

Протокол лабораторної роботи №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав

Студент 3 курсу

Групи ФБ-13

Короткевич Іван

Мета роботи: засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

Від самого початку планувалося виконати лабораторну роботу на мові C++, але виникла проблема з кодуванням символів, тому роботу було виконано мовою Python.

Також чомусь програма зараховує якесь порожнє місце(не пробіл) як символ(перший рядок у таблицях) і зараховує літеру б два рази.

Ентропія літер з пробілом: 4.417851
Ентропія літер без пробіла: 4.488255
Ентропія біграм з пробілом: 4.030182
Ентропія біграм без пробіла: 4.184205
Ентропія біграм(не перетинаються) з пробілом: 4.029912
Ентропія біграм(не перетинаються) без пробіла: 4.18347

Частота та ентропія всіх букв і біграм наведені у файлі lab1_table.
Значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$ були розраховані програмою CoolPinkProgram:

$H^{(10)}$

The screenshot shows the CoolPinkProgram interface with the following data:

- Произвольная часть текста:** удъ_возьмет_на_себя_труд_сравнить_учения_о_морали_господствовавшие_скажем_в
- Использованные буквы:** ч, и,
- Порядок n-граммы:** A list of n-grams from 5 to 50, with 10 selected.
- Введенный символ:** e
- Символ по счету:** 3
- Номер эксперимента:** 51
- Неравенство для энтропии:** $2.30611777048377 < H < 2.97533318523349$
- Двоичная таблица угаданных символов:** A 32x2 grid of 0s and 1s.
- Вероятности:** A list of probabilities $q[1]$ through $q[32]$.
- Строка состояния:** Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

$H^{(20)}$

1

Произвольная часть текста:
жен_уступать_тебе_дай_мне_кусочек_твоего_апельсина_я_давал_тебе_от_своего_д

Использованные буквы:
о, _, р,

Порядок n-граммы:
5
10
15
20
25
30
35
40
45
50

Введенный символ: а
Символ по счету: 4
Номер эксперимента: 51
Поле ввода символов:
а

Неравенство для энтропии:
1,75286696036016 < H < 2,41095528071352
Двоичная таблица угаданных символов:
00000001000000000000000000000000
00000000000000000000000000000000
00000000000000000000000000000000
00100000000000000000000000000000
10000000000000000000000000000000

Вероятности:
q[1] = 0,6078431
q[2] = 0,0392156
q[3] = 0,0392156
q[4] = 0,0588235
q[5] = 0,0196078
q[6] = 0,0196078
q[7] = 0
q[8] = 0,0196078
q[9] = 0
q[10] = 0
q[11] = 0
q[12] = 0,019607
q[13] = 0
q[14] = 0,019607
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0,019607
q[19] = 0,019607
q[20] = 0,019607
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0,019607
q[25] = 0
q[26] = 0
q[27] = 0,039215
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0,039215

Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

$H^{(30)}$

1

Произвольная часть текста:
_же_не_имело_бы_смысла_говорить_что_футбольный_игрок_допустил_нарушение_есл

Использованные буквы:
л, м,

Порядок n-граммы:
5
10
15
20
25
30
35
40
45
50

Введенный символ: т
Символ по счету: 3
Номер эксперимента: 51
Поле ввода символов:
т

Неравенство для энтропии:
1,73083777411109 < H < 2,60092288971848
Двоичная таблица угаданных символов:
10000000000000000000000000000000
01000000000000000000000000000000
00000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000

Вероятности:
q[1] = 0,5294117
q[2] = 0,1176470
q[3] = 0,0784313
q[4] = 0,0588235
q[5] = 0,0196078
q[6] = 0,0196078
q[7] = 0,0196078
q[8] = 0,0196078
q[9] = 0,0196078
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0,019607
q[17] = 0
q[18] = 0,019607
q[19] = 0
q[20] = 0
q[21] = 0,019607
q[22] = 0
q[23] = 0,019607
q[24] = 0
q[25] = 0
q[26] = 0,019607
q[27] = 0,019607
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0

Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

$$H^{(10)}: 2,03611777048377 < H < 2,97533318523349$$

$$H^{(20)}: 1,75286696036016 < H < 2,41095528071352$$

$$H^{(30)}: 1,73083777411109 < H < 2,60092288971848$$

Надлишковість джерела відкритого тексту:

$$H_0 = \log_2 34 \text{ (для алфавіту з пробілом)}$$

$$H_0 = \log_2 33 \text{ (для алфавіту без пробіла)}$$

Для букв з пробілом:

$$R_1 = 1 - \frac{H_1}{H_0} = 1 - \frac{4.4178519973618835}{5.087} = 0,13154$$

Для біграм з пробілом:

$$R_2 = 1 - \frac{H_2}{H_0} = 1 - \frac{4.030182621075429}{5.087} = 0,20774$$

Для букв без пробіла:

$$R_1 = 1 - \frac{H_1}{H_0} = 1 - \frac{4.4882552037331696}{5.044} = 0,11018$$

Для біграм без пробіла:

$$R_2 = 1 - \frac{H_2}{H_0} = 1 - \frac{4.184205702319153}{5.044} = 0,17045$$

Для біграм(не перетинаються) з пробілом:

$$R_2 = 1 - \frac{H_2}{H_0} = 1 - \frac{4.0299126670791425}{5.087} = 0,20780$$

Для біграм(не перетинаються) без пробіла:

$$R_2 = 1 - \frac{H_2}{H_0} = 1 - \frac{4.1834700614518585}{5.044} = 0,17060$$

$$H^{(10)}: 1 - \frac{2,03611777048377}{5} < R < 1 - \frac{2,97533318523349}{5}$$

$$0,59277 > R > 0,40493$$

$$H^{(20)}: 1 - \frac{1,752866960360167}{5} < R < 1 - \frac{2,41095528071352}{5}$$

$$0,64942 > R > 0,51780$$

$$H^{(30)}: 1 - \frac{1,73083777411109}{5} < R < 1 - \frac{2,60092288971848}{5}$$

$$0,65383 > R > 0,47981$$

Висновок

З наведених вище даних можна зробити висновок, що при збільшені n , R буде збільшуватись.