ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ
БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ПРОФЕССИОНАЛЬНОГООБРАЗОВАНИЯ

«СИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ТЕЛЕКОММУНИКАЦИЙ И ИНФОРМАТИКИ»

Теория информации

Практическая работа №3

«Побуквенное кодирование текстов»

Выполнил:
студент гр. ИП-911
Мироненко К.А.

Проверила:
доцент кафедры ПМиК
Мачикина Е.П.

Новосибирск
2022-2023 уч.год

Результат работы

```
PS C:\Users\User\Desktop\oh shit, here we go again\8 семестр\Information_theory\labs\lab3> python .\lab3.py
Текст:  1984 - George Orwell. Русский язык. Часть первая
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Размер алфавита: 43
H0 =log2(43) = 5.426264754702098

Энтропия оригинального текста: 4.383726570658277

Шенон:
 : 0.1579 - 000
о: 0.0937 - 0010
е: 0.0693 - 0110
а: 0.0654 - 0101
и: 0.0606 - 01111
н: 0.0600 - 01110
т: 0.0542 - 10010
с: 0.0457 - 10001
л: 0.0429 - 10011
р: 0.0368 - 10100
в: 0.0365 - 10101
м: 0.0271 - 101110
к: 0.0267 - 110010
д: 0.0253 - 110001
у: 0.0239 - 110011
п: 0.0237 - 110100
ы: 0.0177 - 110110
ь: 0.0153 - 1101111
я: 0.0151 - 1111001
г: 0.0146 - 1110010
з: 0.0145 - 1110100
б: 0.0141 - 1110110
ч: 0.0120 - 1111111
й: 0.0091 - 1111001
ж: 0.0084 - 1111011
х: 0.0072 - 11111110
ш: 0.0065 - 11111010
ю: 0.0043 - 11111011
ц: 0.0035 - 111111010
щ: 0.0032 - 111111011
```

```
7: 0.0001 - 111111111111101
6: 0.0000 - 111111111111110
Средняя длина кодовых слов (L ср.): 4.913257051140997
--
Текст:  coded_shanon.txt
Размер алфавита: 2
H0 =log2(2) = 1.0

Энтропия для 1 символа(ов) подряд: 0.9990161977934344
Энтропия для 2 символа(ов) подряд: 0.9989806888404302
Энтропия для 3 символа(ов) подряд: 0.9974138678154706



Хаффман:
 : 0.1579 - 110
о: 0.0937 - 000
е: 0.0693 - 1011
а: 0.0654 - 1001
и: 0.0606 - 1000
н: 0.0600 - 0111
т: 0.0542 - 0101
с: 0.0457 - 11111
л: 0.0429 - 11110
р: 0.0368 - 11101
в: 0.0365 - 11100
м: 0.0271 - 01001
к: 0.0267 - 01000
д: 0.0253 - 00111
у: 0.0239 - 00101
п: 0.0237 - 00100
```

```
ь: 0.0153 - 101000
я: 0.0151 - 011011
г: 0.0146 - 011010
з: 0.0145 - 011001
б: 0.0141 - 011000
ч: 0.0120 - 001100
й: 0.0091 - 1010110
ж: 0.0084 - 1010011
х: 0.0072 - 1010010
ш: 0.0065 - 0011010
ю: 0.0043 - 10101110
ц: 0.0035 - 00110111
щ: 0.0032 - 00110110
э: 0.0023 - 101011110
ф: 0.0014 - 1010111111
1: 0.0002 - 1010111110111
ъ: 0.0002 - 1010111110101
4: 0.0001 - 1010111110100
0: 0.0001 - 1010111110011
3: 0.0001 - 1010111110010
9: 0.0001 - 1010111110001
8: 0.0001 - 10101111101101
2: 0.0001 - 10101111100001
5: 0.0001 - 10101111100000
7: 0.0001 - 101011111011001
6: 0.0000 - 101011111011000
Средняя длина кодовых слов (L ср.): 4.411975613860233
--
Текст:  coded_huffman.txt
Размер алфавита: 2
H0 =log2(2) = 1.0

Энтропия для 1 символа(ов) подряд: 0.9959960313994722
Энтропия для 2 символа(ов) подряд: 0.9959625925236723
Энтропия для 3 символа(ов) подряд: 0.99567284022491
PS C:\Users\User\Desktop\oh shit, here we go again\8 семестр\Information_theory\labs\lab3>
```

Метод кодирования	Название текста	Оценка избыточности кодирования	Оценка энтропии выходной последовательности (частоты символов)	Оценка энтропии выходной последовательности (частоты пар символов)	Оценка энтропии выходной последовательности (частоты троек символов)
Код Шеннона	1984 — George Orwell. Русский язык. Часть первая	0.5295305	0.9990161	0.9989806	0.9974138
Код Хаффмана	1984 — George Orwell. Русский язык. Часть первая	0.0274710	0.9959960	0.9959625	0.9956728

При кодированиях были получены префиксные код, в которых используется избыточность сообщения (коды более частых символов состоят из коротких последовательностей, а коды более редких символов – из более длинных).

Можно увидеть, что данные методы кодирования обладают высокой избыточностью. Энтропия полученных последовательностей близка к единице, что говорит о том, что на один символ приходится один бит информации. При том, при выборе пар или троек символов энтропия почти не меняется. Это говорит о том, что символы в получившихся кодах равновероятны, что подтверждает эффективность кодирования.