中国科学技术大学

学士学位论文

密码学导论课程实践

题目一统计分析工具

作者姓名： 史昊霖

学科专业： 电子信息工程

导师姓名： 李卫海 副教授

完成时间： 二〇二一年五月二十六日

University of Science and Technology of China

A dissertation for bachelor's degree

A experimental report of the Introduction to Cryptology course practice topic 1: Statistical analysis tools

Author: Shi Haolin

Speciality: Electronic Information Engineering

Supervisor: Asso. Prof. Weihai Li

Finished time: May 26, 2021

中文内容摘要

本科生需要手动将摘要置于目录后。

本文为密码学导论课程实践的题目一：统计分析工具的实验报告。使用 Python 作为编程工具，第一个程序以字符形式读取数据，并对字母频率进行统计分析；第二个程序是简单的维吉尼亚密码加密程序；第三个程序通过查找第一个程序中统计出的字符串位置来尝试破解密钥长度。

关键词：中国科学技术大学；课程论文；密码学导论；统计分析；实验报告

Abstract

This paper is a experimental report of the Introduction to Cryptology course practice topic 1: Statistical analysis tools. Python is used as a programming tool, The first program reads data in the form of characters and makes statistical analysis of the frequency of letters; The second program is a simple Virginia encryption program; The third program tries to crack the key length by looking up the string position counted in the first program.

Key Words: University of Science and Technology of China (USTC); Course Paper; Introduction to Cryptology; Statistical analysis; Experimental Report

