Лабораторная работа №1

Дисциплина: Основы кодирования

Информация, алфавит источника, вероятность, количество информации и энтропия

Введение

Данная лабораторная работа открывает курс «Основы кодирования» и формирует фундамент для всех последующих лабораторных работ и курсового проекта. В работе последовательно рассматриваются понятия источника информации, алфавита, вероятности появления символов, количества информации и энтропии. Особое внимание уделяется влиянию выбора алфавита и структуры источника на эффективность кодирования.

Цель работы

Изучить и practically освоить методы статистического анализа источника информации, научиться вычислять энтропию и избыточность, а также подготовить исходные данные для построения оптимальных кодов (Шеннона–Фано и Хаффмана) в последующих лабораторных работах.

Теоретические сведения

1. Понятие информации

В теории информации под информацией понимают меру уменьшения неопределённости о состоянии некоторой системы или результате события. Информация всегда связана с выбором одного исхода из множества возможных.

Чем больше возможных исходов и чем менее предсказуем конкретный результат, тем больше информации он несёт.

2. Источник информации

Источник информации — это объект или процесс, который генерирует последовательность символов. Примерами источников информации являются текстовые сообщения, аудиосигналы, изображения, данные датчиков и потоки сетевых пакетов.

В рамках лабораторной работы источник информации моделируется в виде текстовой строки, сформированной из персональных данных.

3. Алфавит источника и его мощность

Алфавит источника информации — это конечное множество различных символов, которые может порождать источник.

Мощность алфавита |A| — это количество различных символов в алфавите. Мощность алфавита зависит не от теоретически возможных символов, а только от тех, которые реально присутствуют в сообщении.

4. Частота и вероятность символов

Частота n_i символа x_i — это количество появлений данного символа в сообщении. Если длина сообщения равна N, то вероятность появления символа определяется как отношение частоты к длине сообщения.

$$p_i = n_i / N$$

Сумма вероятностей всех символов алфавита всегда равна единице:

$$\Sigma p_i = 1$$

5. Количество информации

Количество информации $I(x_i)$ характеризует информативность отдельного символа. Редкие символы несут больше информации, чем часто встречающиеся.

$$I(x_i) = - \log_2(p_i)$$

6. Энтропия источника информации

Энтропия источника информации — это среднее количество информации, приходящееся на один символ сообщения.

$$H = - \Sigma p_i \cdot \log_2(p_i)$$

Энтропия характеризует степень неопределённости источника: чем выше энтропия, тем менее предсказуем источник.

7. Максимальная энтропия

Максимальная энтропия достигается в случае, когда все символы алфавита имеют одинаковые вероятности.

$$H_{max} = \log_2(|A|)$$

8. Избыточность источника информации

Если реальная энтропия источника меньше максимальной, источник обладает избыточностью. Избыточность позволяет применять методы эффективного и оптимального кодирования.

$$R = 1 - H / H_{max}$$

9. Влияние алфавита на характеристики источника

Выбор алфавита влияет на мощность алфавита, распределение вероятностей символов и, следовательно, на энтропию и избыточность источника.

Даже при неизменном содержании сообщения переход от кириллицы к латинице (транслитерация) изменяет статистические характеристики источника.

10. Отдельные источники и метаданные

В реальных информационных системах данные часто поступают от нескольких источников. Объединение источников увеличивает разнообразие символов и энтропию.

Метаданные — это служебная информация, добавляемая к основным данным. Добавление метаданных изменяет алфавит и статистику источника, что необходимо учитывать при кодировании.

11. Связь с оптимальным кодированием

Статистические характеристики источника информации лежат в основе построения оптимальных кодов. Таблица частот, полученная в лабораторной работе №1, используется для построения кодов Шеннона–Фано и Хаффмана.

Задание

Источник информации формируется на основе персональных данных студента и двух членов семьи:
– фамилия, имя, отчество (без пробелов);
– дата рождения в формате ДДММГГГГ.

