Задание 7

Тема. Кодирование и сжатие данных методами без потерь.

Цель. Получение практических навыков и знаний по выполнению сжатия данных рассматриваемыми методами

Требуется выполнить три задания

Задание 1. Применение алгоритма группового сжатия текста

Сжать текст, используя метод RLE (run length ecoding/кодирование длин серий/групповое кодирование).

1) Описать процесс сжатия алгоритмом RLE.
2) Придумать текст, в котором есть длинные (в разумных пределах) серии из повторяющихся символов. Выполнить сжатие текста. Рассчитать коэффициент сжатия.
3) Придумать текст, в котором много неповторяющихся символов и между ними могут быть серии. Выполнить групповое сжатие, показать коэффициент сжатия. Применить алгоритм разделения текста при групповом кодировании, позволяющий повысить эффективность сжатия этого текста. Рассчитать коэффициент сжатия после применения алгоритма.
4) В отчете представьте ответы на вопросы пунктов задания с 1 по 3.

Задание 2. Исследование алгоритмов сжатия Лемпеля –Зива (LZ77), LZ78 на примерах.

Тексты для сжатия по вариантам в таб1. Столбцы 2 и 3.

1) Выполнить каждую задачу варианта, представив алгоритм решения в виде таблицы и указав результат сжатия. Примеры оформления решения представлены в Приложении1 этого документа.
2) Описать процесс восстановления сжатого текста.
3) Сформировать отчет, включив задание, вариант задания, результаты выполнения задания варианта.

Задание 3. Разработать программы (или только алгоритмы на псевлокоде или словесно) сжатия и восстановления текста методами Шеннона-Фано и Хаффмана.

1. Сформировать отчет по разработке каждого алгоритма в соответствии с требованиями.

1.1. По методу Шеннона-Фано. Данные для выполнения задания: таб.1, ваш вариант, текст столбца 1.

1) Привести постановку задачи, описать алгоритм формирования префиксного дерева и алгоритм кодирования, декодирования.
2) Представить таблицу формирования кода.
3) Изобразить префиксное дерево.
4) Рассчитать коэффициент сжатия.

1.2. По методу Хаффмана Данные для выполнения задания: ваша фамилия имя отчество.

1) Привести постановку задачи, описать алгоритм формирования префиксного дерева и алгоритм кодирования, декодирования.
2) Построить таблицу частот встречаемости символов в исходной строке для чего сформировать алфавит исходной строки и посчитать количество вхождений (частот) символов и их вероятности появления.
3) Изобразить префиксное дерево Хаффмана.
4) Упорядочить построенное дерево слева-направо (при необходимости) и изобразить его.
5) 2.8 Провести кодирование исходной строки по аналогии с примером:

```
 п     у     п     к    и    н    « »   в     а     с    и    л    и    й
1010 11000 1010 1011 00 11001 010 011 11010 11011 00 100 00 11100
« »   к    и     р    и    л    л     о     в    и    ч
010 1011 00 11101 00 100 100 11110 011 00 11111
```

6) Рассчитать коэффициенты сжатия относительно кодировки ASCII и относительно равномерного кода.
7) Рассчитать среднюю длину полученного кода и его дисперсию.
8) По результатам выполненной работы сделать выводы и сформировать отчет. Отобразить результаты выполнения всех требований, предъявленных в задании и оформить разработку программы: постановка, подход к решению, код, результаты тестирования.

1.3. Реализовать и отладить программу. Применить алгоритм Хаффмана для архивации данных текстового файла. Выполнить практическую оценку сложности алгоритма Хаффмана. Провести архивацию этого же файла любым архиватором. Сравнить коэффициенты сжатия разработанного алгоритма и архиватора.

Таблица 1. Варианты задания 1

Вариант	Закодировать фразу методами Шеннона–Фано	Сжатие данных по методу Лемпеля– Зива LZ77 Используя двухсимвольный алфавит (0, 1) закодировать следующую фразу:	Закодировать следующую фразу, используя код LZ78
	1	2	3
1	Ана, дэус, рики, паки, Дормы кормы констунтаки, Дэус дэус канадэус – бац!	0001010010101001101	кукуркукурекурекун
2	One, two, Freddy's coming for you Three, four, better lock your door Five, six, grab a crucifix Seven, eight, gonna stay up late.	0100100010010000101	упупапекапекаупуп
3	Эне-бене, рики-таки, Буль-буль-буль, Караки-шмаки Эус-деус-краснодеус бац	0100101010010000101	лорлоралоранранлоран
4	Кони-кони, коникони, Мы сидели на балконе, Чай пили, воду пили, По-турецки говорили.	0100001000000100001	пропронепронепрнепронас