致　谢

感谢李卫海老师在课程方面的认真讲解和指导帮助。

感谢助教们批改作业、习题课的讲解，也感谢占用了助教课余时间的问题解答。

目 录

第一章　实验程序介绍

第一节　算法流程图

一、文本分析统计工具

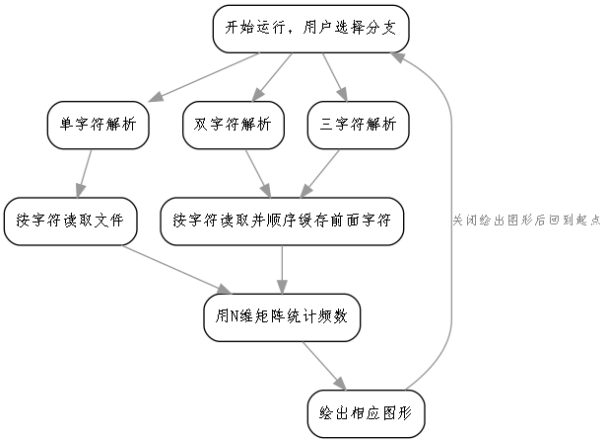

图 1.1　文本分析统计工具-程序流程图

二、维吉尼亚算法加密

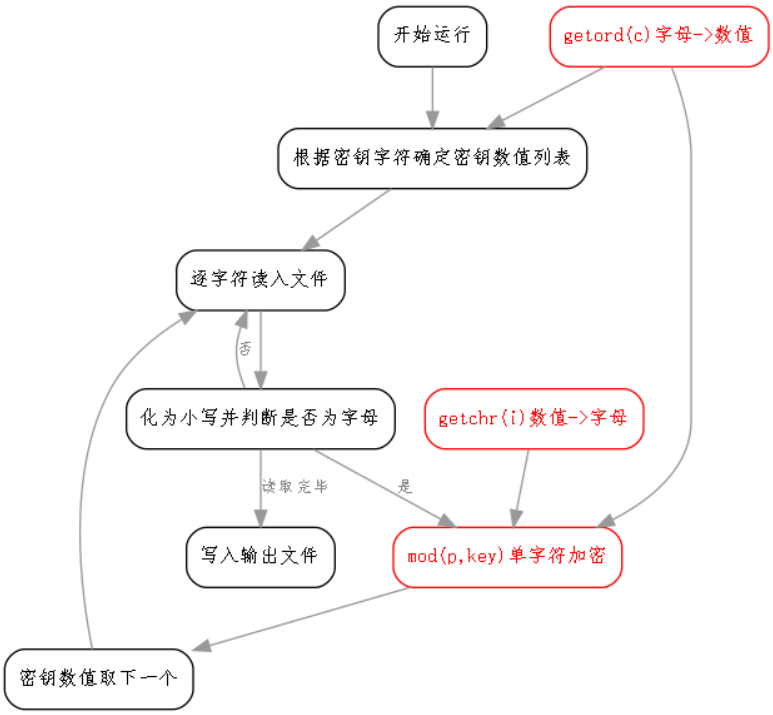

图 1.2　维吉尼亚算法加密-程序流程图

三、Kasiski 方法分析

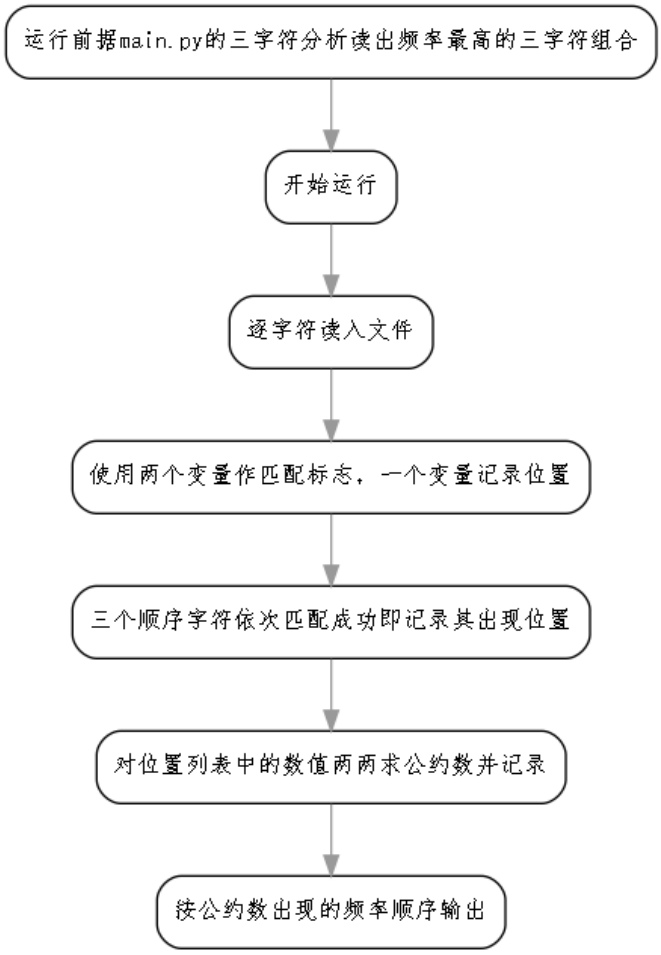

图 1.3　Kasiski 方法分析-程序流程图

第二节　运行效果图

选用的文本文件为《老人与海》的英文版 TXT 文本文档。分析时只取英文字母，并将大写字母转换为小写。

文档大小：134KB

一、原文/密文分析

密文使用维吉尼亚算法加密，密钥为“omymarblues”，11 个字符。

可以看出加密一定程度上令字符的分布更加均匀了，同时改变了字母与频度的对应关系。

这样可以增大字频统计攻击的难度，因此 Vigenere 加密是对凯撒密码的有效改进。

1. 单字符解析

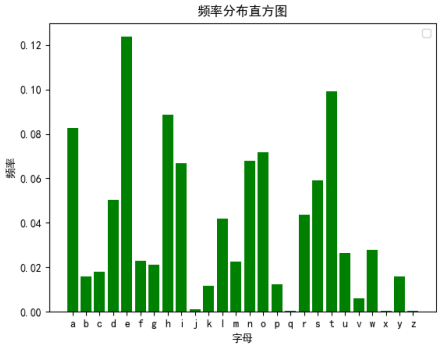

图 1.4　明文

最频繁：e 12640 次

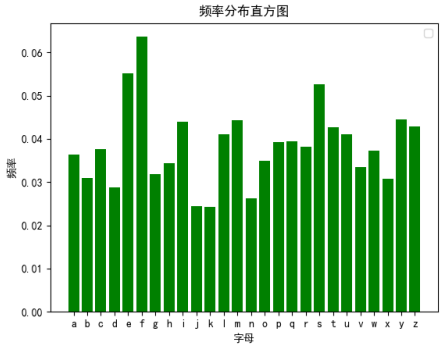

图 1.5　密文

最频繁：f 6350 次

2. 双字符组合解析

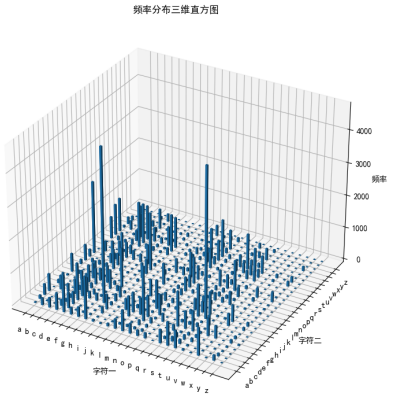

图 1.6　明文

最频繁：he 4.6174%

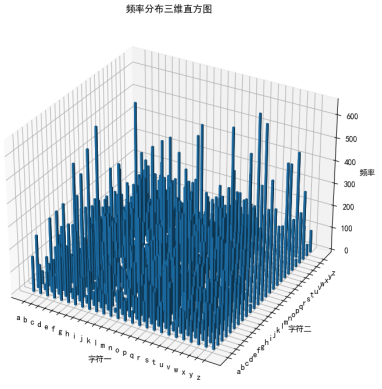

图 1.7　密文

最频繁：fh 0.6561%

3. 三字符组合解析

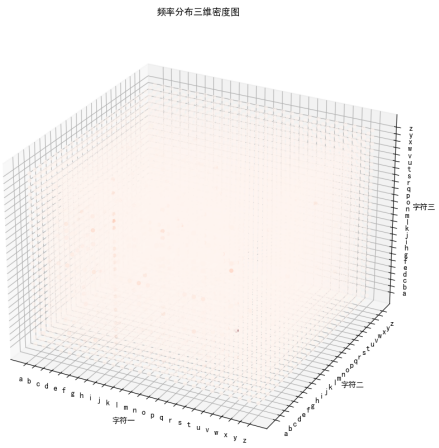

图 1.8　明文

最频繁：the 3.1318%

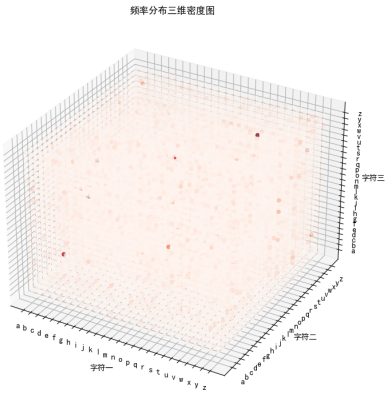

图 1.9　密文

最频繁：htc 0.2935%

第二章　实验结果分析

第一节　课程内简介

一、维吉尼亚密码

代换密码：维吉尼亚密码（Vigenère Cipher）是最简单的多表替换密钥，由多个凯撒替换表循环构成。

加密算法：Ci = E(K, P_i) = (P_i + $K_{imod d}$) mod 26

解密算法：Pi = D(K, C_i) = (C_i - $K_{imod d}$) mod 26

二、Kasiski 方法

Kasiski 方法是一种寻找维吉尼亚密码密钥长度的方法。

1. 假设

明文中存在重复字段。

当重复字段的间隔是 d 的整数倍时，将得到重复的密文。

不同的明文获得相同密文的巧合很少发生。

2. 操作过程

在密文中寻找重复字段。

计算重复字段的间距。

密钥长度 d 应是这些间距的公约数。

3. 缺点

查找算法运算量大，耗时长。

偶尔发生的巧合影响机器判断。

第二节　题目要求