Требуется:
1. Определить алфавит источника и его мощность.
2. Подсчитать частоты и вероятности символов.
3. Рассчитать количество информации каждого символа.
4. Рассчитать энтропию источника.
5. Сравнить кириллицу и латиницу (транслитерация).
6. Сравнить мощность алфавитов, энтропию и избыточность.
7. Рассматривать каждого человека как отдельный источник.
8. Рассчитать индивидуальные и объединённую энтропии.
9. Добавить служебные символы (метаданные).
10. Подготовить таблицу частот для построения кода Хаффмана.

Пример выполнения

Исходные данные для эталонного примера

Студент: АЛПАМЫССЕРИКНУРАЛЫЕВИЧ12042000

Мать: ДАНИЯАЙГУЛСАПАРОВНА05071975

Отец: ЕРНАРТАЛГАТМУХТАРОВИЧ23121970

Объединённый источник (кириллица):

АЛПАМЫССЕРИКНУРАЛЫЕВИЧ12042000ДАНИЯАЙГУЛСАПАРОВНА05071975ЕРНАРТ АЛГАТМУХТАРОВИЧ23121970

Вариант А: полный расчёт для кириллицы

N = 86, |A| = 29

Формулы:

$p_i = n_i / N$

$I(x_i) = - \log_2(p_i)$

$H = - \Sigma\, p_i \cdot \log_2(p_i)$

$H_{max} = \log_2(|A|)$

$R = 1 - H / H_{max}$

Таблица 1. Частоты и вероятности (кириллица), отсортировано по убыванию частоты

Символ	n_i	p_i	$I(x_i)$, бит	$p_i \cdot I(x_i)$
А	12	0.139535	2.841302	0.396461
0	7	0.081395	3.618910	0.294562
Р	6	0.069767	3.841302	0.267998
1	4	0.046512	4.426265	0.205873
2	4	0.046512	4.426265	0.205873
И	4	0.046512	4.426265	0.205873
Л	4	0.046512	4.426265	0.205873
Н	4	0.046512	4.426265	0.205873
7	3	0.034884	4.841302	0.168883
В	3	0.034884	4.841302	0.168883
Е	3	0.034884	4.841302	0.168883
С	3	0.034884	4.841302	0.168883
Т	3	0.034884	4.841302	0.168883

У	3	0.034884	4.841302	0.168883
5	2	0.023256	5.426265	0.126192
9	2	0.023256	5.426265	0.126192
Г	2	0.023256	5.426265	0.126192
М	2	0.023256	5.426265	0.126192
О	2	0.023256	5.426265	0.126192
П	2	0.023256	5.426265	0.126192
Ч	2	0.023256	5.426265	0.126192
Ы	2	0.023256	5.426265	0.126192
З	1	0.011628	6.426265	0.074724
4	1	0.011628	6.426265	0.074724
Д	1	0.011628	6.426265	0.074724
Й	1	0.011628	6.426265	0.074724
К	1	0.011628	6.426265	0.074724
Х	1	0.011628	6.426265	0.074724
Я	1	0.011628	6.426265	0.074724

Итого: N = 86. Проверка: $\Sigma p_i = 1.000000$ (с учётом округления).

Итоговые расчёты (кириллица):

H = 4.534286 бит/символ

Hmax = $\log_2(29)$ = 4.857981 бит/символ

R = 1 − H/Hmax = 0.066631 (то есть 6.66%)

Вариант B: полный расчёт для латиницы (транслитерация)

Латинское представление (после транслитерации):

