ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ
БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ПРОФЕССИОНАЛЬНОГООБРАЗОВАНИЯ

«СИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ТЕЛЕКОММУНИКАЦИЙ И ИНФОРМАТИКИ»

Теория информации

Практическая работа №1-2

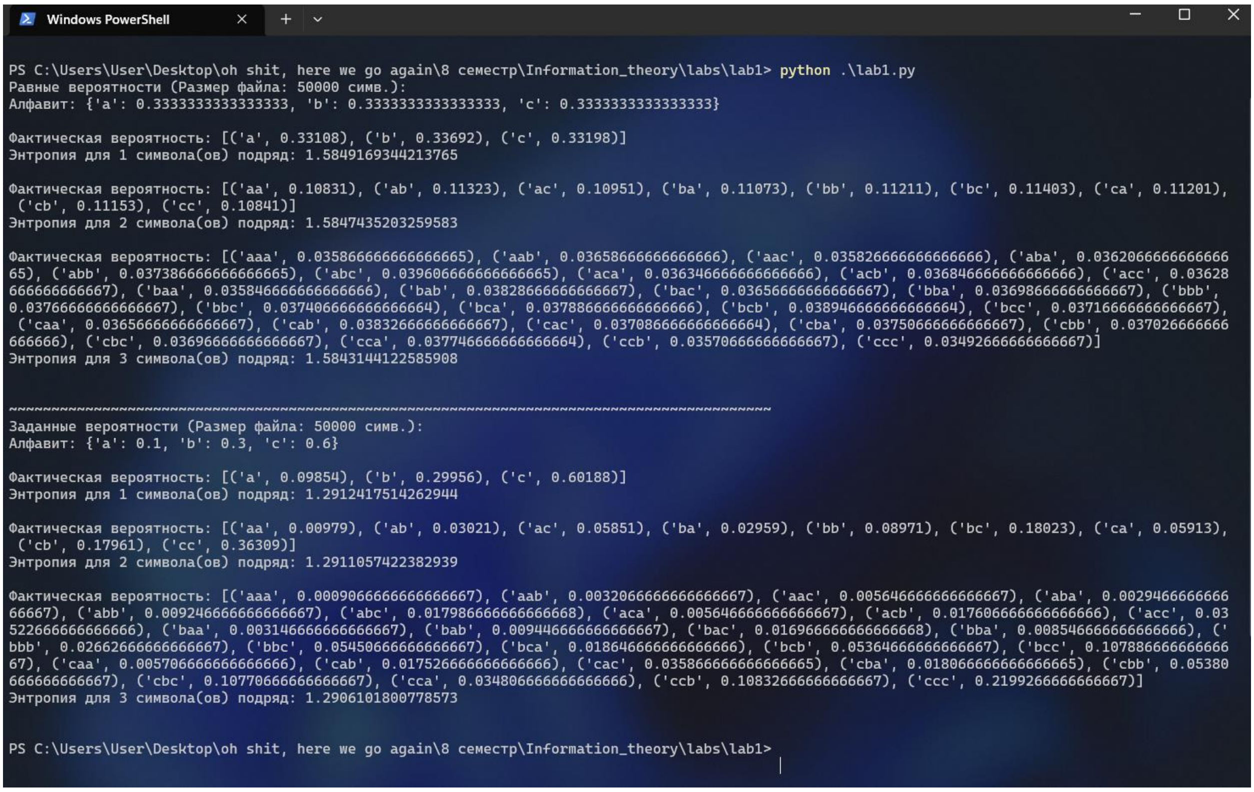
«Вычисление энтропии Шеннона»

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

Проверила:
доцент кафедры ПМиК
Мачикина Е.П.

