Контрольное домашнее задание

Задача

(1) Разработать с использованием языка C++ программу, реализующую алгоритмы сжатия данных без потерь, для упаковки и распаковки файлов различного типа:

1. алгоритм Хаффмана (простой),
2. алгоритм Шеннона-Фано (простой),
3. алгоритм Лемпеля-Зива 77 года LZ77 (со скользящим окном),
4. алгоритм Лемпеля-Зива-Велча LZW ***Бонус-задача.** Включение данного алгоритма в исследование не обязательно. Его реализация и включение в состав экспериментального исследования позволяет получить дополнительные баллы (см. раздел Оценивание данного задания).

Форматы имен файлов:

1. исходный файл <name>
2. метод упаковки, использующий алгоритм Хаффмана, упакованный файл <name>.haff
3. метод распаковки, использующий алгоритм Хаффмана, распакованный файл <name>.unhaff
4. метод упаковки, использующий алгоритм Шеннона-Фано, архивированный файл <name>.shan
5. метод распаковки, использующий алгоритм Шеннона-Фано, разархивированный файл <name>.unshan
6. метод упаковки, использующий алгоритм LZ77 (
   - размер скользящего окна 5 Кб, размер словаря 4 Кб архивированный файл <name>.lz775,
   - размер скользящего окна 10 Кб, размер словаря 8 Кб архивированный файл <name>.lz7710,
   - размер скользящего окна 20 Кб, размер словаря 16 Кб архивированный файл <name>.lz7720,
7. метод распаковки, использующий алгоритм LZ77;
   - размер скользящего окна 5 Кб, размер словаря 4 Кб архивированный файл <name>.unlz775,
   - размер скользящего окна 10 Кб, размер словаря 8 Кб архивированный файл <name>.unlz7710,
   - размер скользящего окна 20 Кб, размер словаря 16 Кб архивированный файл <name>.unlz7720,
8. *метод упаковки, использующий алгоритм LZW, архивированный файл <name>.lzw
9. *метод распаковки, использующий алгоритм LZW, разархивированный файл <name>.unlzw.

Рекомендуем входной файл для упаковки рассматривать как поток байтов.

В файле main.cpp указать в комментариях (в самом начале файла):

```
// КДЗ по дисциплине Алгоритмы и структуры данных? 2017-2018 уч.год
// ФИО студента, группа БПИ-XXX, дата (XX.XX.2018)
// Среда разработки, состав проекта (файлы *.cpp и *.h)
// Что сделано, а что нет
```

Другие необходимые комментарии могут быть в коде.

Алгоритмы Хаффмана и Шеннона-Фано работают *в два прохода*. Сначала строится *таблица частот встречаемости символов* в конкретном упаковываемом файле. Затем строится *кодовое дерево*. По нему определяются коды символов и с их помощью упаковывается файл.
Для распаковки алгоритмам требуется знать *таблицу частот встречаемости символов*/ *кодовое дерево*, которые использовались при упаковке. Как вариант: соответствующая таблица должна сохраняться в начале упакованного файла и использоваться при распаковке. В начале пишется количество различных символов *n*, имеющихся в сжимаемом файле, а затем символы по убыванию частоты встречаемости символа в сжимаемом файле.

Алгоритм LZ77 работает *в один проход*. Используется скользящее окно для динамического построения словаря, который, в свою очередь, используется для кодирования содержимого упаковываемого файла. Материалы с описанием алгоритма LZ77 имеются в LMS.

> **Подсказка:** Вы уже разрабатывали необходимые реализации алгоритмов при выполнении еженедельных домашних заданий. С нуля разрабатывать алгоритмы придётся только в том случае, если домашние задачи не были сданы.

*****Алгоритм LZW является развитием алгоритма LZ78. Данный алгоритм был реализован всеми, кто решился выполнить бонусную задачу.

(2) Провести вычислительный эксперимент с целью оценки реализованных алгоритмов сжатия без потерь (упаковка/распаковка). Для проведения эксперимента с алгоритмами сжатия без потерь необходимо использовать набор из **36** файлов с именами “1”...”36”, выданных вместе с заданием. Рекомендуется для упаковки рассматривать исходный файл как поток байт.

Вычислить:

- *энтропию* исходных файлов; определяется общее количество различных символов m, вычисляется их частотная встречаемость w_i в файле, и энтропия файла по формуле $H = -\sum_{i=1}^{m} w_i log_2 w_i$; значения близкие к 1 характеризуют данный файл, как файл с близкой к равночастотной встречаемостью символов;
- коэффициент сжатия.

