# 实验三 恶意代码特征提取

班级：1703202

学号：1170300421

专业：信息安全

日期：2020.11.8

# 实验三 恶意代码特征提取

## 一、实验项目描述

1. 基于特征匹配的误用检测技术原理和方法
2. 掌握多模式匹配方法
3. 掌握基于双数组-自动机原理的多模式匹配原理
4. 基于双数组-AC算法的多模式特征匹配算法实现
5. 双数组-自动机的预处理：Next表、Base表、Check表、失效函数、输出函数构建
6. 双数组-自动机的特征扫描流程
7. 合理的数据结构
8. 利用构建的自动机扫描目标文件

（1）扫描文件发现命中某个模式，需明确给出命中的模式和该模式在文件中的位置

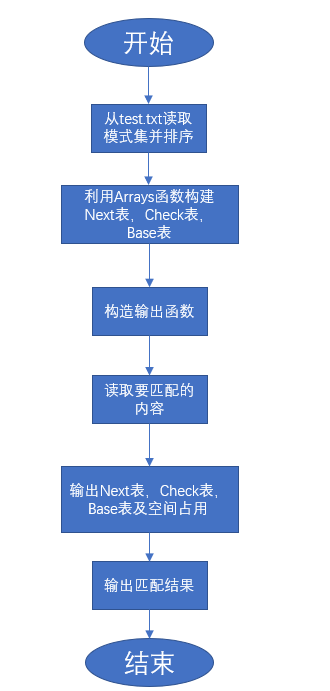
（2）如命中多个模式，需全部列出

## 二、实验要求

1. 实验数据准备。实验数据应简单实用：利用ASCII字符集做为输入集，不考虑多字节编码的中文、英文字符集。模式和待匹配文本可以只包含英文。
2. 程序的输入部分（模式、待匹配文本）需以文件方式提供或者其它非固定的输入途径。
3. 实验结果和实验数据一起给出：包括设定的模式有哪些？Next表、Base表、Check表是什么？待匹配文本是什么？输出结果是什么。注意：仅给出匹配结果是不够的，必须在输入多模式后，给出Next表、Base表、Check表内容，Next表中空间占用的百分比(Next表中状态占用的空间/Next表大小)，然后输入待检测文本，再输出检测结果。
4. 程序本身需要提交。

## 三、实验结果

1、程序流程图



2. 列出构建自动机所定义的数据结构，简单说明其功能

Next表：转向函数表，下标是位置偏移量，输出是状态值

Base表： key为状态值，value是base值

Check表：key是状态值，value是该状态父状态的值

Fail表：key为当前失效状态，value为失效值。

Output表： key为接收状态，value是匹配成功的模式集

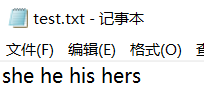
3. 根据这些数据结构说明Next表、Base表、Check表、失效函数、输出函数的构建过程

首先初始化一个AC自动机类用于后续构建。首先初始化一个TreeMap结构用于后续存放模式串，从文件中读入模式串，将每个模式串存入map中之后，调用AC自动机中的Array函数，传入这个map之后构建自动机。构建一个双数组Trie树，通过不断循环更新current\_state来更新Next、Base、Check表。将深度为1的节点的Fail设为根节点，为深度 > 1 的节点建立Fail表，这是一个bfs。构建完毕后传入check，调用parseText进行匹配。不断更新当前位置和状态，调用hit来进行匹配。同时更新output数组。

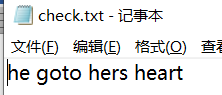
首先尝试按success表转移，失败了则按照Fail表回去，成功了继续按success表转移，直到失败跳转failure表，或者遇到Output表中标明的“可输出状态”然后输出所有数据。

4.实验结果

模式集：



测试集：



结果输出：