Новосибирск
2022-2023 уч.год

Результат работы первой лабораторной



```
Windows PowerShell
PS C:\Users\User\Desktop\oh shit, here we go again\8 семестр\Information_theory\labs\lab1> python .\lab1.py
Равные вероятности (Размер файла: 50000 симв.):
Алфавит: {'a': 0.3333333333333333, 'b': 0.3333333333333333, 'c': 0.3333333333333333}

Фактическая вероятность: [('a', 0.33108), ('b', 0.33692), ('c', 0.33198)]
Энтропия для 1 символа(ов) подряд: 1.5849169344213765

Фактическая вероятность: [('aa', 0.10831), ('ab', 0.11323), ('ac', 0.10951), ('ba', 0.11073), ('bb', 0.11211), ('bc', 0.11403), ('ca', 0.11201),
 ('cb', 0.11153), ('cc', 0.10841)]
Энтропия для 2 символа(ов) подряд: 1.5847435203259583

Фактическая вероятность: [('aaa', 0.035866666666666665), ('aab', 0.036586666666666666), ('aac', 0.035826666666666666), ('aba', 0.0362066666666666
65), ('abb', 0.037386666666666665), ('abc', 0.039606666666666665), ('aca', 0.036346666666666666), ('acb', 0.036846666666666666), ('acc', 0.03628
666666666667), ('baa', 0.035846666666666666), ('bab', 0.03828666666666667), ('bac', 0.03656666666666667), ('bba', 0.03698666666666667), ('bbb',
0.03766666666666667), ('bbc', 0.037406666666666664), ('bca', 0.037886666666666666), ('bcb', 0.038946666666666664), ('bcc', 0.03716666666666667),
 ('caa', 0.03656666666666667), ('cab', 0.03832666666666667), ('cac', 0.037086666666666664), ('cba', 0.03750666666666667), ('cbb', 0.037026666666
666666), ('cbc', 0.03696666666666667), ('cca', 0.037746666666666664), ('ccb', 0.03570666666666667), ('ccc', 0.03492666666666667)]
Энтропия для 3 символа(ов) подряд: 1.5843144122585908


~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Заданные вероятности (Размер файла: 50000 симв.):
Алфавит: {'a': 0.1, 'b': 0.3, 'c': 0.6}

Фактическая вероятность: [('a', 0.09854), ('b', 0.29956), ('c', 0.60188)]
Энтропия для 1 символа(ов) подряд: 1.2912417514262944

Фактическая вероятность: [('aa', 0.00979), ('ab', 0.03021), ('ac', 0.05851), ('ba', 0.02959), ('bb', 0.08971), ('bc', 0.18023), ('ca', 0.05913),
 ('cb', 0.17961), ('cc', 0.36309)]
Энтропия для 2 символа(ов) подряд: 1.2911057422382939

Фактическая вероятность: [('aaa', 0.0009066666666666667), ('aab', 0.0032066666666666667), ('aac', 0.005646666666666667), ('aba', 0.0029466666666
66667), ('abb', 0.009246666666666667), ('abc', 0.017986666666666668), ('aca', 0.005646666666666667), ('acb', 0.017606666666666666), ('acc', 0.03
522666666666666), ('baa', 0.0031466666666666667), ('bab', 0.009446666666666667), ('bac', 0.016966666666666668), ('bba', 0.008546666666666666), ('
bbb', 0.02662666666666667), ('bbc', 0.05450666666666667), ('bca', 0.018646666666666666), ('bcb', 0.05364666666666667), ('bcc', 0.10788666666666
67), ('caa', 0.005706666666666666), ('cab', 0.017526666666666666), ('cac', 0.035866666666666665), ('cba', 0.018066666666666665), ('cbb', 0.05380
666666666667), ('cbc', 0.10770666666666667), ('cca', 0.034806666666666666), ('ccb', 0.10832666666666667), ('ccc', 0.21992666666666667)]
Энтропия для 3 символа(ов) подряд: 1.2906101800778573


PS C:\Users\User\Desktop\oh shit, here we go again\8 семестр\Information_theory\labs\lab1>
```

	Оценка энтропии (частоты отдельных символов)	Теоретическое значение энтропии (отдельные символы)	Оценка энтропии (частоты пар символов)	Теоретическое значение энтропии (для пар символов)
Равномерное	≈ 1.584917	≈1.584962	≈1.584962	≈1.584743
P(0,1; 0,3; 0,6)	≈1.291241	≈1.295462	≈1.295461	≈1.291105

Практические значения были получены в результате работы программы.

