МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение высшего образования

УРАЛЬСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

имени первого Президента России Б.Н. Ельцина  
ИНСТИТУТ ЕСТЕСТВЕННЫХ НАУК И МАТЕМАТИКИ

Департамент математики, механики и компьютерных наук

**СИСТЕМА ПОИСКА ЗАГРУЖЕННЫХ WEB-ШЕЛЛОВ**

**НА WEB-СЕРВЕРЕ**

Направление подготовки

02.03.01 «Математика и компьютерные науки»

|  |  |
| --- | --- |
| Директор департамента:  к. ф.-м. н. Е. С. Пьянзина  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | Выпускная квалификационная  работа бакалавра  **Арцыбашева Григория Игоревича**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| Нормоконтролер:  О. А. Суслова  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | Научный руководитель:  к. т. н., Р. В. Гибилинда  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |

Екатеринбург 2024

**РЕФЕРАТ**

Арцыбашев Г. И., Система поиска загруженных web-шеллов на web-сервере, Выпускная квалификационная работа: стр. 36, рис. 3, библ. 8 назв., 1 прил..

Ключевые слова: web-шелл, преобразование, энтропия, избыточность языка обфускация, расстояние Левенштейна, редакционное расстояние

Цели работы:

* выработка подходов для обнаружения web-шеллов;
* разработка системы поиска загруженных web-шеллов, реализующей выработанные подходы.

Методы работы: разработка программы с графическим интерфейсом на языке Python с использованием PyQt (набора расширений графического фреймворка Qt) и СУБД SQLite.

Результат: разработана система поиска загруженных web-шеллов.

**МЕСТО ВЫПОЛНЕНИЯ РАБОТЫ**

Работа выполнена по месту прохождения преддипломной практики в Департаменте математики, механики и компьютерных наук.

**СОДЕРЖАНИЕ**

[ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ 6](#_Toc168855827)

[ВВЕДЕНИЕ 9](#_Toc168855828)

[ОСНОВНАЯ ЧАСТЬ 11](#_Toc168855829)

[1. Постановка задачи 11](#_Toc168855830)

[2. Характерные особенности web-шеллов 11](#_Toc168855831)

[2.1. Подозрительные функции и команды 11](#_Toc168855832)

[2.2. Преобразование текста программы 12](#_Toc168855833)

[2.2.1. Обоснование выбора характеристики 12](#_Toc168855834)

[2.2.2. Способы выявления преобразования 13](#_Toc168855835)

[2.2.3. Информационная энтропия 13](#_Toc168855836)

[2.2.4. Избыточность языка 14](#_Toc168855837)

[2.2.5. Избыточность преобразованного текста 15](#_Toc168855838)

[2.3. Обфускация кода программы 15](#_Toc168855839)

[2.3.1. Цели применения обфускации 15](#_Toc168855840)

[2.3.2. Способы выявления обфускации 16](#_Toc168855841)

[2.3.3. Расстояние Левенштейна 17](#_Toc168855842)

[3. Реализация алгоритмов обнаружения web-шеллов 18](#_Toc168855843)

[3.1. Реализация алгоритма обнаружения преобразования 18](#_Toc168855844)

[3.1.1. Описание алгоритма 18](#_Toc168855845)

[3.1.2. Поиск и удаление непреобразованной части текста 18](#_Toc168855846)

[3.1.3. Взаимодействие со словарём 19](#_Toc168855847)

[3.1.4. Расчёт энтропии 19](#_Toc168855848)

[3.1.5. Вынесение вердикта по значению энтропии 20](#_Toc168855849)

[3.2. Реализация алгоритма обнаружения обфускации 21](#_Toc168855850)

[3.2.1. Поиск имён 21](#_Toc168855851)

[3.2.2. Поиск похожего слова в словаре 21](#_Toc168855852)

[3.2.3. Обнаружение обфускации 23](#_Toc168855853)

[3.3. Реализация алгоритма поиска подозрительных функций и команд 24](#_Toc168855854)

[3.3.1. Классификация подозрительных инструкций 24](#_Toc168855855)

[3.3.2. Поиск подозрительных инструкций 24](#_Toc168855856)

[4. Проверка корректности и эффективности работы алгоритмов 25](#_Toc168855857)

[4.1. Алгоритм обнаружения преобразования 25](#_Toc168855858)

[4.1.1. Описание выборки 25](#_Toc168855859)

[4.1.2. Проверка работы алгоритма 25](#_Toc168855860)

[4.2. Алгоритм обнаружения обфускации 26](#_Toc168855861)

[4.2.1. Описание выборки 26](#_Toc168855862)

[4.2.2. Проверка работы алгоритма 26](#_Toc168855863)

[5. Разработка системы поиска загруженных web-шеллов 27](#_Toc168855864)

[5.1. Интерфейс и поведение программы 27](#_Toc168855865)

[5.2. Анализатор файла 29](#_Toc168855866)

[5.3. Сверка хешей 30](#_Toc168855867)

[5.4. Использование базы данных 30](#_Toc168855868)

[5.4.1. Сущности 30](#_Toc168855869)

[5.4.2. Получение и запись данных 31](#_Toc168855870)

[5.5. Основной сервис 32](#_Toc168855871)

[5.6. Настройка программы 32](#_Toc168855872)

[ЗАКЛЮЧЕНИЕ 34](#_Toc168855873)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ 35](#_Toc168855874)

[ПРИЛОЖЕНИЕ 36](#_Toc168855875)

# ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящей работе применяются следующие обозначения и сокращения:

AES – Advanced Encryption Standard – симметричный алгоритм блочного шифрования, принятый в качестве стандарта шифрования правительством США и активно использующийся в настоящее время

ASP.NET – платформа для разработки динамических сайтов и web-приложений, созданная компанией Microsoft

Bash – Bourne again shell – одна из наиболее популярных современных разновидностей командной оболочки UNIX

DDoS-атака (Distributed Denial of Service) – разновидность DoS-атаки, при которой сетевой ресурс перегружают трафиком, исходящим из нескольких источников

DoS-атака (Denial of Service) – атака, направленная на создание помех или полную остановку работы сетевого ресурса путём его перегрузки большим количеством запросов

C# – язык программирования общего назначения, применяемый в том числе для создания web-приложений

CSS – Cascading Style Sheets – формальный язык описания внешнего вида web-страницы

DES – Data Encryption Standard –симметричный алгоритм блочного шифрования, утверждённый правительством США в 1977 году как официальный стандарт

Hex – hexadecimal – шестнадцатеричный код

HTML – HyperText Markup Language – стандартизированный язык гипертекстовой разметки документов