Измерить для каждого файла и для каждого алгоритма:

- время упаковки;
- время распаковки.

Время измерять в тактах ЦП или в наносекундах (как на учебной практике летом 2017 года). Для получения достоверных результатов упаковку и распаковку каждого файла каждым методом выполнить не менее 20 раз, после чего вычислить среднее время работы алгоритма. Количество экспериментальных измерений времени **не менее** (20 раз * 10 (или 12) алгоритмов * 36 файлов) = **720** (учтены алгоритмы упаковки/распаковки LZ77 с разным размером окна).

(3) Подготовить отчет по итогам работы, содержащий постановку задачи, описание алгоритмов и задействованных структур данных, описание реализации, обобщенные результаты измерения эффективности алгоритмов, описание использованных инструментов (например, если использовались скрипты автоматизации), оценку соответствия результатов экспериментальной проверки теоретическим оценкам эффективности исследуемых алгоритмов.

Отчет также должен содержать описание аппаратных средств и показатели качества архивации (коэффициент сжатия = отношение размеров выходного и входного файлов), данные о времени работы каждого алгоритма с каждым файлом из тестового набора.

В отчете необходимо ***явно*** указать, какие части задания были сделаны, а какие нет!

Указания

Результаты работы надо *загрузить в ЛМС* (проект КДЗ) в виде архива, содержащего:

1. Отчет в форматах pdf *и* doc/docx/другой исходник (обязательно включающий план эксперимента, см. ниже),
2. Исходный код проекта.

Отчет по работе

Типовое содержание отчета о работе:

1. Титульный лист
2. Оглавление
3. Постановка задачи
4. Описание алгоритмов и использованных структур данных
5. Описание плана эксперимента
6. Результаты экспериментов - таблицы и графики (подробнее см. далее в этом документе),
7. Сравнительный анализ алгоритмов
8. Заключение (выводы кратко)
9. Использованные источники

Результаты выполнения экспериментов *необходимо* оформить в виде таблицы следующего вида:

Имя файла	H*	Алгоритм Хаффмана			Алгоритм Шеннона-Фано			Алгоритм LZ77, окно 4 Кб			Алгоритм LZ77, окно 8 Кб			Алгоритм LZ77, окно 16 Кб		
		K*	tp*	tu*	K*	tp*	tu*	K*	tp*	tu*	K*	tp*	tu*	K*	tp*	tu*
1																
2																
...																

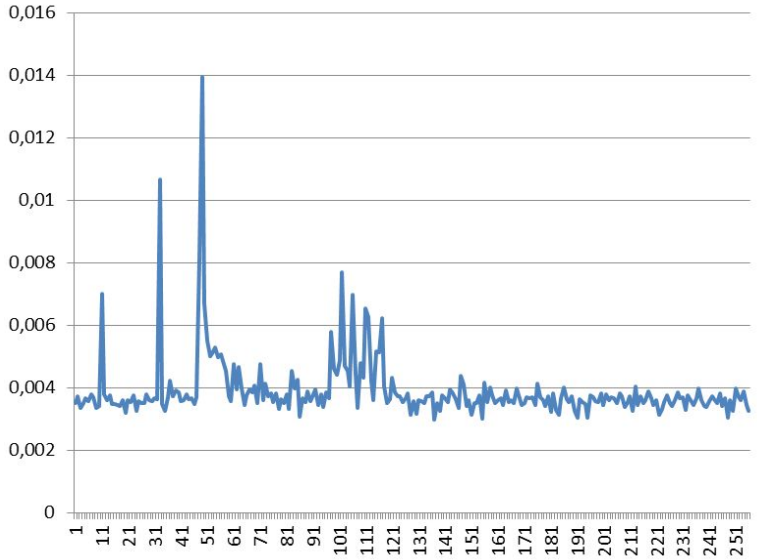
где
* H - энтропия исходного файла,
* K - коэффициент сжатия,
* tp - время упаковки (nanoseconds),
* tu - время распаковки (nanoseconds).

Каждая строка таблицы содержит результаты выполнения эксперимента для одного из файлов тестового набора (всего 36).

Таблицу *можно* оформить в альбомной ориентации (см. "разрыв разделов").

Отчет *должен содержать* следующие графики и иллюстрации.