5	Прибавь к ослиной голове Еще одну, получишь две. Но сколько б ни было ослов, Они и двух не свяжут слов.	10100010010101000101 1	какатанекатанекатата

6	По-турецки говорили. Чяби, чяряби Чяряби, чяби-чяби. Мы набрали в рот воды.	000101110110100111	менменаменаменатеп
7	Тише, мыши, кот на крыше, А котята ещё выше. Кот пошёл за молоком, А котята кувырком.	11010101100110000100 1	долделдолдилделдил
8	Мой котёнок очень странный, Он не хочет есть сметану, К молоку не прикасался И от рыбки отказался.	01011011011010001000 1	sarsalsarsanlasanl 33
9	Эни-бени рити-Фати. Дорба, дорба сентибрати. Дэл. Дэл. Кошка. Дэл. Фати!	00010010110010001000 1	kloklonkolonklonkl
10	Самолёт-вертолёт! Посади меня в полёт! А в полёте пусто – Выросла капуста.	11101001101100011101	tertrektekertektrek

11	Кот пошёл за молоком, А котята кувырком. Кот пришёл без молока, А котята ха-ха-ха.	10101001101100111010	bigbonebigborebigbo
12	Цветом мой зайчишка – белый, А ещё, он очень смелый! Не боится он лисицы, Льва он тоже не боится.	00010010101010011101	commercommecommerce
13	Эне, бене, лики, паки, Цуль, буль-буль, Калики-цваки, Эус-беус, кликмадеус, бокс...	01011011001010101011	webwerbweberweberweb
14	Ана-дэус-рики-паки, Дормы-кормыконсту-таки, Энус-дэус-кана-дэусБАЦ!	00101001100100000001	porpoterpoterporter
15	Раз, два – упала гора; три, четыре – прицепило; пять, шесть – бьют шерсть; семь, восемь – сено косим.	10110111100110001101	mantopmentopomantomen

16	Зуба зуба, зуба зуба, Зуба дони дони мэ, А шарли буба раз два три, А ми раз два три замри.	0100101010010000101	roporopoterropoterter
17	Плыл по морю чемодан, В чемодане был диван, На диване ехал слон. Кто не верит – выйди вон!	00010000101010011101	webwerbweberweberweb
18	Дрынцы-брынцыбубен-цы, Раз-звонилисьудальцы, Диги-диги-диги-дон, Выхо-ди-скорее-вон!	11101001101110011101	sionsinossionsinos
19	Перводан, другодан, На колоде барабан; Свистель, коростель, Пятерка, шестерка, утюг.	00010000101010011101	comconcomconacom
20	Эни бэни рики паки Турбаурбасентибряки . Может – выйдет, может – нет, В общем – полный Интернет	0100101010010000101	mantopmentopomantomen

Приложение 1. Примеры оформления выполнения заданий 2 и 3

1. Оформление задания 2

1.1. Для метода Лемпеля –Зива LZ77 для сжатия двоичного кода

Исходный текст	0000000111111111110 00000000011011110
LZ-код	0.00.100.001.011.1011.1101. 1010.0110.10010.10001.10110.
R	2 3 4
Вводимые коды	– 10 11 100 101 110 111 1000 1001 1010 1011 1100

Где LZ – сжатый текст (в данном примере в связи с небольшим размером исходного текста размер текста не уменьшился)

R отмечает шаги кодирования, после которых происходит переход на представление кодов A увеличенным числом разрядов R. Так, на первом шаге вводится код 10 для комбинации 00, и поэтому на следующих двух шагах R = 2, после третьего шага R = 3, после седьмого шага R = 4, т.е. в общем случае R = K после шага 2^{K-1} –1.

1.2. Для метода Лемпеля –Зива LZ78 для сжатия текста

В отличие от LZ77, работающего с уже полученными данными, LZ78 ориентируется на данные, которые только будут получены (LZ78 не использует скользящее окно), он хранит словарь из уже просмотренных фраз.

Пример 1. Дан текст ababaaabb. Сжать текст, используя метод LZ78
Словарь

Ссылка на символ	Символы словаря	код
1	a	<0,a>
2	b	<0,b>
3	ab	<1,b>
4	aa	<1,a>
5	abb	<3,b>

Содержимое словаря	Содержимое считанной строки	Код

a	a	<0,a>
a, b	b	<0,b>
a, b, ab	ab	<1,b>
a, b, ab ,aa	aa	<1,a>
a, b, ab ,aa, abb	abb	<3,b>

Код 0a0b1b1a3b

ababaaabb

Пример 2. Сжать текст cababababababm. Другая форма таблицы.