ALPAMYSSERIKNURALYEVICH12042000DANIYAAYGULSAPAROVNA05071975ERNARTALGATMUKHTAROVICH23121970

N = 90, |A| = 27

Таблица 2. Частоты и вероятности (латиница), отсортировано по убыванию частоты

Символ	n_i	p_i	$I(x_i)$, бит	$p_i \cdot I(x_i)$
A	13	0.144444	2.791413	0.403204
0	7	0.077778	3.684498	0.286572
R	6	0.066667	3.906891	0.260459
1	4	0.044444	4.491853	0.199638
2	4	0.044444	4.491853	0.199638
I	4	0.044444	4.491853	0.199638
L	4	0.044444	4.491853	0.199638
N	4	0.044444	4.491853	0.199638
Y	4	0.044444	4.491853	0.199638
7	3	0.033333	4.906891	0.163563
E	3	0.033333	4.906891	0.163563
H	3	0.033333	4.906891	0.163563
S	3	0.033333	4.906891	0.163563
T	3	0.033333	4.906891	0.163563
U	3	0.033333	4.906891	0.163563
V	3	0.033333	4.906891	0.163563
5	2	0.022222	5.491853	0.122041
9	2	0.022222	5.491853	0.122041
C	2	0.022222	5.491853	0.122041
G	2	0.022222	5.491853	0.122041
K	2	0.022222	5.491853	0.122041
M	2	0.022222	5.491853	0.122041

O	2	0.022222	5.491853	0.122041
P	2	0.022222	5.491853	0.122041
3	1	0.011111	6.491853	0.072132
4	1	0.011111	6.491853	0.072132
D	1	0.011111	6.491853	0.072132

Итого: N = 90. Проверка: $\Sigma p_i = 1.000000$ (с учётом округления).

Итоговые расчёты (латиница):

H = 4.485729 бит/символ

$H_{max} = \log_2(27) = 4.754888$ бит/символ

$R = 1 - H/H_{max} = 0.056607$ (то есть 5.66%)

Сравнение кириллица vs латиница

Параметр	Кириллица	Латиница
Длина сообщения N	86	90
Мощность алфавита |A|	29	27
Энтропия H (бит/символ)	4.534286	4.485729
Макс. энтропия Hmax (бит/символ)	4.857981	4.754888
Избыточность R	0.066631 (6.66%)	0.056607 (5.66%)

Вывод: выбор алфавита влияет на |A|, Hmax и R. При транслитерации часть букв разворачивается в 2 символа (например, Ч→СН, Х→КН), что меняет длину сообщения и распределение частот.

Вариант C: отдельные источники + объединение

Источник	N	|A|	H (бит/символ)
Студент	30	18	4.015061
Мать	27	19	4.028605
Отец	29	20	4.116265

Таблицы частот по каждому источнику приведены ниже.

Таблица 3. Частоты и вероятности (Студент)

Символ	n_i	p_i	$I(x_i)$, бит	$p_i \cdot I(x_i)$
0	4	0.133333	2.906891	0.387585
А	3	0.100000	3.321928	0.332193
2	2	0.066667	3.906891	0.260459
Е	2	0.066667	3.906891	0.260459
И	2	0.066667	3.906891	0.260459
Л	2	0.066667	3.906891	0.260459
Р	2	0.066667	3.906891	0.260459
С	2	0.066667	3.906891	0.260459
Ы	2	0.066667	3.906891	0.260459
1	1	0.033333	4.906891	0.163563
4	1	0.033333	4.906891	0.163563
В	1	0.033333	4.906891	0.163563
К	1	0.033333	4.906891	0.163563
М	1	0.033333	4.906891	0.163563
Н	1	0.033333	4.906891	0.163563
П	1	0.033333	4.906891	0.163563
У	1	0.033333	4.906891	0.163563
Ч	1	0.033333	4.906891	0.163563

Итого: N = 30. Проверка: $\Sigma p_i = 1.000000$ (с учётом округления).

Таблица 4. Частоты и вероятности (Мать)

Символ	n_i	p_i	$I(x_i)$, бит	$p_i \cdot I(x_i)$
А	5	0.185185	2.432959	0.450548
0	2	0.074074	3.754888	0.278140

5	2	0.074074	3.754888	0.278140
7	2	0.074074	3.754888	0.278140
Н	2	0.074074	3.754888	0.278140
1	1	0.037037	4.754888	0.176107
9	1	0.037037	4.754888	0.176107
В	1	0.037037	4.754888	0.176107
Г	1	0.037037	4.754888	0.176107
Д	1	0.037037	4.754888	0.176107
И	1	0.037037	4.754888	0.176107
Й	1	0.037037	4.754888	0.176107
Л	1	0.037037	4.754888	0.176107
О	1	0.037037	4.754888	0.176107
П	1	0.037037	4.754888	0.176107
Р	1	0.037037	4.754888	0.176107
С	1	0.037037	4.754888	0.176107
У	1	0.037037	4.754888	0.176107
Я	1	0.037037	4.754888	0.176107

Итого: N = 27. Проверка: $\Sigma p_i = 1.000000$ (с учётом округления).