JS – JavaScript –язык программирования, обычно применяемый в браузерах как язык сценариев для придания интерактивности web-страницам

LZMA – Lempel-Ziv-Markov chain-Algorithm – алгоритм сжатия данных

MD-5 - 128-битный алгоритм хеширования

PHP – скриптовый язык общего назначения, обычно применяемый для разработки web-приложений

PNG – portable network graphics – формат сжатия изображений

Python – язык программирования общего назначения, применяемый, в том числе, для создания web-приложений

Ruby – язык программирования общего назначения, применяемый, в том числе, для создания web-приложений

SHA-256 – 256-битный алгоритм хеширования из семейства криптографических алгоритмов SHA-2 (Secure Hash Algorithm Version 2)

SQL – Structured Query Language – декларативный язык программирования, применяемый для создания, модификации и управления данными в реляционной БД

tar – tape archive – формат архивации файлов

WebP – WEB Pictures - формат сжатия изображений

Web-шелл (веб-оболочка) – командная оболочка для удаленного управления web-сервером

XML – eXtensible Markup Language – расширяемый язык разметки

zip – формат архивации файлов и сжатия данных без потерь

БД – база данных – упорядоченный набор структурированной информации или данных, которые хранятся в электронном виде в компьютерной системе

Дефейс сайта – изменение содержимого сайта, публикация нежелательных сообщений или полное его блокирование

Исходный код (код) – текст компьютерной программы на каком-либо удобочитаемом языке программирования, определяющий поведение компьютера

Кибератака – несанкционированное действие против компьютерной инфраструктуры, которое ставит под угрозу конфиденциальность, целостность или доступность её содержимого

Командная оболочка ОС – программа, с помощью которой пользователь или администратор управляет операционной системой и установленными программами, используя командную строку

Обфускация – процесс изменения кода программы, в результате которого он приобретает вид, трудный для понимания – при этом программа сохраняет свои функции

ОС – операционная система

ПО – программное обеспечение

Сервер – компьютер для выполнения сервисного ПО

СУБД – система управления базами данных – комплекс программ, позволяющих создать базу данных и манипулировать данными

Хеш – последовательность символов фиксированной длины, полученная путем преобразования произвольных исходных данных при помощи специального математического алгоритма, которая однозначно соответствует этим исходным данным, но не позволяет их восстановить

Хеширование – процесс вычисления хеша

Шифрование – такое обратимое преобразование информации, что неавторизованным лицам, не обладающим ключом для дешифровки, невозможно её понять

# ВВЕДЕНИЕ

Современное общество все больше зависит от информационных технологий, применяемых уже в большинстве сфер жизни. Эта растущая зависимость является основным фактором, вызывающим уязвимость к кибератакам, поскольку практически во всех компьютерных системах есть ошибки, допущенные разработчиками этих систем, которые могут быть использованы злоумышленниками. Кибератаки становятся все более распространенными и опасными, потери от них могут быть значительными как для организаций, так и для отдельных индивидуумов. Поэтому противодействие различным кибератакам является крайне актуальной задачей, борьба с ними – это непрерывная работа, поскольку угрозы постоянно эволюционируют, становятся более изощренными и сложными, что затрудняет их обнаружение и удаление.

Web-шелл, как командная оболочка, может использоваться для выполнения легитимных задач, однако чаще всего применяется в кибератаках, поскольку позволяет установить обратную связь с атакующим, давая ему полный контроль над сервером [1]. Web-шелл может быть написан на любом языке программирования, поддерживаемом сервером. Чаще всего используется PHP из-за его широкого использования для создания web-приложений. Также могу использоваться программы, написанные при помощи ASP.NET, Python, Ruby, JavaScript. Для внедрения web-шеллов злоумышленники используют различные методы, эксплуатируя уязвимости как самого сервера, так и связанных с ним web-сайтов или приложений. После установки и запуска web-шелл, предоставляя преступнику доступ к файловой системе и базам данных сервера, позволяет прослушивать трафик и удаленно выполнять команды [1].

Злоумышленники могут применять web-шеллы с различными целями, среди которых:

* дефейс сайта;
* кража данных, то есть получение доступа к конфиденциальной информации;
* объединение скомпрометированных серверов в сеть для проведения DDoS-атак;
* дальнейшее проникновение в сеть жертвы, включая устройства, недоступные напрямую из публичной сети, но доступные с зараженного web-шеллом сервера;
* заражение устройств посетителей скомпрометированного web-ресурса.

Таким образом, web-шелл дает злоумышленникам мощный инструмент для нанесения ущерба различным системам, кражи данных и распространения вредоносного ПО.

Существуют разные подходы в обнаружении web-шеллов. Наличие в системе web-шелла под управлением злоумышленников может, например, проявляться в следующем [1]:

* присутствие на сервере неизвестных файлов;
* присутствие на сервере файлов с необычным временем создания (например, более новых, чем последние обновления установленных приложений);
* длительные периоды большой нагрузки на сервер (злоумышленник может загружать очень большой объём данных).

Методы обнаружения, предлагаемые ниже в работе, основаны на анализе содержимого файла с целью выявления в нём характерных для типичного web-шелла особенностей. Целью работы является реализация этих методов в виде программы с графическим интерфейсом, которую сможет использовать системный администратор для быстрого поиска подозрительных файлов. Программа должна «подсказать», какие из файлов могут оказаться web-шеллами. Таким образом, программа играет роль вспомогательного инструмента при поиске загруженных web-шеллов, при этом web-шеллу не обязательно быть активированным, то есть введённым в действие.

# ОСНОВНАЯ ЧАСТЬ

## **1. Постановка задачи**

Для выполнения обозначенных целей работы были определены следующие задачи:

* выявление характерных особенностей типичных web-шеллов;
* выработка подходов в поиске выявленных особенностей в исходном коде web-приложений, реализованных на языках программирования PHP, JavaScript, Python, Ruby, C#;
* реализация и проверка корректности работы алгоритмов поиска загруженных web-шеллов;
* разработка системы поиска загруженных web-шеллов.

## **2. Характерные особенности web-шеллов**

В ходе изучения исходных кодов web-шеллов [2] были определены следующие характерные особенности:

* подозрительные функции и команды;
* наличие преобразования (подстановки, шифрование) над текстом программы;
* обфускация (один из видов преобразования) текста программы.

Обоснование отдельного рассмотрения обфускации будет приведено позже.

Важно отметить, что и легитимные файлы могут иметь указанные особенности. Поэтому однозначно определить, является ли определённый файл web-шеллом, может быть затруднительно.

