Анализ статистических свойств языков

Лабораторная работа

Ревизия: 0.3

Последняя модификация: Vlad Kovtun

Дата последней модификации: 15.09.2011 0:30:00

История изменений

20.09.2009 – Версия 0.1. Первичный документ. Владислав Ковтун

06.09.2010 – Версия 0.2. Уточнены дополнительные языки для исследования, а также более четко сформулировано задание на работу. Владислав Ковтун

14.09.2011 – Версия 0.3. Уточнено задание для исследования. Владислав Ковтун.

Содержание

История изменений 2
Содержание 3
Лабораторная работа 1. Анализ статистических свойств языков 4
- Вопросы 4
- Постановка задачи 4
  - Цель 4
  - Задачи 4
  - Вход 4
  - Пример 5
  - Вывод 5
  - Требования 6
- Методические указания для самостоятельной работы 6
  - Алгоритм 6
  - Тестирование 7
- Теоретические сведения 7
- Содержимое отчета 7

Литература 7

Лабораторная работа 1. Анализ статистических свойств языков

Вопросы

1. Общие положения.
2. Постановка задачи.
3. Методические рекомендации.

Постановка задачи

Цель

1. Разработать консольное приложение, которое формирует для заданного текстового (двоичного) файла или набора файлов, статистику появления в нем символов (байт) языка.

2. Исследовать статистические свойства языков: русский, украинский, английский, немецкий, французский, итальянский.

3. Разработать консольное приложение, которое реализует простейший алгоритм зашифровывания (простая замена или перестановка).

4. Разработать консольное приложение, которое реализует алгоритм статистического криптоанализа на основе информации о статистических свойствах языка.

5. Исследовать подходы к распознаванию языка по его статистическим свойствам, а также рассмотреть особенности реализации алгоритма статистического криптоанализа.

Задачи

1. Ознакомиться с основами теории информации.

2. Разработать консольное приложение для анализа статистических свойств языка. Рассмотреть вариант статистических свойств символов текста для заданного языка и вариант статистических свойств алфавита. При составлении портрета следует построить частоты появления: единичных символов, пар и троек на основе больших текстов (размер текстовой выборки не менее 10 Мбайт):

- Русский.
- Украинский.
- Английский.
- Немецкий.
- Итальянский.

Французский

3. Разработать консольное приложение для зашифровывания текстового файла посредством операции XOR с некоторой константой.

4. Провести анализ статистических свойств зашифрованного и исходного файла, на основе результатов анализа произвести распознавание языка по его статистическому портрету и произвести замену зашифрованных символов на соответствующие им символы языка с заданными статистическими свойствами.

5. Оформить отчет.

Вход

Для управления приложением анализа предлагается использовать ключи командной строки:

- /id:<filedir> – полный путь к директории, в которой хранятся файлы-книги для анализа. Данный параметр является обязательным, если он не указан, то происходит вывод на экран соответствующего сообщения и подсказки по использованию данного приложения.
- /if:<filenames> - перечисление имен файлов-книг через пробел, которые необходимо проанализировать. Файлы обязаны, находится в директории

указанной с помощью ключа /id. Данный ключ позволяет осуществлять выборочный анализ файлов-книг.

- /o:<filename> – полный путь к файлу, либо имени файла, который будет хранить сформированную статистику. Если данный параметр не указывается, то вывод производится на экран.
- /? - вывод информации о допустимых ключах командной строки.

Для управления приложением зашифровывания, предлагается использовать ключи командной строки:

- /if:<filename> - полный путь к файлу, либо имени файла –книги, который необходимо зашифровать. Данный ключ позволяет осуществлять выборочный анализ файлов-книг.
- /k:<key> - число-ключ, в диапазоне от 1 до 255. Используется для сложения по модулю 2 с символом книги.
- /o:<filename> – полный путь к файлу, либо имени файла, который будет хранить зашифрованный файл.

/? - вывод информации о допустимых ключах командной строки.

Для управления приложением расшифровывания (криптоанализа), предлагается использовать ключи командной строки:

- /if:<filename> - полный путь к файлу, либо имени файла, который необходимо зашифровать.
- /s:<filename> - полный путь к файлу, либо имени файла, который содержит статистику языка для замены зашифрованных символов на обладающие аналогичными (близкими) статистическими свойствами.
- /se:<filename> - полный путь к файлу, либо имени файла, который содержит статистику появления байт зашифрованных символов.
- /r:<range> - действительное число из диапазона $0,1$, которое хранит допустимое отклонение частоты появления исходного символа от зашифрованного. Подбирается эмпирическим путем.
- /o:<filename> – полный путь к файлу, либо имени файла, который будет хранить восстановленный, в результате криптоанализа, файл.

Пример

Анализ текстового файла war-n-peace.txt, который содержит книгу «Война и мир», результат анализа выводится в файл statistics.dat.

C:/>analyser.exe /id:c:\data /if:war-n-peace_vol1.txt war-n-peace_vol2.txt war-n-peace_vol3.txt /o:statistics.dat

Зашифровывание файла war-n-peace.txt, который содержит книгу «Война и мир», результат зашифровывания ключом 123, выводится в файл war-n-peace.enc.

C:/>encrypter.exe /if:war-n-peace.txt /k:123 /o: war-n-peace.enc

Криптоанализ файла war-n-peace.enc, который содержит зашифрованную книгу «Война и мир». Для криптоанализа используется файл статистики языка statistics.dat, а также файл статистики зашифрованного файла stat_enc.dat, допустимый диапазон отклонения частоты зашифрованного символа от содержащегося в файле статистики 0,001, результат криптоанализа выводится в war-n-peace.dec.