Содержимое словаря		Содержимое считанной строки	Код
	1	*c*	<0,c>
c,	*2*	*a*	<0,a>
c,a	*3*	*b*	<0,b>
c, a, b	*4*	*ab*	<2,b>
c, a, b, ab	*5*	*aba*	<4,a>
c, a, b, ab, aba	*6*	*ba*	<3,a>
c, a, b, ab, aba, ba	*7*	*bab*	<6,b>
c, a, b, ab, aba, ba, abab		*m*	<0,m>

Результат сжатия: 0c0a0b2b4a3a5b0m

2. Оформление отчета Задание 3

2.1. По методу Шеннона-Фано

1. Оформление таблицы метод Шеннона–Фено. Закодирована фраза «Тише, мыши, тише, кот на крыше», используя метод Шеннона–Фено.

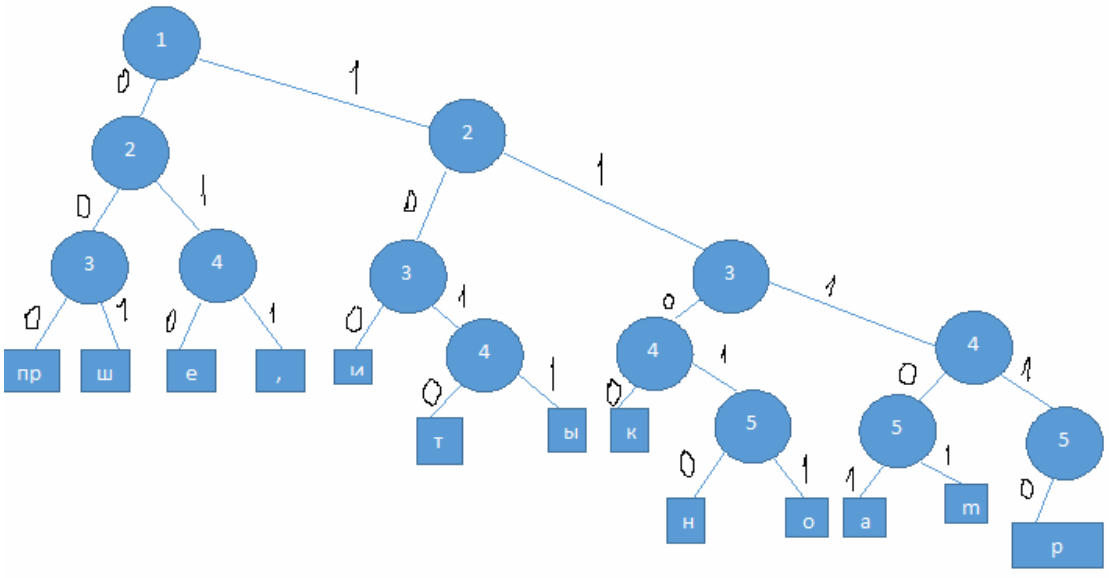
Таблица 1

Символ	Кол-во	1-я цифра	2-я цифра	3-я цифра	4-я цифра	5-я цифра	Код	Кол-во бит
пробел	5	0	0	0			000	15
ш	4	0	0	1			001	12
е	3	0	1	0			010	9
,	3	0	1	1			011	9

и	3	1	0	0			100	9
т	3	1	0	1	0		1010	12
ы	2	1	0	1	1		1011	8
к	2	1	1	1	0		1110	8
н	1	1	1	1	1		1111	4
о	1	1	1	0	0	0	11000	5
а	1	1	1	0	0	1	11001	5
м	1	1	1	0	1	0	11010	5
р	1	1	1	0	1	1	11011	5
								106

2. Оценка объема и коэффициента сжатия

Объем незакодированной фразы – 30*8 бит = 240 бит.
Объем закодированной фразы –106 бит.
Коэффициент сжатия: 106/240=0,44
Дерево префиксного кода Фано



2.2. По методу Хаффмана

Провести кодирование(сжатие) исходной строки символов «Фамилия Имя Отчество» с использованием алгоритма Хаффмана. Исходная строка символов, таким образом, определяет индивидуальный вариант задания для каждого студента.