### **2.1. Подозрительные функции и команды**

К подозрительным функциям можно отнести функции языков программирования, которые могут быть использованы злоумышленником для:

1. преобразования текста (например, обратное преобразование текста с помощью функции *base64\_decode* в PHP);
2. вычисления и выполнения некоторого выражения во время работы программы (например, передача команды форматирования диска для командной оболочки ОС с помощью функции *system* в PHP);
3. взаимодействия с файловой системой и файлами в частности;
4. загрузки и отправки файлов по сети;
5. взаимодействия с функциями ОС (например, обращения к переменным среды).

Также в исходном коде web-шелла могут содержаться команды для оболочки ОС сервера, позволяющие, например, взаимодействовать с файлами. К таким можно отнести, например, *ls*, *rmdir*, *touch*. Также подозрительными можно считать команды *arp*, *awk*, *ipconfig* и другие.

Все указанные типы функций и команд сами по себе абсолютно безвредны и часто используются в легитимных программах, как, например, функции для преобразования текста. Однако некоторые из функций достаточно редко встречаются в исходных кодах приложений, чего нельзя сказать про web-шеллы. Яркий тому пример – функции для вычисления и выполнения некоторого выражения во время работы программы. Чем шире набор подозрительных функций и команд в исходном коде программы, тем выше вероятность, что это web-шелл.

### **2.2. Преобразование текста программы**

#### **2.2.1. Обоснование выбора характеристики**

Для затруднения обнаружения вредоносного кода злоумышленник может преобразовать его так, что его предназначение становится непонятным для человека или программы-анализатора. На практике web-шеллы часто содержат в себе преобразованные фрагменты кода, над которыми совершается обратное преобразование, после чего они выполняются.

Таким образом, поиск подозрительных функций или команд может дать нулевой результат при подобной организации web-шелла. Поэтому необходимо уметь выявлять наличие различного рода преобразований в коде программы.

Важно отметить, что преобразование само по себе не является признаком вредоносной программы. Преобразование может использоваться при сжатии информации (архивы, изображения, данные), а также при передаче легитимных веб-скриптов по сети. Поэтому обнаружение преобразования в коде ещё не говорит о том, что он вредоносен.

#### **2.2.2. Способы выявления преобразования**

Для выявления преобразования будет использоваться два подхода: простой поиск hex-кода и вычисление энтропии текста.

Нередко используется достаточно простое преобразование, когда исходный текст преобразуется в советующий ему hex-код. Такое преобразование легко распутать, однако текст в таком виде становится нечитаемым и непонятным, а именно этого и хочет добиться злоумышленник. Обнаружить hex-код нетрудно ввиду особенностей его написания в коде программы, для этого будет достаточно обычного поиска по тексту специфических последовательностей символов.

Более сложные преобразования (например, base64) обнаружить таким образом куда сложнее, в этом случае используется другой подход. Предполагается, что преобразование изменяет статистические свойства текста, и, следовательно, энтропия текста может быть использована для его выявления. Энтропия в данном контексте относится к статистической мере неопределенности символов в тексте.

Ниже будет приведено обоснование применения энтропии в контексте обнаружения преобразований.

#### **2.2.3. Информационная энтропия**

Информационная энтропия представляет собой меру неопределенности в системе, особенно в отношении непредсказуемости появления символов первичного алфавита. В контексте передачи сообщений, энтропия численно соответствует количеству информации, приходящемуся на каждый символ сообщения при отсутствии информационных потерь.

Для расчета энтропии в работе будет использоваться формула Шеннона [3]. Пусть для каждого символа алфавита определена вероятность его появления () в информационном сообщении. Тогда энтропия сообщения, построенного на алфавите , будет вычисляться по формуле:

,

где – некоторая положительная константа.

Это соотношение было получено Шенноном для определения среднего количества информации в сообщении с произвольными вероятностями появления значений символов [4].

#### **2.2.4. Избыточность языка**

Избыточность языка – это статистическая величина, обозначающая избыточность информации, содержащейся в тексте на определённом языке. Для формального определения избыточности введем понятия нормы (энтропии) и абсолютной нормы (абсолютной энтропии) языка [5].

Для данного языка норма (энтропия) языка равна:

,

где – энтропия сообщения , – длина сообщения. Для примера, норма обычного английского языка при больших принимает различные значения от 1,0 до 1,5 бит на букву.

Абсолютная норма (абсолютная энтропия) языка равна максимальному количеству битов, которое может быть передано каждым символом при условии, что все последовательности символов равновероятны. Если в алфавите языка используется различных букв, то абсолютная норма равна:

.

Для английского языка с 26 буквами в алфавите абсолютная норма равна  ~ 4,7 бит на букву. Действительная норма английского языка намного меньше абсолютной, поскольку естественные языки обладают высокой избыточностью.

Избыточность языка, обозначаемая , определяется так:

.

Принимая норму английского языка за 1,3, избыточность составит 3,4 бит на букву, то есть каждая английская буква содержит 3,4 бита избыточной информации. Следовательно, у сообщения ASCII (где каждый символ кодируется одним байтом), состоящего только из английских букв, количество избыточной информации на каждый байт составляет 6.7 бит.

#### **2.2.5. Избыточность преобразованного текста**

Избыточность достаточна высока в текстах, написанных на естественных языках. В тексте, представляющем собой исходный код программы, где гораздо активнее используются знаки пунктуации, цифры, а также специальные символы, избыточность ниже. Но еще ниже избыточность в преобразованных (например, зашифрованных или сжатых) текстах, поскольку одно из характерных свойств алгоритма преобразования – перераспределение частот встречаемости значений байтов таким образом, что вероятность появления каждого из них примерно одинакова.

Более-менее одинаковые частоты встречаемости приводят к заметному росту энтропии, что означает сильное уменьшение избыточности. И чем больше символов будет использоваться для преобразования текста, тем ближе к нулю будет избыточность и тем ближе будет значение энтропии к значению абсолютной энтропии.

На этом факте и основана идея обнаружения преобразования. Чем меньше избыточность текста, тем вероятнее, что он преобразован.

### **2.3. Обфускация кода программы**

#### **2.3.1. Цели применения обфускации**

Обфускация может применяется в следующих целях:

* защита интеллектуальной собственности (сокрытие алгоритмов и логики работы программы);
* конфиденциальность разработки (предотвращение кражи и дальнейшего анализа кода);
* предотвращение кибератак (сокрытие уязвимостей, которые могут быть использованы злоумышленниками для взлома системы).

Ранее уже было отмечено, что преобразования могут быть использованы злоумышленником для сокрытия вредоносного кода. Обфускация является преобразованием, которое сохраняет вычисляемую программой функцию, но при этом придает программе такую форму, что извлечение из текста программы ключевой информации об алгоритмах и структурах данных, реализованных в этой программе, становится трудоёмкой задачей [6]. Это обстоятельство делает обфускацию одним из методов, применяемых в разработке вредоносных программ, в том числе и web-шеллах.