Таблица 5. Частоты и вероятности (Отец)

Символ	n_i	p_i	$I(x_i)$, бит	$p_i \cdot I(x_i)$
А	4	0.137931	2.857981	0.394204
Р	3	0.103448	3.273018	0.338588
Т	3	0.103448	3.273018	0.338588
1	2	0.068966	3.857981	0.266068
2	2	0.068966	3.857981	0.266068

0	1	0.034483	4.857981	0.167517
3	1	0.034483	4.857981	0.167517
7	1	0.034483	4.857981	0.167517
9	1	0.034483	4.857981	0.167517
В	1	0.034483	4.857981	0.167517
Г	1	0.034483	4.857981	0.167517
Е	1	0.034483	4.857981	0.167517
И	1	0.034483	4.857981	0.167517
Л	1	0.034483	4.857981	0.167517
М	1	0.034483	4.857981	0.167517
Н	1	0.034483	4.857981	0.167517
О	1	0.034483	4.857981	0.167517
У	1	0.034483	4.857981	0.167517
Х	1	0.034483	4.857981	0.167517
Ч	1	0.034483	4.857981	0.167517

Итого: N = 29. Проверка: $\Sigma p_i = 1.000000$ (с учётом округления).

Объединённый источник (без метаданных):

Нобщ = 4.534286 бит/символ при N = 86 и |A| = 29.

Метаданные (служебные символы)

Источник с метаданными:

<S>АЛПАМЫССЕРИКНУРАЛЫЕВИЧ12042000</S><M>ДАНИЯАЙГУЛСАПАРОВНА05071975</M><F>ЕРНАРТАЛГАТМУХТАРОВИЧ23121970</F>

N = 107, |A| = 35

Таблица 6. Частоты и вероятности (с метаданными)

Символ	n_i	p_i	$I(x_i)$, бит	$p_i \cdot I(x_i)$
А	12	0.112150	3.156504	0.354001

0	7	0.065421	3.934112	0.257372
<	6	0.056075	4.156504	0.233075
>	6	0.056075	4.156504	0.233075
Р	6	0.056075	4.156504	0.233075
1	4	0.037383	4.741467	0.177251
2	4	0.037383	4.741467	0.177251
И	4	0.037383	4.741467	0.177251
Л	4	0.037383	4.741467	0.177251
Н	4	0.037383	4.741467	0.177251
/	3	0.028037	5.156504	0.144575
7	3	0.028037	5.156504	0.144575
В	3	0.028037	5.156504	0.144575
Е	3	0.028037	5.156504	0.144575
С	3	0.028037	5.156504	0.144575
Т	3	0.028037	5.156504	0.144575
У	3	0.028037	5.156504	0.144575
5	2	0.018692	5.741467	0.107317
9	2	0.018692	5.741467	0.107317
F	2	0.018692	5.741467	0.107317
М	2	0.018692	5.741467	0.107317
S	2	0.018692	5.741467	0.107317
Г	2	0.018692	5.741467	0.107317
М	2	0.018692	5.741467	0.107317
О	2	0.018692	5.741467	0.107317
П	2	0.018692	5.741467	0.107317

Ч	2	0.018692	5.741467	0.107317
Ы	2	0.018692	5.741467	0.107317
З	1	0.009346	6.741467	0.063004
4	1	0.009346	6.741467	0.063004
Д	1	0.009346	6.741467	0.063004
Й	1	0.009346	6.741467	0.063004
К	1	0.009346	6.741467	0.063004
Х	1	0.009346	6.741467	0.063004
Я	1	0.009346	6.741467	0.063004

Итого: N = 107. Проверка: $\Sigma p_i = 1.000000$ (с учётом округления).

