

Міністерство освіти і науки України Національний  
технічний університет України "Київський політехнічний  
інститут імені Ігоря Сікорського" Фізико-технічний  
інститут

## Криптографія

### Комп'ютерний практикум No1

Експериментальна оцінка ентропії на символ джерела  
відкритого тексту

Виконав студент групи  
ФБ-11 Анучін Максим

Київ - 2023

## 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.

## Порядок виконання роботи

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку  $H_1$  та  $H_2$  за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення  $H_1$  та  $H_2$  на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення  $H_1$  та  $H_2$  на тому ж тексті, в якому вилучено всі пробіли.

Буду наводити по 5 рядків з таблиць, ознайомитись з повним набором, можна в відповідних csv файлах

Букви:

Без пробілів:  $H_1 = 4.467235290398882$   $R = 0.10655294192022358$

table\_letters\_nospaces

Біграма	Частота	Вірогідність
о	98914	0.10982134629241652
е	81893	0.09092342349844174
а	67731	0.07519976551076352
н	60469	0.06713697746482938
и	59977	0.06659072412985285

З пробілами  $H_1 = 4.463449196370976$   $R = 0.10731016072580479$

table\_letters\_spaces

Біграма	Частота	Вірогідність
	166416	0.1559520830814818
о	98914	0.09269447857130139
е	81893	0.07674372620296
а	67731	0.06347220543212098
н	60469	0.056666825977394746

Біграми без пробілів, що перетинаються:

$$H_2 = 4.145794847240513; R = 0.5854205152759487$$

table\_bigrams\_nospaces\_cross

Біграма	Частота	Вірогідність
ст	6662	0.014793267309144202
то	6297	0.013982768574854554
но	5561	0.012348447839410224
не	5402	0.011995381267486787
ен	5204	0.011555713460940623

Біграми без пробілів, що не перетинаються:

$$H_2 = 4.1219309719166555; R = 0.5878069028083345$$

table\_bigrams\_nospaces\_nocross

Біграма	Частота	Вірогідність
ст	6765	0.01514259557294523
то	6489	0.014524804534049016
но	6276	0.01404803101490085
не	5539	0.012398349871181614
ен	5061	0.011328407419759911

Біграми з пробілами, що перетинаються:

$H_2 = 4.06477904182616$ ;  $R = 0.593522095817384$

table\_bigrams\_spaces\_cross

Біграма	Частота	Вірогідність
о	10976	0.020571719882747194
е	9289	0.017409867528319852
и	8971	0.016813857422387488
с	8290	0.015537496157796487
п	8106	0.015192634964426818

Біграми з пробілами, що не перетинаються:

$H_2 = 4.0449804653305925$ ;  $R = 0.5955019534669408$

table\_bigrams\_spaces\_nocross

Біграма	Частота	Вірогідність
о	10662	0.020064132962988056
е	10391	0.019554155469743845
и	9502	0.017881203471610627
с	8026	0.01510361387740969
п	7855	0.014781819961008363

2. За допомогою програми CoolPinkProgram оцінити значення  $H^{(10)}$ ,  $H^{(20)}$ ,  $H^{(30)}$ .

Лабораторная работа №1

Произвольная часть текста:  
лишенные\_музыкального\_слуха\_но\_рассматривая\_человечество\_в\_целом\_люди\_полаг

Использованные буквы:  
т, р, п, к, я, ч, с,

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

Введенный символ: м

Символ по счету: 8

Номер эксперимента: 60

Поле ввода символов:  
м

Продолжить

Другой

Неравенство для энтропии:  
1,96281979349206 < H < 2,82828238106548

Двоичная таблица угаданных символов:

Вероятности:  
q[1] = 0,5  
q[2] = 0,133333  
q[3] = 0,05  
q[4] = 0,05  
q[5] = 0,0166666  
q[6] = 0,0166666  
q[7] = 0  
q[8] = 0,0166666  
q[9] = 0  
q[10] = 0,016666  
q[11] = 0,016666  
q[12] = 0,016666  
q[13] = 0,016666  
q[14] = 0  
q[15] = 0,016666  
q[16] = 0,016666  
q[17] = 0,016666  
q[18] = 0  
q[19] = 0,016666  
q[20] = 0  
q[21] = 0  
q[22] = 0,033333  
q[23] = 0  
q[24] = 0  
q[25] = 0  
q[26] = 0  
q[27] = 0,033333  
q[28] = 0  
q[29] = 0  
q[30] = 0  
q[31] = 0  
q[32] = 0,016666