Для выполнения задания 3 необходимо выполнить следующие действия:

Исходная строка примера **пупкин василий кириллович**

1. Таблица частот

Алфавит	п	у	к	и	н	« »	в
Кол. вх.	2	1	2	6	1	2	2
Вероятн.	0.08	0.04	0.08	0.24	0.04	0.08	0.08

Алфавит	а	с	л	й	р	о	ч
Кол. вх.	1	1	3	1	1	1	1
Вероятн.	0.04	0.04	0.12	0.04	0.04	0.04	0.04

(скобки < > обозначают пробел в исходной строке)

2. Таблица отсортированных частот

Алфавит	и	л	п	к	« »	в	у
Кол. вх.	6	3	2	2	2	2	1
Вероятн.	0.24	0.12	0.08	0.08	0.08	0.08	0.04

Алфавит	н	а	с	й	р	о	ч
Кол. вх.	1	1	1	1	1	1	1
Вероятн.	0.04	0.04	0.04	0.04	0.04	0.04	0.04

3. Построить дерево кодирования Хаффмана, в данном примере оно имеет вид:

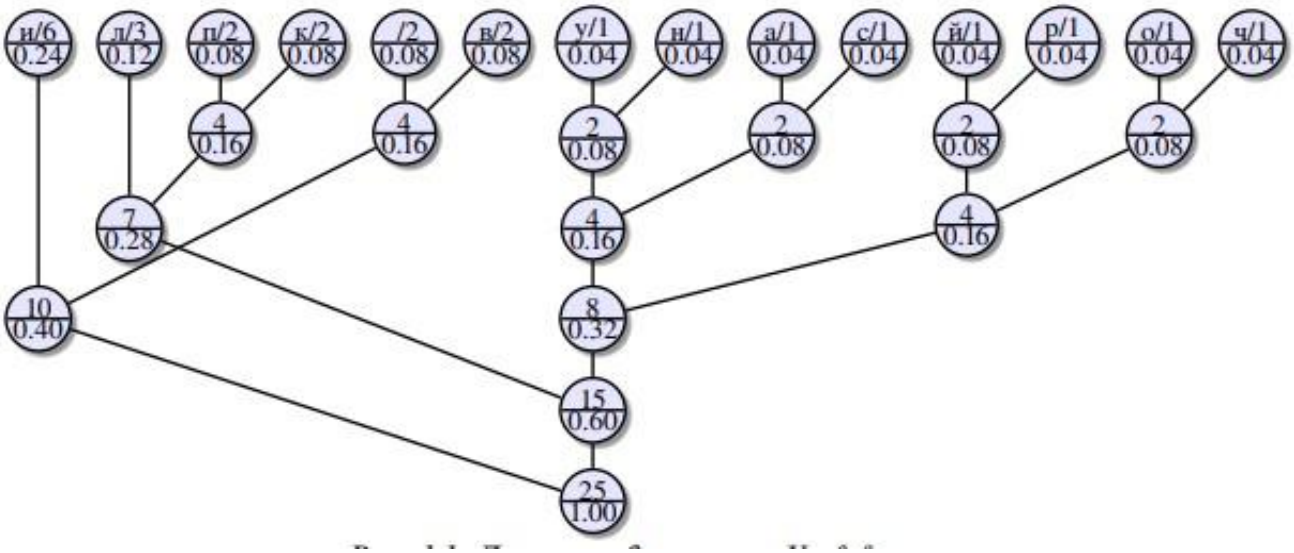


Рисунок Дерево кодирования Хаффмана (указаны частоты)

4. Упорядочить построенное дерево слева-направо (при необходимости).

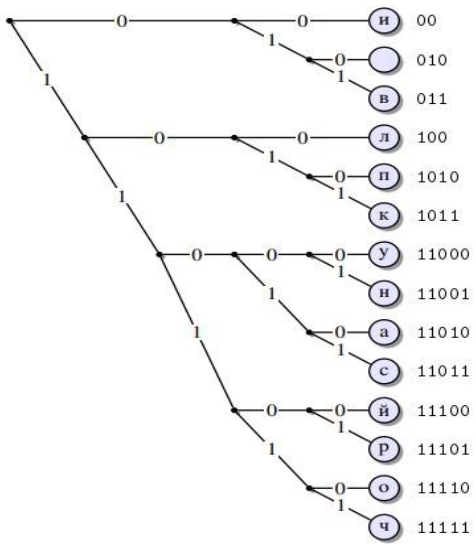


Рисунок Упорядоченное дерево кодирования Хаффмана

5. Привести таблицу с кодами символов

6. Провести кодирование исходной строки по аналогии с примером:

```
  п     у     п     к     и     н    « »    в     а     с     и     л     и     й
1010 11000 1010 1011 00 11001 010 011 11010 11011 00 100 00 11100
« »    к     и     р     и     л     л     о     в     и     ч
010 1011 00 11101 00 100 100 11110 011 00 11111
```

7. Рассчитать коэффициенты сжатия относительно кодировки ASCII и относительно равномерного кода.