C:/>cryptoanalyser.exe /if:war-n-peace.enc /s:statistics.dat /se:stat_enc.dat /r:0,001 /o: war-n-peace.dec

Вывод

Во время работы приложения, рекомендуется выводить информацию о статусе приложения, например % проанализированной информации, а также о корректности его работы на консоль.

Допускается перенаправление вывода консоли в текстовый файл, например:

C:/>analyser.exe /id:c:\data /if:war-n-peace_vol1.txt war-n-peace_vol2.txt war-n-peace_vol3.txt /o:statistics.dat >report.txt

Требования

- Архитектура приложения строится по модульному принципу.
- За основу принимается стандартная библиотека C++ (в случае разработки на C++).
- Рекомендуется использовать защищенные ресурсы и указатели.
- Во время работы приложения обязательным является отображение информации о статусе приложения.

Методические указания для самостоятельной работы

При подготовке к лабораторной работе необходимо:

- Ознакомиться с функциями для работы с файлами.
- Ознакомиться с функциями стандартной библиотеки.

Алгоритм

Кратко опишем алгоритм приложения анализа статистических свойств языка:

1. Получение размера файла (файлов) и вычисление объема работ.

2. Инициализация счетчиков встречи символов (байт).

3. Цикл чтения файлов с данными. Чтение файла и проверка его на корректность.

4. Цикл чтения блоков байт файла, например размер 4096 байт, из файла. Последовательный анализ каждого блока и увеличение соответствующих счетчиков.

5. Закрытие файла с входными данными. Проверка на корректность.

6. Создание файла с результирующими данными. Проверка на корректность.

7. Запись в результирующий файл значений символов и количество их встреч, а также частоту их появления в данном тексте.

8. Закрытие файлов с выходными данными. Проверка на корректность.

Алгоритм зашифровывания текстового файла:

1. Открытие исходного файла, его чтение и проверка на корректность.

2. Открытие результирующего файла, проверка его на корректность.

3. Цикл чтения блоков байт исходного файла, например размер 4096 байт. Последовательное сложение по модулю 2 каждого символа блока исходного файла с байтом-ключом. Запись сформированного блока в результирующий файл.

4. Закрытие файла с входными данными. Проверка на корректность.

5. Закрытие файла с результирующими данными. Проверка на корректность.

Алгоритм криптоанализа зашифрованного файла. Для простоты, будем рассматривать лишь частоты встречи одного символа, а двойки и тройки использовать в приложении криптоанализа не следует.

1. Открытие файла статистики исходного языка, проверка его на корректность.

2. Чтение частоты появления одиночных символов языка в память и формирование таблицы анализа.

3. Закрытие файла статистики. Проверка на корректность.

4. Открытие файла статистики зашифрованного файла, проверка его на корректность.

5. Чтение частоты появления одиночных символов языка в память и формирование таблицы анализа.

6. Закрытие файла статистики. Проверка на корректность.

7. Открытие зашифрованного файла, его чтение и проверка на корректность.

8. Открытие файла с результирующими данными. Проверка на корректность.

9. Цикл чтения блоков байт зашифрованного файла, например размер 4096 байт.

9.1. Для каждого байта блока следует произвести его поиск в таблице частот (статистика зашифрованного файла).

9.2. Произвести поиск соответствующего ему символа языка в таблице частот (статистика языка).

9.3. Записать восстановленный блок в результирующий файл.

10. Закрытие файла с входными данными. Проверка на корректность.

11. Закрытие файла с результирующими данными. Проверка на корректность.

Тестирование

Тестирование приложения осуществляется в несколько этапов:

- Разработка Unit Test в рамках проекта. Данный подход позволит проверить корректность реализации алгоритма статистического анализа текста.

Теоретические сведения

Необходимая информация по стандартной библиотеке C++, доступна в [1].

Содержимое отчета

1. Краткие теоретические сведения о статистических свойствах языков и возможности их применения для их распознавания и эффективного криптоанализа.

2. Детальное описание алгоритмов каждого приложения:

- Приложения сбора статистических свойств языка.
- Приложения для тривиального зашифровывания текстовых файлов.
- Приложение для эффективного криптоанализа.

Наличие блок-схем в Microsoft Visio – обязательно.

3. Описание программной реализации алгоритмов каждого приложения.

4. Описание проведенных экспериментов и их результатов, которые содержат статистический портрет для языка (русского, украинского, английского, французского, итальянского и немецкого) в виде частот появления:

- Единичных, двоек и троек символов текста на языке.
- Единичных, двоек и троек букв языка.

5. Описание проведенного эксперимента по зашифровыванию файла заданного языка. Сбора статистических свойств результирующего файла (содержащего криптограмму) для дальнейшего определения языка исходного языка посредством сравнения статистического портрета, с известными языками используя критерии χ^2 и критерия Пирсона.

Файл электронных таблиц Microsoft Excel с результатами вычислений согласно статистических критериев.

6. Описание проведенного эксперимента по процессу криптоанализа, с детализацией последовательности действий, полученных результатов.

7. Выводы.

Литература

1. Microsoft Developer Network. Доступно по адресу: www.msdn.com

2. Теоретические материалы доступны по адресу: www.nrjetix.com/r-and-d/lectures