

# Лабораторна робота №1

## Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

ФБ-23 Литвин Руслан

ФБ-23 Ващаєв Тимофій

### 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.
2. За допомогою програми CoolPinkProgram оцінити значення  $H^{(10)}$ ,  $H^{(20)}$ ,  $H^{(30)}$ .
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

### Хід роботи

Початковою і головною задачею було фільтрація тексту російською мовою для подальшого його аналізу. Виконання цього етапу не викликало жодних труднощів, усі символи, окрім текстових, були вилучені, прописні літери замінилися на відповідні стрічні, а послідовність пробілів та інших розділових знаків було видалено.

Виконання наступного етапу, що полягав у підрахунку частот букв у тексті, відбувалося у декілька кроків:

1. Спочатку, використовуючи цикл `for`, рахуємо кількість появи кожного символу в обраному тексті, зберігаючи результати в словник.  

```
for item in text:
    if item not in letters.keys():
        letters[item] = 1
    else:
        letters[item] += 1
```
2. Далі знаходимо загальну кількість символів у тексті.  

```
text_size = len(text)
```
3. Тепер підраховуємо частоту букв, що є відношенням кількості появи певного символу до загальної кількості символів у тексті.  

```
for key, val in letters.items():
    letters[key] = val / text_size
```

Наступним завданням став підрахунок частот біграм у тексті. Виконання цього етапу, як і підрахунок частот букв, відбувалося у декілька кроків:

1. Спочатку, використовуючи цикл `for`, рахуємо кількість появи кожної біграми в тексті
2. Далі знаходимо загальну кількість біграм у тексті  
`bigrams_count = sum(bigrams.values())`
3. Тепер підраховуємо частоту біграм, що є відношенням кількості появи певної біграми до загальної кількості біграм у тексті  
`for key, val in bigrams.items():  
 bigrams[key] = val / bigrams_count`

Основною складністю цього етапу було те, що при підрахунку частот біграм треба розглядати як пари букв, що перетинаються, так і пари букв, що не перетинаються. Для вирішення даної проблеми було додано параметр `cross`, що визначає крок руху вздовж тексту.

```
for i in range(0, len(text) - 1, 1 if cross else 2):  
    item = text[i] + text[i + 1]  
    if item in bigrams.keys():  
        bigrams[item] += 1
```

На наступному етапі обчислюємо значення  $H_1$  та  $H_2$  на обраному тексті. Для того, щоб обчислити значення  $H_1$  та  $H_2$  скористаємося наступною формулою:

$$H_n = \frac{1}{n} H(x_1, x_2, \dots, x_n),$$

де  $H(x_1, x_2, \dots, x_n)$  – ентропія  $n$ -грами відкритого тексту  $(x_1, x_2, \dots, x_n)$

```
def entropy(bigrams, n = 1):  
    return -sum(p * log2(p) for p in bigrams.values() if p > 0) / n
```

Для обчислення значення  $H_1$  знаходимо частоту букв у тексті та використовуємо наступну формулу:

$$H_1 = - \sum_{i=1}^n p_i * \log_2 p_i,$$

де  $p_i$  – частота появи певної літери в тексті, а  $n$  – загальна кількість символів.

Для обчислення значення  $H_2$  знаходимо частоту біграм у тексті та використовуємо наступну формулу:

$$H_2 = - \frac{1}{2} \sum_{i=1}^n p_i * \log_2 p_i,$$

де  $p_i$  – частота появи певної біграми в тексті, а  $n$  – загальна кількість біграм.

На останньому етапі обчислюємо надлишковість джерела відкритого тексту, використовуючи наступну формулу:

$$R = 1 - \frac{H_\infty}{H_0},$$

де  $H_\infty$  – ентропія джерела  $(H_1, H_2)$ , а  $H_0 = \log_2 m$  ( $m$  – кількість букв в алфавіті).

```
def redundancy(h, alphabet):  
    return 1 - (h / log2(len(alphabet)))
```

## Результати розрахунків

Усі таблиці, що продемонстровані, зберігаються у відповідних .csv файлах

Таблиця частот літер

Літера	Частота (без пробілу)	Частота (з пробілом)
	-	0.164662
о	0.116239	0.097099
е	0.087349	0.072966
а	0.074936	0.062597
н	0.066793	0.055795
и	0.063623	0.053146
т	0.062762	0.052427
с	0.055835	0.046641
л	0.049871	0.041659
р	0.042513	0.035513
в	0.042078	0.035149
м	0.035170	0.029379
к	0.032909	0.027490
д	0.031532	0.026340
у	0.027092	0.022631
п	0.025749	0.021509
я	0.021210	0.017718
ы	0.020290	0.016949
ь	0.019107	0.015961
г	0.018455	0.015416
б	0.017562	0.014670
ч	0.016347	0.013655
з	0.016177	0.013513
ж	0.010604	0.008858
й	0.010476	0.008751
х	0.008463	0.007070
ш	0.007825	0.006537
ю	0.006303	0.005265
э	0.005050	0.004219
ц	0.003093	0.002583
щ	0.002933	0.00245
ф	0.001322	0.001104
ъ	0.000291	0.000243
ё	0.000039	0.000033

### Таблиці частот біграм

У наступних таблицях наведено 20 найбільш уживаних біграм (через великий розмір таблиці, додати усі біграми неможливо).