Обфусцирование кода может иметь разную степень сложности. В простейшем случае, это изменение имён переменных, функций, методов, классов и т.п. Более сложные способы обфускации в работе рассматриваться не будут.

#### **2.3.2. Способы выявления обфускации**

Для обнаружения обфускации можно так же использовать расчёт энтропии, однако в ходе выполнения работы было установлено, что в общем случае значение энтропии текста, преобразованного, например, популярными алгоритмами кодирования base64 или ASCII85, отличается от значения энтропии обфусцированного кода в большую сторону. Для отделения преобразованных текстов от непреобразованных необходимо установить границу значений энтропии, переходя которую, текст будет считаться преобразованным. Устанавливая границу такой, чтобы преобразованными считались в том числе и обфусцированные коды программ, довольно немалая часть непреобразованных текстов (кодов) будут признаваться преобразованными, то есть будет много ложных срабатываний. Чтобы уменьшить их число, для обнаружения обфусцированного кода будет использоваться другой подход.

Изменение имён выглядит следующим образом: оригинальное имя, составленное, как правило, из осмысленных слов английского алфавита, преобразуется в бессмысленный набор символов, допустимых для именования. Поскольку изменение имён почти всегда используется при обфускации, именно оно и послужит её признаком.

Способ обнаружения обфускации заключается в следующем:

1. простым поиском найти в коде имена переменных, констант, функций, методов и классов;
2. попытаться вычленить из каждого имени осмысленные слова;
3. если имя состоит преимущественно из осмысленных слов, обфускация к нему не применялась;
4. в конечном счёте по количеству обфусцированных имен станет ясно, была ли примерена обфускация к коду программы, или нет.

Применение отдельного подхода к поиску обфусцированного кода позволяет повысить точность работы алгоритмов.

#### **2.3.3. Расстояние Левенштейна**

Для определения того, является ли отдельно взятое слово осмысленным, будет использоваться следующий подход: в заранее подготовленном словаре будет искаться слово, похожее на рассматриваемое. Степень похожести определяется допустимым расстоянием Левенштейна между словами.

Расстояние Левенштейна – это метрика, измеряющая по модулю разность между двумя последовательностями символов. Она определяется как минимальное количество односимвольных операций (а именно вставки, удаления, замены), необходимых для превращения одной последовательности символов в другую [7].

Пусть и – два слова длиной и соответственно, тогда расстояние Левенштейна рассчитывается по следующей рекуррентной формуле:

, где

*deleteCost*, *insertCost*, *replaceCost* – стоимости удаления, вставки и замены символа соответственно. В нашем случае нет смысла выставлять приоритет операций, поэтому все стоимости будут равны единице.

## **3. Реализация алгоритмов обнаружения web-шеллов**

### **3.1. Реализация алгоритма обнаружения преобразования**

#### **3.1.1. Описание алгоритма**

На практике анализируемый текст может быть преобразован лишь частично, например, на 5%. Эта часть текста будет характеризоваться высокой энтропией. Остальные 95% будут иметь невысокую энтропию. В результате может получится так, что энтропия по тексту в целом тоже будет невысокой, так как 95% «затрут» остальные 5%. Во избежание подобной ситуации было придумано два решения: расчёт энтропии методом «скользящего» окна и игнорирование при расчёте всех английских[[1]](#footnote-1) слов и некоторых синтаксических конструкций используемого языка. Это позволит локализовать преимущественно высокоэнтропийные области в тексте.

Работу алгоритма можно поделить на следующие этапы:

1. удаление из текста всех символов, не относящихся к преобразованным фрагментам;
2. вычисление энтропии текста в целом;
3. вычисление энтропии текста методом «скользящего» окна;
4. вынесение вердикта по результатам вычисленных значений энтропии.

В следующих разделах каждый этап алгоритма будет рассмотрен отдельно.

#### **3.1.2. Поиск и удаление непреобразованной части текста**

Для поиска непреобразованной части текста используется словарь английских слов объёмом примерно 3800 слов.

Алгоритм поиска и удаления заключается в следующем:

* + - 1. весь текст разбивается на токены;
      2. все буквенные токены разбиваются на слова, которые в свою очередь сопоставляются со словарём слов: если большая часть слов была найдена в словаре, то значение токена признается непреобразованным, в противном случае значение определяется как преобразованное;
      3. формируется последовательность токенов:
         1. если предыдущий и текущий буквенные токены были признаны преобразованными, то эти токены, а также все небуквенные токены между ними, будут сохранены;
         2. если один из двух указанных буквенных токенов был признан непреобразованным, сохранены будут только те токены, которые ближе к преобразованному;
         3. в случае, когда оба буквенных токена определены как непреобразованные, то они, а также все небуквенные токены между ними, будут проигнорированы, и их значения не попадут в результирующий текст;
      4. из последовательности токенов формируется итоговый вариант текста.

Токен представлен классом *Token*, хранит в себе значение (последовательность байт) и тип (буквы, цифры, символ нижнего подчеркивания, прочие символы). Вычлененное из токена слово оборачивается в класс *Word*. Для формирования последовательности токенов и вычленения из них слов используются классы *TokenExtractor* и *WordExtractor.*

#### **3.1.3. Взаимодействие со словарём**

Взаимодействие со словарём осуществляется посредством класса *WordDictService*, позволяющий эффективно (сложность ) проверить наличие слова в словаре.

#### **3.1.4. Расчёт энтропии**

Класс *EntropyCalculator* производит расчёт энтропии. Класс *EntropyAnalyzer* использует *EntropyCalculator* для расчёта энтропии по тексту в целом и методом «скользящего» окна. Размер окна составляет часть от размера текста, величина сдвига окна – от его размера. Метод *EntropyAnalyzer.window\_analyze* двигает окно вдоль всего текста, в каждом «окне» рассчитывает энтропию попавших в него символов и выдает результаты расчётов в виде массива.

Указанные размер окна и величина его сдвига были получены в ходе «обучения» программы и признаны оптимальными.

