篇幅不是很大时，光凭借频率分析是不足以破译全部密文的。往往还要熟知该种语言的人，经过对可能出现的词汇及字母组合进行分析，才能完整地破译密码。

第一阶段问题：假设明文是由现代通常使用的英语写成的。现在我们获取了一些由单字母加密方法加密的密文。请你建立合理的数学模型，设计一个算法，来自动化地破译密文。并设计一个衡量破译能力的标准，来评价破译算法的破译能力。为了问题简单起见，我们假设密码表仅是针对 26 个字母的，每个单词之间的空格，以及标点符号仍然会保留。在设计算法时，如果需要，可以参考英文语料库的数据，例如免费的 COCA[1] 等相关资料。

[1] http://corpus.byu.edu/full-text/formats.asp

第二阶段问题：在密码的实际使用中，我们获取的密文往往是经过各种干扰的。例如密文不完整，或者在通信传输的过程中，密文存在被丢失、添加和篡改字符的可能。为简单起见，我们假设密文是通过一个带有噪声干扰的信道传输的。在通信过程中，每个字符经过信道传输的结果都属于如下四种情形之一：

1. 该字符在传输过程中被丢失，其概率为 p_1；
2. 该字符本身正常传输，在其之后添加了一个随机字符，其概率为 p_2；
3. 该字符在传输过程中被篡改为一个随机字符，其概率为 p_3；
4. 该字符正常传输，其概率为 $1 - p_1 - p_2 - p_3$。

在加密后，单词之间的间隔和标点符号全部被删去。请你改进之前的破译算法，使其能够在这种带有干扰的条件下完成破译工作，并推广对破译能力的评价指标，使其能应用于这种带有干扰的条件。