编写一个软件，实现以下功能：统计一段文字中单字符、双字符、三字符的出现频率，分别用直方图、三维直方图、三维密度图表示出来。字符范围应可以指定；画图部分可以包含在你的软件之内，也可以单独使用某个工具实现。

自选一本英文书籍，用该工具分析字母统计分布（字母改为小写，删除所有非字母的符号）。

用维吉尼亚密码加密这本书，对得到的密文进行统计分析。

根据密文统计结果，用 Kasiski 方法分析密钥长度。（你发现了多少可用的串？是否遇到了偶然的干扰?）

第三节　Kasiski 方法分析结果

先用统计分析工具得出最常出现的三字符为“htc”。

利用参数“htc”对上述密文文件进行分析后得到:

11	24960
22	6208
33	3107
44	1389
66	918
55	808
77	560

表 2.1　公因数出现次数

可以看出 11 是其最常出现的公因数，次数远超其他。

并且很容易看出其他元素也含有 11 这个元素作为因数，因此可以确定 11 为密钥长度。

尝试换用其他密钥加密尝试，如 24 个字符“ohmygodthatisinteresting”。

统计分析工具三字符统计结果:

字符串：zvl 0.0015526861470343694

字符串：lpr 0.0015326514870726356

字符串：zvh 0.0014525128472257004

字符串：atc 0.0013523395474170316

字符串：gai 0.0013423222174361645

字符串：gns 0.0013423222174361645

字符串：lam 0.0013423222174361645

字符串：rns 0.0013122702274935638

这次最常出现的三字符为“zvl”，频率为 0.0015526861470343694。

统计后发现公因数 1 出现了 10203 次，可以看出这三个字符不是我们需要的。

频率第二的三字符为“lpr”，频率为 0.0015326514870726356。

进行尝试后发现公因数 12 出现了 5077 次，可以认为这个字符串查找的结果是有效的。分析得到：

12	5077
4	1336
36	633
1	610
60	252
132	111
84	102

表 2.2　公因数出现次数

可以看出频率高的公因数多为 12 的倍数，可以合理猜测密钥长度也是 12 的倍数，指定的密钥长度是 24，很可能在少数次验证后被猜测出来。

第三章　算法详细与结论

第一节　算法实现

程序的三种需求（统计、加密、查找密钥长度）分别使用三个 Python 文件实现。

实验程序源码地址：https://github.com/lclichen/Crypt2021/tree/master/CH1-FREQ

第二节　结论说明

维吉尼亚密码是一种对凯撒密码的有效改进加密方法，可以将密文的字符频率较大程度地均匀化，但还是可以看出一定趋势。

Kasiski 方法在密钥长度为质数或其质因数仅出现一次时效果最好，若质因数多次出现，会产生很大的干扰。

因此使用维吉尼亚密码时密钥长度可以尽量使用较大的合数，可以有效减少被破解的可能。

通过程序进行大量数据的统计可以对古典密码学中密文的破译、密钥长度的寻找提供极大的帮助。

参 考 文 献