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

$$1.96281979349206 < H < 2.82828238106548$$

Лабораторная работа №1

Произвольная часть текста:  
али\_и\_не\_сделали\_что\_ж\_вы\_никогда\_не\_стали\_бы\_связывать\_себя\_словом\_если\_бы

Использованные буквы:

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

Введенный символ: о

Символ по счету: 1

Номер эксперимента: 60

Поле ввода символов:  
о

Продолжить

Другой

Неравенство для энтропии:  
1,65160632344393 < H < 2,3821437990614

Двоичная таблица угаданных символов:

Вероятности:  
q[1] = 0,483333  
q[2] = 0,25  
q[3] = 0,0333333  
q[4] = 0,0333333  
q[5] = 0,0166666  
q[6] = 0,0333333  
q[7] = 0,0333333  
q[8] = 0  
q[9] = 0  
q[10] = 0,016666  
q[11] = 0,033333  
q[12] = 0  
q[13] = 0,016666  
q[14] = 0  
q[15] = 0  
q[16] = 0  
q[17] = 0  
q[18] = 0  
q[19] = 0  
q[20] = 0,016666  
q[21] = 0  
q[22] = 0  
q[23] = 0  
q[24] = 0  
q[25] = 0  
q[26] = 0,033333  
q[27] = 0  
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

$$1.65160632344393 < H < 2.3821437990614$$

Лабораторная работа №1

Произвольная часть текста:  
асны\_в\_том\_что\_не\_следует\_ставить\_на\_первое\_место\_самого\_себя\_эгоизм\_никогд

Использованные буквы:  
н, б, д, й, у, ы,

Порядок n-граммы:  
☐ 5 символов  
☐ 10 символов  
☐ 15 символов  
☐ 20 символов  
☐ 25 символов  
☒ 35 символов  
☐ 40 символов  
☐ 45 символов  
☐ 50 символов

Введенный символ: в

Символ по счету: 7

Номер эксперимента: 60

Неравенство для энтропии:  
 $1.5864087627454 < H < 2.28147550239062$

Двоичная таблица угаданных символов:

Поле ввода символов:  
в

Продолжить Другой

Вероятности:

q[1]	= 0,5833333
q[2]	= 0,1333333
q[3]	= 0,0333333
q[4]	= 0,0666666
q[5]	= 0,0166666
q[6]	= 0
q[7]	= 0,0166666
q[8]	= 0
q[9]	= 0
q[10]	= 0
q[11]	= 0,0166666
q[12]	= 0
q[13]	= 0
q[14]	= 0
q[15]	= 0
q[16]	= 0,0166666
q[17]	= 0
q[18]	= 0,0333333
q[19]	= 0
q[20]	= 0
q[21]	= 0
q[22]	= 0
q[23]	= 0
q[24]	= 0
q[25]	= 0
q[26]	= 0
q[27]	= 0
q[28]	= 0,0166666
q[29]	= 0,0333333
q[30]	= 0
q[31]	= 0,0166666
q[32]	= 0,0166666

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

$$1.5864087627454 < H < 2.28147550239062$$

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

$$R = 1 - H_{\infty}/H_0; H_0 = \log_2 32 = 5$$

$$H^{(10)}: 1 - 1.96281979349206/5 \approx 0,61 < R < 1 - 2.82828238106548/5 \approx 0,43$$

$$H^{(20)}: 1 - 1.65160632344393/5 \approx 0,67 < R < 1 - 2.3821437990614/5 \approx 0,52$$

$$H^{(30)}: 1 - 1.5864087627454/5 \approx 0,68 < R < 1 - 2.28147550239062/5 \approx 0,54$$

## Проблеми, які виникли у ході роботи

1. Розібратись з методичкою. Виділити час на лабу). Знайти нормальний txt файл з книжкою, Фінансист Т.Д., бо багато з кривою кирилицею.

## Висновки

Під час виконання лабораторної роботи ми ознайомились з поняттями ентропії на символ джерела та його надлишковості, виконали практичну перевірку за допомогою написаної власноруч програми. Поклацали CoolPinkProgram, для мануального тесту ентропії та вірогідності вгадування + можна було під час експерименту побачити, найбільш вживані біграми та в принципі найбільш вживані літери. Експериментально визначили ентропію, щоб більш наочно все перевірити.