Расчет теоретических значений:

1) Теоретическое значение энтропии (отдельные символы):

1.1) Для файла, где все цифры генерируются последовательно и независимо с равными вероятностями:
Вероятность выпадения каждого символа равна 1/3;
-p*log(p) = 0.5283208335737187, где p = 1/3
H_1 (0,33; 0,33; 0,33) ≈1.584962

1.2) Для файла, где все цифры последовательности генерируются с заданными вероятностями:

a – 0,1
b – 0,3
c – 0,6
H_1 (0,1; 0,3; 0,6) ≈1.295462

2) Теоретическое значение энтропии (для пар символов):

2.1) Для файла F1:

H_2(0.1111; 0.1111; 0.1111; 0.1111; 0.1111; 0.1111; 0.1111; 0.1111; 0.1111) = (0.3522138890491458 * 9) / 2 ≈1.584743

2.2) Для файла F2:

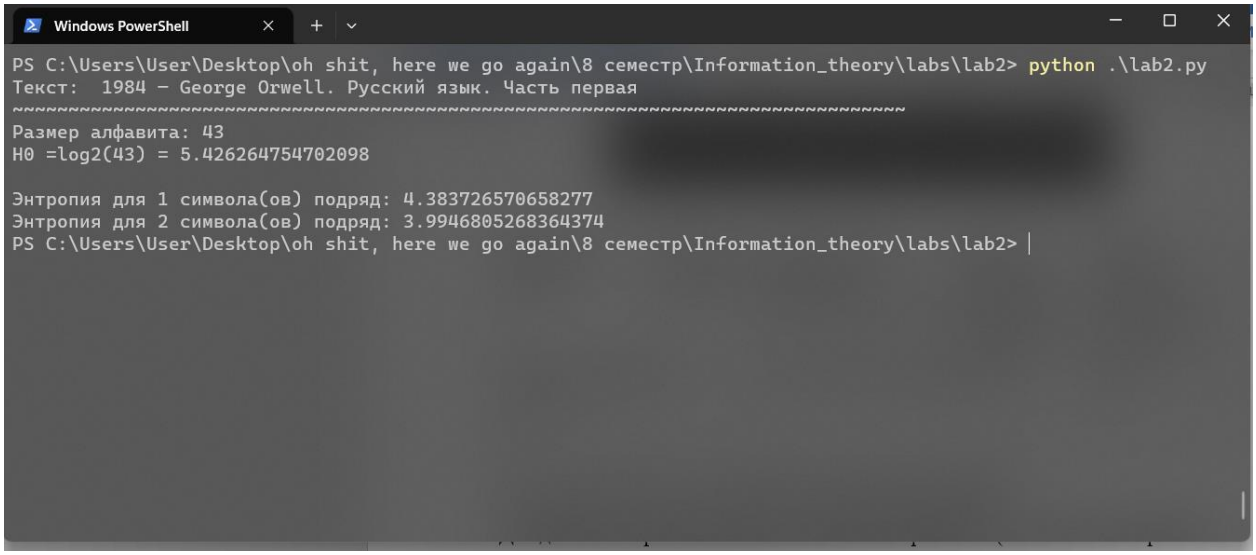
aa – 0,1 * 0,1 = 0,01
ab – 0,1 * 0,3 = 0,03
ac – 0,1 * 0,6 = 0,06
ba – 0,3 * 0,1 = 0,03
bb – 0,3 * 0,3 = 0,09
bc – 0,3 * 0,6 = 0,18
ca – 0,6 * 0,1 = 0,06
cb – 0,6 * 0,3 = 0,18
cc – 0,6 * 0,6 = 0,36

H_2 (0,01; 0,03; 0,06; 0,03; 0,09; 0,18; 0,06; 0,18; 0,36) = 2.5909236884766433/ 2 ≈1.291105

Вывод:

Сравнив теоретические и практические значения энтропии, можно сказать, что они очень близки. Из этого можно сделать вывод, что программа работает верно.

Результат работы второй лабораторной



Название файла	Размер алфавита	Максимальное возможное значение энтропии	Оценка энтропии (одиночные символы)	Оценка энтропии (частоты пар символов)
1984 — George Orwell.	43	≈5.4262647	≈4.383726	≈3.99468

Расчет максимального возможного значения энтропии:
H = log(m), где m – количество символов в алфавите
Для данного алфавита количество символов равно 43 (32 символов алфавита + 10 цифр + пробел).
H = log(43) = 5.426264754702098

Вывод:

В отличии от первой лабораторной — во второй энтропия для одиночных символов и пар сильно отличается, т.к. в художественном тексте у символов больший разброс по их частоте появления, из-за чего неопределённость появления для некоторых букв меньше, чем для других, а некоторые сочетания букв встречаются еще реже, поэтому неопределённость уменьшается еще сильнее.