Не перехресні, без пробілу	
Біграма	Частота
то	0.017132
ст	0.013966
на	0.012998
но	0.012421
не	0.011810
он	0.011128
ен	0.010956
по	0.010286
ос	0.009853
ко	0.009537
ов	0.009478
го	0.009398
ли	0.009166
от	0.009158
ал	0.009069
ер	0.008918
ни	0.008665
ом	0.008236
ка	0.008202
ло	0.008173

Не перехресні, з пробілом	
Біграма	Частота
о	0.023161
е	0.017899
и	0.017181
с	0.016232
н	0.016214
в	0.015275
а	0.015156
п	0.014811
то	0.014178
я	0.011881
о	0.011399
ст	0.011343
на	0.010397
но	0.010207
ь	0.009655
не	0.009634
и	0.009204
по	0.008853
м	0.008413
к	0.008336

Перехресні, без пробілу	
Біграма	Частота
то	0.017208
ст	0.013945
на	0.012806
но	0.012322
не	0.011613
он	0.011122
ен	0.011065
по	0.010352
ко	0.009844
ос	0.009796
ов	0.009663
го	0.009512
ал	0.009095
от	0.009082
ер	0.009006
ли	0.008916
ни	0.008665
ло	0.008280
ка	0.008196
ра	0.008122

Перехресні, з пробілом	
Біграма	Частота
о	0.023393
е	0.017802
и	0.016983
с	0.016357
н	0.016274
в	0.015377
а	0.015032
п	0.014980
то	0.014083
я	0.011920
о	0.011577
ст	0.011404
на	0.010664
но	0.010026
ь	0.009762
не	0.009648
и	0.009018
по	0.008640
м	0.008337
к	0.008288

Розрахуємо надлишковість джерела відкритого тексту, використовуючи наступну формулу:

$$R = 1 - \frac{H_{\infty}}{H_0},$$

де  $H_0 = \log_2 33 = 5.044$  для тексту без пробілів і  $H_0 = \log_2 34 = 5.087$  для тексту з пробілами.

Також обчислимо питому ентропію ( $H_n$ ) для кожної моделі відкритого тексту.

Модель відкритого тексту	Ентропія	Надлишковість
$H_1$ з пробілами	4.370231	0.140980
$H_1$ без пробілів	4.459134	0.116022
$H_2$ перехресні біграми з пробілами	3.968223	0.220000
$H_2$ перехресні біграми без пробілів	4.144162	0.178462
$H_2$ не перехресні біграми з пробілами	3.967991	0.220045
$H_2$ не перехресні біграми без пробілів	4.142745	0.178743

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

Модель відкритого тексту	Надлишковість
$H^{(10)}$	$1.646608 < H^{(10)} < 2.332097$
$H^{(20)}$	$1.292996 < H^{(20)} < 2.006961$
$H^{(30)}$	$1.097476 < H^{(30)} < 1.884361$

Виконаємо 50 експериментів для того, щоб знайти значення  $H^{(10)}$

Лабораторная работа №1

Произвольная часть текста:  
е\_не\_имел

Использованные буквы:

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:

Символ по счету:

Номер эксперимента: 51

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:  
1.6466075546266 < H < 2.33209668899588

Двоичная таблица угаданных символов:  
10000000000000000000000000000000  
00001000000000000000000000000000  
10000000000000000000000000000000  
10000000000000000000000000000000  
10000000000000000000000000000000

Вероятности:  
q[1]=0.48  
q[2]=0.18  
q[3]=0.06  
q[4]=0.12  
q[5]=0.06  
q[6]=0  
q[7]=0.02  
q[8]=0  
q[9]=0  
q[10]=0  
q[11]=0  
q[12]=0  
q[13]=0.04  
q[14]=0  
q[15]=0  
q[16]=0  
q[17]=0  
q[18]=0  
q[19]=0.02  
q[20]=0  
q[21]=0.02  
q[22]=0  
q[23]=0  
q[24]=0  
q[25]=0  
q[26]=0  
q[27]=0  
q[28]=0  
q[29]=0  
q[30]=0  
q[31]=0  
q[32]=0

Строка состояния:

Виконаємо 50 експериментів для того, щоб знайти значення  $N^{(20)}$

[illegible]

Виконаємо 50 експериментів для того, щоб знайти значення  $N^{(30)}$

[illegible]



## Висновки

У результаті виконання лабораторної роботи ми ознайомилися з поняттям ентропії на символ джерела та його надлишковості, а також навчилися визначати частоти літер і біграм на довільному тексті, розраховували ентропію та надлишковість мови в різних моделях джерела.

Використовуючи отримані значення, ми помітили, що питома ентропія  $H_1$  з пробілами більша за  $H_1$  без пробілів. Також, аналізуючи перехресні та не перехресні біграми з пробілами (перехресні та не перехресні біграми без пробілів), було помітно, що значення ентропії  $H_2$  та їх надлишковість майже не відрізняються.

Також, використовуючи програму CoolPinkProgram, ми помітили, що значення умовної ентропії джерела  $H^{(n)}$  зменшується тоді, коли  $n$  зростає, тобто чим довший текст дано, тим легше вгадати наступну літеру.