#### **3.1.5. Вынесение вердикта по значению энтропии**

Для обнаружения преобразования в тексте используется следующий подход. Сначала проверяется уровень энтропии текста в целом. Если он оказывается ниже определённого значения, при котором текст считался бы преобразованным, проверяется, сколько «окон» можно отнести к преобразованным, то есть какой процент «окон» содержит в себе текст с высокой энтропией. Ниже приведены следующие поля класса *EncryptionDeterminatorByEntropy*, который непосредственно отвечает за обнаружение преобразования:

* *\_window\_encryption\_border* – все «окна» с энтропией равной или большей значения этого поля будут считаться «окнами» с высокой энтропией;
* *\_unconditional\_lower\_bound\_of\_entropy* – текст с энтропией равной или большей значения этого поля будет считаться преобразованным;
* *\_conditional\_lower\_bound\_of\_entropy* – текст с энтропией равной или большей значения этого поля будет считаться преобразованным, если процент окон с высокой энтропией будет не меньше, чем *\_percent\_of\_entropy\_vals\_for\_window*;
* *\_percent\_of\_entropy\_vals\_for\_window* – см. выше;
* *\_upper\_bound\_of\_entropy\_strict*, *\_upper\_bound\_of\_entropy\_optimal* – текст с энтропией равной или большей значения одного из этих полей (какое из полей будет выбрано, определяется режимом работы алгоритма: строгим или оптимальным) будет считаться непреобразованным; эти границы необходимы, чтобы исключить ложные срабатывания на высокоэнтропийных, но легитимных файлах (например, сжатые архивы или изображения).

Значения полей, указанных выше, получены в ходе «обучения» программы и были признаны оптимальными на основании метрик.

### **3.2. Реализация алгоритма обнаружения обфускации**

#### **3.2.1. Поиск имён**

Для поиска имён переменных и констант, функций и методов, классов используются регулярные выражения, инкапсулированные в классы *VariableSearcher*, *FunctionSearcher* и *ClassSearcher* соответственно. Эти классы являются наследниками класса *AbstractSearcher*. Метод *AbstractSearcher.get\_name\_iter* принимает на вход анализируемый текст, а возвращает итератор по именам, найденным в нём. Все имена представлены экземплярами класса *Name*.

Прежде, чем передать имя (*Name*) на анализ алгоритму обнаружения обфускации, оно переводится в экземпляр класса *NameInfo* с помощью класса *NameProcessor*. *NameInfo* хранит в себе информацию о том, какие слова содержит имя, сколько в имени букв и цифр. Вычленение слов из имени осуществляется посредством описанных ранее классов: *Token*, *TokenExtractor*, *Word*, *WordExtractor*.

#### **3.2.2. Поиск похожего слова в словаре**

Ранее было описано, что для определения того, является ли слово осмысленным, для него ищется похожее слово в словаре, имеющее расстояние Левенштейна не больше указанного.

Расчёт метрики Левенштейна осуществляет класс *CalculatorLevenshteinMetric*, принимая на вход два слова (экземпляры класса *Word*) и возвращая целое число – редакционное расстояние между ними. Для расчёта используется алгоритм Вагнера-Фишера. Данная реализация имеет временную сложность и пространственную – , где – длины передаваемых слов.

Поиск похожего слова в словаре осуществляет класс *SearcherByLevenshteinMetric*. Алгоритм поиска:

1. проверить, есть ли проверяемое слово в словаре (сложность ), если да, алгоритм завершает работу;
2. если слово не нашлось, алгоритм продолжает работу: на основе длины проверяемого слова и заранее переданного множителя рассчитывается максимально допустимое редакционное расстояние (расстояние равно округленному до ближайшего целого значению произведения длины слова и множителя);
3. далее выстраивается порядок длин слов, среди которых будет искаться похожее, при этом в первую очередь будут рассматриваться слова, «близкие» к проверяемому слову (например, если проверяемое слово содержит 7 букв, а максимально допустимое расстояние равно 3, то будет выстроен следующий порядок: 7, 6, 8, 5, 9, 4, 10; то есть сначала будут рассмотрены все слова той же длины (7), что и проверяемое слово, затем слова, отличающиеся по длине на один символ (6 и 8), на два символа (5 и 9) и, наконец, на три символа (4 и 10); рассматривать слова меньшей/большей длины (начиная с 3 и 11 соответственно) не имеет смысла, поскольку расстояние Левенштейна для таких слов будет уже больше 3);
4. выстроив порядок длин слов, алгоритм в соответствии с ним перебирает все слова, пытаясь найти похожее;
5. как только будет найдено слово, с расстоянием не больше указанного, алгоритм завершит свою работу, вернув слово и расстояние; если же были рассмотрены все допустимые слова, и при этом не было найдено ни одного похожего, алгоритм вернет пустое значение, означающее, что похожих слов в словаре нет.

Словарь, используемый классом *SearcherByLevenshteinMetric*, это тот же словарь, который используется для поиска и удаления непреобразованной части текста в задаче обнаружения преобразования. Работа со словарём осуществляется с помощью класса *WordDictService*. Метод класса *get\_words\_with\_len* возвращает множество всех слов в словаре указанной длины за .

#### **3.2.3. Обнаружение обфускации**

Алгоритм обнаружения обфускации инкапсулирован в классе *ObfuscationDeterminator*. Класс содержит три поля, определяющих работу алгоритма:

* *\_obf\_text\_border* – если доля обфусцированных имен в коде окажется выше этого значения, текст признается обфусцированным;
* *\_obf\_name\_border* – если доля неосмысленных слов в имени окажется выше этого значения, имя признается обфусцированным;
* *\_max\_non\_obf\_count\_digits* – максимальное количество цифр в имени, при котором оно ещё не считается обфусцированным, и по нему будет проводится проверка на наличие осмысленных слов; если цифр в имени больше значения этого поля, оно признается обфусцированным.

Значения полей, указанных выше, получены в ходе «обучения» программы и были признаны оптимальными на основании метрик.

Работу алгоритма можно поделить на следующие этапы:

1. найти все имена в коде;
2. для каждого имени определить, является ли оно обфусцированным (для этого используется характеристика имени или поиск похожего слова в словаре);
3. по количеству обфусцированных имён определить, является ли код обфусцированным.

### **3.3. Реализация алгоритма поиска подозрительных функций и команд**

#### **3.3.1. Классификация подозрительных инструкций**

Для удобства все подозрительные инструкции были разнесены в различные группы. Для этого было придумано две классификации: по уровню угрозы и по типу или области действия.

Классификация по уровню угрозы:

* опасные (DANGEROUS) – инструкции, представляющие наибольшую опасность (в среде программистов существуют общие рекомендации не использовать подобные инструкции в коде своих программ во избежание возможного появления в них уязвимостей);
* подозрительные (SUSPICIOUS) – инструкции, не представляющие сами по себе какой-либо угрозы, но которые часто используются и применяются в web-шеллах;
* требующие внимания (PAY\_ATTENTION) – инструкции, которые очень часто используются в легитимных программах, однако могут быть использованы злоумышленником (например, загрузка файлов по сети).