Итоговые расчёты (с метаданными):

H = 4.830396 бит/символ

Hmax = $\log_2(35)$ = 5.129283 бит/символ

R = 0.058271 (то есть 5.83%)

Подготовка данных для кода Хаффмана (входная таблица)

Ниже приведена таблица символов объединённого источника (кириллица, без метаданных), отсортированная по возрастанию частоты. Она используется как входные данные для ЛР №6 (Хаффман).

Таблица 7. Входные данные для Хаффмана (сортировка по возрастанию n_i)

Символ	n_i	p_i	$I(x_i)$, бит	$p_i \cdot I(x_i)$
З	1	0.011628	6.426265	0.074724
4	1	0.011628	6.426265	0.074724
Д	1	0.011628	6.426265	0.074724
Й	1	0.011628	6.426265	0.074724
К	1	0.011628	6.426265	0.074724
Х	1	0.011628	6.426265	0.074724

Я	1	0.011628	6.426265	0.074724
5	2	0.023256	5.426265	0.126192
9	2	0.023256	5.426265	0.126192
Г	2	0.023256	5.426265	0.126192
М	2	0.023256	5.426265	0.126192
О	2	0.023256	5.426265	0.126192
П	2	0.023256	5.426265	0.126192
Ч	2	0.023256	5.426265	0.126192
Ы	2	0.023256	5.426265	0.126192
7	3	0.034884	4.841302	0.168883
В	3	0.034884	4.841302	0.168883
Е	3	0.034884	4.841302	0.168883
С	3	0.034884	4.841302	0.168883
Т	3	0.034884	4.841302	0.168883
У	3	0.034884	4.841302	0.168883
1	4	0.046512	4.426265	0.205873
2	4	0.046512	4.426265	0.205873
И	4	0.046512	4.426265	0.205873
Л	4	0.046512	4.426265	0.205873
Н	4	0.046512	4.426265	0.205873
Р	6	0.069767	3.841302	0.267998
0	7	0.081395	3.618910	0.294562
А	12	0.139535	2.841302	0.396461

Итого: N = 86. Проверка: $\Sigma p_i = 1.000000$ (с учётом округления).

Пример последовательно используется для анализа базового варианта, сравнения алфавитов, анализа отдельных источников и добавления метаданных.

Требования к программе (Python)

Программа должна:

1) принимать исходные строки (ФИО+дата) для 3 человек;
2) формировать объединённый источник;
3) строить таблицу символов: n_i, p_i, $I(x_i)$, $p_i \cdot I(x_i)$;
4) вычислять H, Hmax, R;
5) выполнять транслитерацию и повторять расчёты для латиницы;
6) вычислять энтропию каждого отдельного источника;
7) добавлять метаданные и повторять расчёт;
8) формировать входную таблицу для Хаффмана (сортировка по n_i).

Необходимо развивать код из ЛР1–ЛР2 как единый проект (рекомендуется модульная структура и функции).

Запрещается использовать готовые функции расчёта энтропии и кодов.

Требования к отчёту

Отчёт должен содержать:

1. Титульный лист.
2. Цель и задание.
3. Теоретические формулы.
4. Исходные данные.
5. Таблицы частот и вероятностей.
6. Расчёт энтропии и избыточности.
7. Сравнение алфавитов.
8. Выводы.
9. Листинг программы.

Контрольные вопросы

1. Что такое источник информации?
2. Что такое алфавит источника и его мощность |A|?
3. Как определяется мощность алфавита?
4. Как интерпретировать вероятность p_i и частоту n_i?
5. Что такое энтропия H и как она вычисляется?
6. Что такое Hmax и когда она достигается?
7. Почему энтропия может быть меньше максимальной?
8. Как выбор алфавита влияет на энтропию?
9. Почему при транслитерации меняются N и распределение частот?
10. Что такое избыточность источника?
11. Зачем добавляются метаданные?