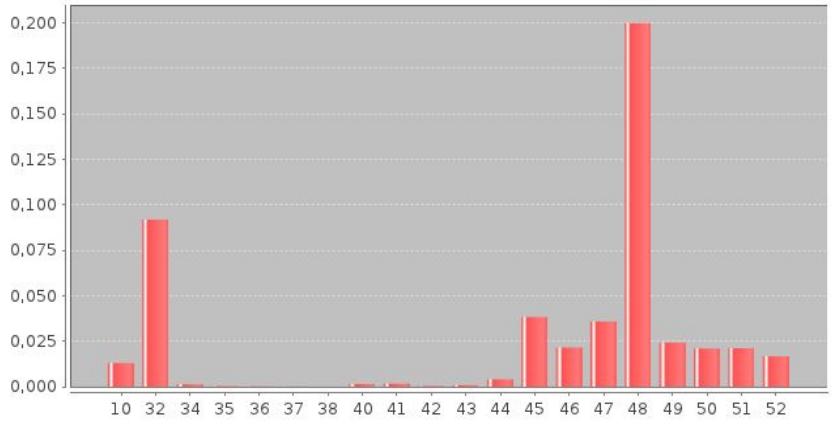
1. 36 диаграмм распределения частот встречаемости символов (байтов) для всех 36 файлов в следующем формате (по оси OX - значения байта, по оси OY - частота):



2. Столбчатые диаграммы, отражающие
   - коэффициент сжатия каждого файла для каждого алгоритма (Ось OX - номер (имя) файла, ось OY - коэффициент сжатия, легенда - название алгоритма),

- время упаковки каждого файла для каждого алгоритма,
- время распаковки каждого файла для каждого алгоритма.

Пример диаграммы, на которой приводятся данные для одного алгоритма и 52 файлов:



Приводимые диаграммы должны позволять сравнить эффективность и скорость алгоритмов на предлагаемых примерах.

Позволяют ли такие диаграммы судить о содержании (формате) тестовых файлов?

> **Подсказка:** Стиль оформления в вашей работе может отличаться от используемого в настоящем задании. Мы представляем не все необходимые графики, а только примеры. **Важно**, чтобы *оси графиков* были соответствующим образом оцифрованы и подписаны, единицы измерения должны быть указаны, приведены легенды графиков, разъясняющие смысл раскраски и начертания линий и т.д.

Приветствуются эксперименты с разными видами и способами представления информации *в дополнение к требуемым*. Можно ли наглядно показать какие-то зависимости в трех измерениях? Можно ли задействовать размер и форму точек на графиках?

Обратите внимание, что для небольших файлов с большим количеством символов *архив может быть больше исходного файла* по размеру. Почему? Ваши эксперименты подтверждают или опровергают это наблюдение?

Отчет *может* содержать дополнительные таблицы/графики, которые студент сочтет информативными и полезными в рамках задачи.

Сроки выполнения КДЗ

Нестрогий срок загрузки проекта и отчета в ЛМС — 19.03.2018 (понедельник), 13:30:00, защита работ — во время семинаров 19 и 20 марта 2018 г. Защита проводится **только** при условии загрузки отчёта по

КДЗ в ЛМС (*при сдаче и защите 19-20 марта 2018 г. полагается бонусный балл за ранее выполнение задания*).

Строгий срок загрузки проекта и отчета в ЛМС — 26.03.2018 (понедельник), 10:00:00, защита работ — 26 марта 2018 г.

Оценивание

Оценка за работу выставляется по итогам *очной защиты* проекта преподавателю.

Составляющая работы	**Балл (макс)**
Реализация упаковки/распаковки алгоритмом Хаффмана	1
Реализация упаковки/распаковки алгоритмом Шеннона-Фано	1
Реализация упаковки/распаковки алгоритмом LZ77	2
Реализация эксперимента, измерение времени работы	2
Анализ результатов, полнота отчета (описание алгоритмов и структур данных, особенностей реализации, наличие всех графиков, осмысленность выводов, дополнительный балл за лучшие решения/отчёты)	2
*Бонус-балл за раннее выполнение задания бонус за красивые решения и т.п.	1 2
*Реализация бонусного алгоритма LZW и выполнение всех элементов КДЗ для этого алгоритма (пункт не является обязательным)	+1 балл к *накопленной* оценке
Итого	**8+1+2**

Заключительные замечания

1. За один день (и даже за три дня) работу хорошо выполнить **невозможно**! Это надо иметь в виду.
2. Плагиат строго наказываем, как и всегда.

Желаем успеха!

Разработано в 2016-2018-м годах
Р. З. Ахметсафиной, А. А. Мицюком
и М. В. Ульяновым

Версия 06.03.2018-2
Изменения и исправления:
- формула энтропии, форматирование