Классификация по типу/области действия:

* команды для оболочек ОС (COMMAND);
* преобразование (ENCRYPT);
* вычисление переданного выражения (EXECUTION);
* взаимодействие с файлами (FILES);
* подключение внешних скриптов/библиотек, вызывающих подозрение (IMPORT);
* загрузка/отправка файлов по сети (NET);
* взаимодействие с функциями ОС (OS);
* остальные подозрительные инструкции (GENERAL);

#### **3.3.2. Поиск подозрительных инструкций**

Задача поиска подозрительных функций и команд возложена на класс *SuspySearcher*. Непосредственно поиск осуществляется с помощью регулярных выражений. Метод *search* получает на вход анализируемый текст, после чего, перебирая шаблоны регулярных выражений, ищет все возможные совпадения тексте. Одновременно с обнаружением подозрительной инструкции проводится её классификация.

## **4. Проверка корректности и эффективности работы алгоритмов**

### **4.1. Алгоритм обнаружения преобразования**

#### **4.1.1. Описание выборки**

Для проверки работы алгоритма обнаружения преобразования была собрана выборка из 92 файлов, из которых 46 преобразованы.

В выборку непреобразованных файлов были включены файлы, представляющие собой исходные коды программ, реализованных с использованием следующих языков программирования: JavaScript, PHP, Python, Ruby, C#. Также представлены файлы HTML, CSS, SQL, XML, bash. Кроме того, в выборку были включены архивы, как сжатые (zip через LZMA), так и не сжатые (zip, tar), а также файлы-изображения (PNG, WebP).

В выборку преобразованных файлов были включены файлы, к которым были применены следующие алгоритмы:

* base32, base64, base85 (ASCII85), base122;
* rot13, rot18, rot47;
* AES;
* DES (triple).

Кроме этого проверялась обнаружение файла, содержащего hex-код.

При этом для каждого способа преобразования, кроме rotXX, было произведено несколько степеней преобразования, файлы были преобразованы на 5%, 10%, 30%, 50% и 100%.

#### **4.1.2. Проверка работы алгоритма**

Результат работы алгоритма по описанной выборке файлов (под срабатыванием подразумевается обнаружение преобразования, под пропуском – не обнаружение):

* верное срабатывание – 43 файла;
* ложное срабатывание – 15 файлов;
* верный пропуск – 31 файл;
* ложный пропуск – 3 файла.

Все сжатые файлы (архивы и изображения) были признаны непреобразованными, поскольку алгоритм работал в оптимальном режиме, при котором файлы с крайне высокой энтропией не включаются в множество преобразованных. Все ложные пропуски обусловлены файлами, преобразованными по алгоритму rotXX. Поскольку rotXX, в отличие от baseXX, AES или DES, используется для простой обфускации текста путём сдвига по алфавиту, энтропия оригинального и преобразованного текстов будет примерно одинакова. Вследствие этого алгоритм определяет файлы, преобразованные rotXX, как непреобразованные.

Метрики по выборке:

* точность – 0,74;
* полнота – 0,93;
* F-метрика – 0,83.

Следует отметить, что теоретически F-метрика могла быть чуть выше, однако параметры алгоритма подобраны так, чтобы полнота была наибольшей, вследствие чего снижается точность.

### **4.2. Алгоритм обнаружения обфускации**

#### **4.2.1. Описание выборки**

Для проверки работы алгоритма обнаружения преобразования была собрана выборка из 26 файлов, из которых 13 обфусцированы.

В выборку представлены файлы исходных кодов программ, реализованных с использованием следующих языков программирования: JavaScript, PHP, Python.

#### **4.2.2. Проверка работы алгоритма**

Результат работы алгоритма по описанной выборке файлов (под срабатыванием подразумевается обнаружение обфускации, под пропуском – не обнаружение):

* верное срабатывание – 13 файлов;
* ложное срабатывание – 1 файл;
* верный пропуск – 12 файлов;
* ложный пропуск – 0 файлов.

Метрики по выборке:

* точность – 0,93;
* полнота – 1.00;
* F-score – 0,96.

Параметры алгоритма обнаружения обфускации также подобраны так, чтобы полнота была наибольшей.

## **5. Разработка системы поиска загруженных web-шеллов**

### **5.1. Интерфейс и поведение программы**

Система представлена графическим интерфейсом, содержащим в себе:

* информацию о текущем каталоге сайта;
* таблицу с результатами анализа файлов в каталоге сайта;
* кнопку «Результаты»;
* кнопку «Сканировать»;
* кнопку «Сохранить»;
* кнопку «Установить новый».

В таблице для каждого файла (приведено его имя) указаны следующие характеристики:

* доверенный он или нет;
* кнопка-флажок для помечивания доверенных файлов;
* наличие преобразования (в том числе и обфускации);
* подозрительные фрагменты кода, их тип и уровень опасности.

По умолчанию все файлы считаются не доверенными, то есть небезопасными. Решение о том, безопасен ли файл, принимает пользователь. При желании, он может пометить файл как безопасный, установив соответствующий флажок.

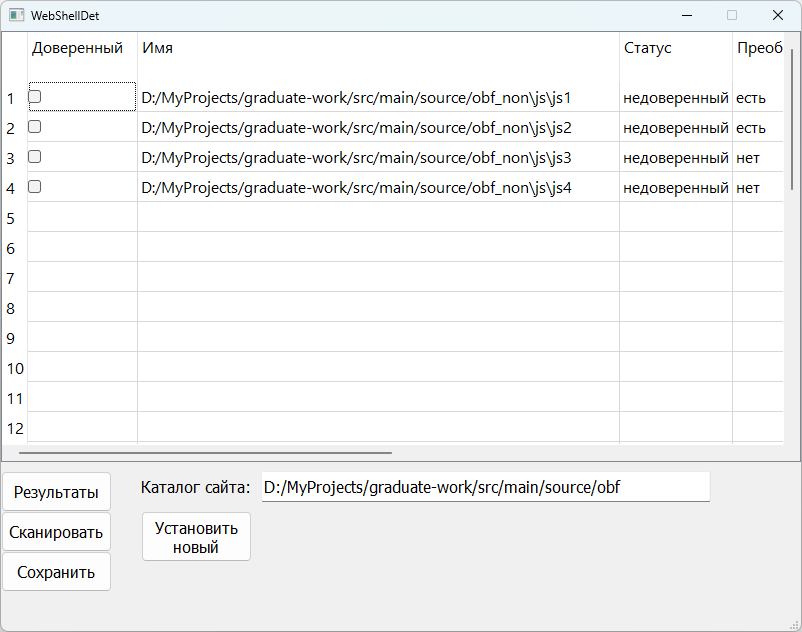


Рисунок 5.1.1 Интерфейс программы, вид столбцов «Доверенный», «Имя» и «Статус»

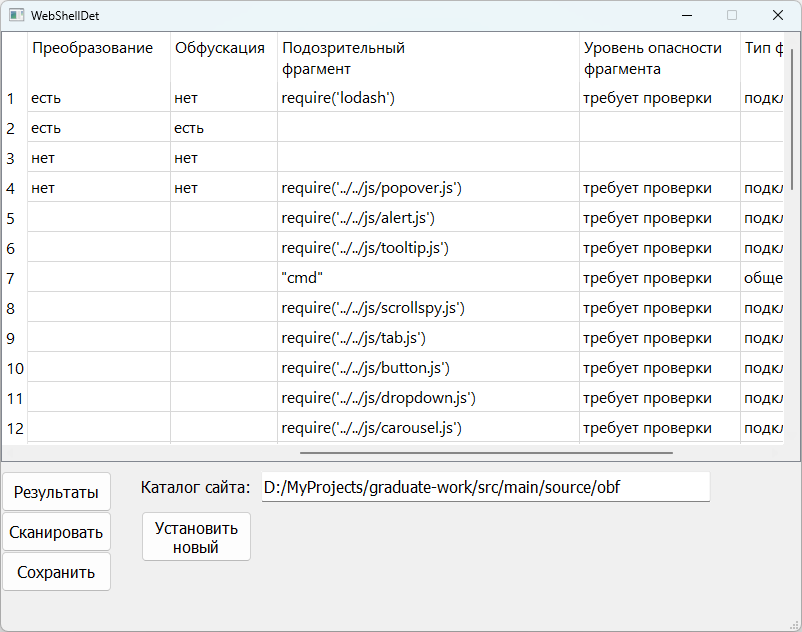


Рисунок 5.1.2 Интерфейс программы, вид столбцов «Преобразование», «Обфускация», «Подозрительный фрагмент» и «Уровень опасности фрагмента»

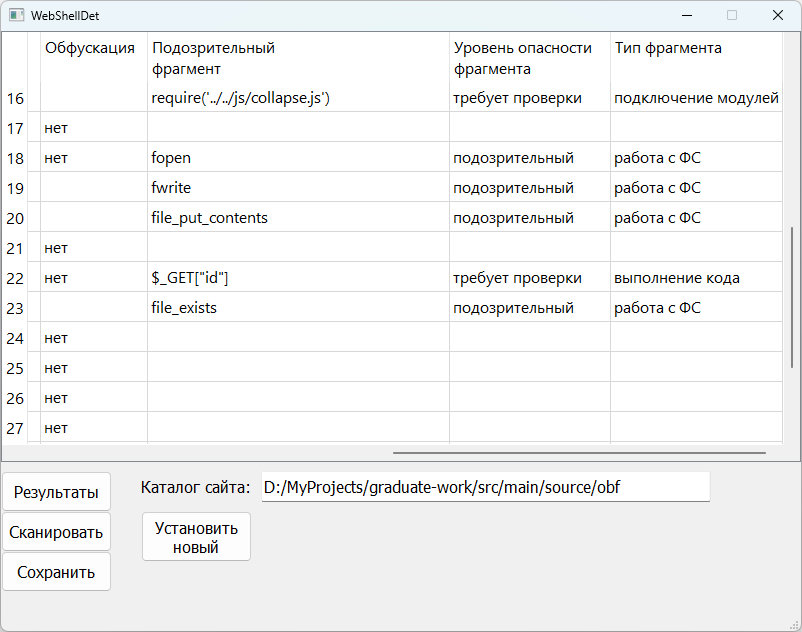


Рисунок 5.1.3 Интерфейс программы, вид столбцов «Обфускация», «Подозрительный фрагмент», «Уровень опасности фрагмента» и «Тип фрагмента»

При нажатии на кнопку «Результаты» в таблице отобразятся последние сохранённые результаты анализа файлов.

При нажатии на кнопку «Сканировать» программа начнёт процесс сканирования всех файлов (в том числе и вложенных) в каталоге сайта. По окончании процесса результаты анализа файлов будут отображены в таблице.

При нажатии на кнопку «Сохранить» программа сохранит текущие результаты анализа из таблицы в базу данных. После этого именно эти результаты будут выдаваться при нажатии на кнопку «Результаты».

При нажатии на кнопку «Установить новый» появится диалоговое окно, в котором пользователь должен выбрать каталог сайта.

### **5.2. Анализатор файла**

Задача анализатора файла – провести анализ его содержимого на наличие в нём подозрительных фрагментов, а также обнаружить преобразование (в том числе и обфускацию).

В своей работе анализатор (представлен классом *Analyzer*) использует уже описанные ранее классы *EncryptionDeterminator*, *ObfuscationDeterminator* и *SuspySearcher*, которые возвращают результаты своей работы в виде экземпляров классов *EncrAnalyzeResult*, *ObfuscationResult* и *SuspiciousCode* соответственно. Результат анализа файла (представлен классом *AnalysisResult*) формируется посредством объединения этих результатов.

### **5.3. Сверка хешей**

Прежде чем провести анализ очередного файла, для него вычисляется хеш. Если с момента последней проверки хеш не изменился, значит, не изменился и файл, поэтому старые результаты анализа всё ещё актуальны. В случае, если хеши не совпали или ранее хеш не вычислялся (файл обрабатывается впервые), проводится анализ файла. По умолчанию хеш считается по алгоритму SHA-256. Изменив файл конфигурации (файл настроек), можно настроить вычисление через алгоритм MD-5.

За вычисление хеша отвечает метод *calc\_hash* класса *Hasher*, результат вычисления оборачивается в класс *HashResult*.

### **5.4. Использование базы данных**

Результаты анализа файлов, а также указания о том, какие из них считаются [пользователем] безопасными, хранятся в локальной базе данных. Ввиду небольшого объёма хранимой информации, а также простоты использования, в качестве СУБД была выбрана SQLite. SQLite – это библиотека на языке C, которая реализует небольшой, быстрый, автономный, высоконадежный, полнофункциональный компонент SQL database engine [8].

#### **5.4.1. Сущности**

В БД отражены три сущности: *ScanResult*, *EncrResult* и *SuspyResult*.

*ScanResult* содержит общую информацию о результате анализа файла и включает в себя:

1. идентификатор результата;
2. имя файла;
3. хеш файла;
4. статус файла (доверенный или не доверенный);
5. информацию о наличии в файле преобразованных или подозрительных фрагментов.

*EncrResult* содержит информацию о преобразованиях в файле и включает в себя:

1. идентификатор результата;
2. имя файла;
3. информацию о наличии в файле высокоэнтропийных областей или содержании в нём подозрительного количества hex-кода.

*SuspyResult* содержит информацию о подозрительных фрагментах в файле и включает в себя:

1. идентификатор результата;
2. идентификатор каждого подозрительного фрагмента;
3. имя файла;
4. уровень опасности фрагмента;
5. тип подозрительного фрагмента;
6. непосредственно фрагмент.

#### **5.4.2. Получение и запись данных**

API для чтения/записи из/в БД представлено классом *DBService*. Все сущности, полученные из БД, инкапсулируются в классе *ResultFromDB*. Класс *ConverterForDB* переводит данные из типа *ResultFromDB* в тип *ResultOfFileAnalysis*. Класс *ResultOfFileAnalysis* содержит в себе следующую информацию:

1. имя файла;
2. результат анализа (*AnalysisResult*)
3. хеш;
4. статус.

Класс *ResultOfFileAnalysis* используется в работе класса *MainService*.

### **5.5. Основной сервис**

Класс *MainService* представляет собой точку сбора всей внутренне логики программы. Класс содержит следующие основные методы:

* *root\_dir* – свойство, хранит в себе путь к каталогу сайта);
* *get\_results\_from\_db* – использует описанные выше классы *DBService* и *ConverterForDB* для получения данных из БД и возвращает их в виде *ResultOfFileAnalysis*;
* *mark\_files\_as\_trusted* – помечает указанные файлы как «доверенные» и обновляет (актуализирует) БД;
* *run* – производит анализ каждого файла из *root\_dir*, используя метод *\_make\_new\_result\_for\_db*, после чего записывает результаты анализа в БД;
* *\_make\_new\_result\_for\_db* – формирует результат (*ResultOfFileAnalysis*) анализа файла.

На основе информации из БД и текущего хеша файла метод *\_make\_new\_result\_for\_db* может принять одно из следующих решений:

* для нового файла (соответствующей записи в БД) провести анализ, записать его результат вместе с текущим хешом и пометкой «не доверенный»;
* для изменённого файла (в БД есть соответствующая запись, однако хеш из БД не совпадает с текущим хешом) провести анализ, записать его результат вместе с текущим хешом и пометкой «модифицированный»;
* для неизменённого файла (в БД есть соответствующая запись, хеш из БД совпадает с текущим хешом) не предпринимать никаких действий, сохранить прежними хеш, результат анализа и статус из БД.

### **5.6. Настройка программы**

Все настройки программы вынесены в конфигурационный файл settings.json, откуда при запуске считываются параметры для корректной работы алгоритмов, пути к файлам БД и словаря слов. Для настроек были созданы классы-обёртки (например, *ObfuscationDetSettings* хранит настройки для класса *ObfuscationDeterminator*).

Также в настройках указан режим работы алгоритма обнаружения преобразований посредством поиска высокоэнтропийных областей в коде. Определены два режима работы алгоритма: строгий и оптимальный. В оптимальном режиме для подозрительных значений энтропии устанавливается не только нижняя граница, но и верхняя. Как уже было отмечено в разделе 3.1.5, верхняя граница необходима, чтобы исключить ложные срабатывания на высокоэнтропийных, но безопасных файлах, как, например, сжатые изображения в формате JPG. Строгий режим работы не предусматривает верхней границы энтропии.

# ЗАКЛЮЧЕНИЕ

В ходе выполнения работы были выявлены характерные особенности web-шеллов, разработаны и реализованы подходы к их обнаружению, а также представлены результаты проведения тестов и метрики, отражающие корректность работы алгоритмов обнаружения.

Результатом выполнения работы является работающая программа (в приложении указана ссылка на репозиторий проекта с исходным кодом) с графическим интерфейсом, выполненная с помощью языка программирования Python, а также с применением PyQt и SQLite, позволяющая обнаруживать подозрительные файлы, содержащие различного рода преобразования (включая обфускацию) и подозрительные фрагменты кода. Программа способна проверять исходные коды программ, выполненных на языках программирования PHP, JavaScript, Python, Ruby, C#.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. Веб-шелл // encyclopedia by Kaspersky [электронный ресурс]. URL: <https://encyclopedia.kaspersky.ru/glossary/web-shell>/ (дата обращения: 28.05.2024).
2. Исходные коды web-шеллов // GitHub [электронный ресурс на английском языке]. URL: https://github.com/alphaSeclab/awesome-webshell/ (дата обращения: 20.05.2024).
3. Шеннон К. Работы по теории информации и кибернетике. — М.: [Издательство иностранной литературы](https://ru.wikipedia.org/wiki/%D0%98%D0%B7%D0%B4%D0%B0%D1%82%D0%B5%D0%BB%D1%8C%D1%81%D1%82%D0%B2%D0%BE_%D0%B8%D0%BD%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D0%BD%D0%BE%D0%B9_%D0%BB%D0%B8%D1%82%D0%B5%D1%80%D0%B0%D1%82%D1%83%D1%80%D1%8B), 1963. — 829 с.
4. Цымбал В.П Теория информации и кодирование: Учебник. — 4-е изд., перераб. и доп. — К.: Вища шк., 1992. — 263 с.:ил.
5. [Шнайер Б.](https://ru.wikipedia.org/wiki/%D0%A8%D0%BD%D0%B0%D0%B9%D0%B5%D1%80,_%D0%91%D1%80%D1%8E%D1%81) Прикладная криптография. Протоколы, алгоритмы, исходные тексты на языке Си — М.: Триумф, 2002. — 816 с.
6. Большая российская энциклопедия: сайт / учредитель АНО «Национальный научно-образовательный центр «Большая российская энциклопедия». URL: <https://bigenc.ru/c/obfuskatsiia-fe1e3c>/ (дата обращения: 28.05.2024).
7. Романовский И.В. Дискретный анализ: Учебное пособие для студентов, специализирующихся по прикладной математике и информатике. — 4-е изд., испр. и доп. — СПб.: Невский Диалект; БХВ-Петербург, 2008. — 336  с.:ил.
8. Официальный сайт SQLite // SQLite [электронный ресурс на английском языке]. URL: <https://www.sqlite.org/> (дата обращения 05.06.2024).

# ПРИЛОЖЕНИЕ

Исходный код проекта представлен на платформе GitHub. Ссылка на репозиторий: https://github.com/GrigoryArcibashev/graduate-work/

1. Поскольку алгоритм предназначен для поиска преобразований в исходных кодах программ, разумно «вырезать» из текста только английские слова, так как в большинстве случаев исходный код либо не содержит буквенные символы каких-либо других естественных языков, либо содержит в небольшом количестве, что не должно сильно влиять на энтропию. [↑](#footnote-ref-1)