

## КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

### Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав: Кандила Микита ФБ-12

**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:

0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
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 видалено 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

### 1. Підрахунок $H(1)$ та $H(2)$

#### 1.1 Фільтрація

Для початку за допомогою функції *text\_filter* з початкового тексту було «видалено» зайві символи, а саме символи переносу рядка і всі букви, які не належать російському алфавіту (включаючи букву «ё»). Далі були видалені зайві пробіли і новий текст був записаний в новий файл (*filtered.txt*).

## 1.2 Підрахунок символів для $H_1$

Після фільтрації були підрахована загальна кількість символів у тексті за допомогою функції `letter_count()`. Відповідні дані були занесені в змінні `dict_no_space` і `dict_with_space`.

## 1.3 Розрахунок $H_1$

Наразі можна легко знайти  $H_1$  використовуючи формулу:

$$H_1 = - \sum_1^n p(i) \cdot \log_2 p(i),$$

де  $p(i)$  – частота появи букви в тексті.

Для знаходження  $H_1$  була використана функція `entropy_h1_calc()` для тексту з пробілами і без.

## 1.4 Підрахунок символів для $H_2$

Для підрахунку біграм з перетину і без на тексті в були використані функції `no_cross` і `with_cross`.

## 1.5 Розрахунок $H_2$

Після підрахунку біграм, для знаходження  $H_2$  була використана формула:

$$H_2 = - \frac{1}{2} \sum_1^n p(i, j) \cdot \log_2 p(i, j),$$

де  $p(i, j)$  – частота появи біграми в тексті.

Була використана функція `entropy_h2_calc()`.

## 2. Результати виконання коду

### 2.1 Таблиця частот літер

З пробілом:

Буква	Кількість	Частота
а	47086	0.06686
б	10721	0.015223
в	24317	0.034529
г	10004	0.014205
д	17799	0.025274
е	50576	0.071815
ж	6505	0.009237
з	10255	0.014562
и	36418	0.051712
й	5470	0.007767
к	21475	0.030493
л	27765	0.039425
м	19223	0.027296
н	38806	0.055103
о	68455	0.097203
п	15871	0.022536
р	22374	0.03177
с	30846	0.0438
т	38288	0.054367
у	17107	0.024291
ф	900	0.001278
х	4969	0.007056
ц	1362	0.001934
ч	9433	0.013394
ш	4613	0.00655
щ	1963	0.002787
ъ	120	0.00017
ы	11036	0.015671
ь	12963	0.018407
э	2002	0.002843
ю	2926	0.004155
я	12442	0.017667
пробіл	120161	0.170622

Без пробіла:

Буква	Кількість	Частота
а	47086	0.080614
б	10721	0.018355
в	24317	0.041632
г	10004	0.017127
д	17799	0.030473
е	50576	0.086589
ж	6505	0.011137
з	10255	0.017557
и	36418	0.06235
й	5470	0.009365
к	21475	0.036767
л	27765	0.047535
м	19223	0.032911
н	38806	0.066438
о	68455	0.117199
п	15871	0.027172
р	22374	0.038306
с	30846	0.05281
т	38288	0.065552
у	17107	0.029288
ф	900	0.001541
х	4969	0.008507
ц	1362	0.002332
ч	9433	0.01615
ш	4613	0.007898
щ	1963	0.003361
ъ	120	0.000205
ы	11036	0.018894
ь	12963	0.022193
э	2002	0.003428
ю	2926	0.00501
я	12442	0.021302

## 2.2 Таблиці частот біграм

Без пробілу без перетину

Біграма	Кількість	Частота
то	5091	0.017432
но	3980	0.013628
не	3871	0.013255
ст	3644	0.012478
на	3584	0.012272
по	3458	0.011841
он	3269	0.011193
ос	3130	0.010718
ко	3076	0.010533
от	3065	0.010495
ни	2970	0.01017
ал	2919	0.009995
ен	2825	0.009673
ов	2794	0.009567
ка	2743	0.009392

З пробілом без перетину

Біграма	Кількість	Частота
о_	8701	0.02471
е_	6964	0.019777
_н	6573	0.018667
а_	6291	0.017866
и_	6159	0.017491
_с	5897	0.016747
_п	5612	0.015938
_в	5399	0.015333
то	4881	0.013862
_о	4580	0.013007
ь_	3999	0.011357
не	3913	0.011113
я_	3909	0.011101
но	3834	0.010888
_к	3775	0.010721

Без пробілу з перетином

Біграма	Кількість	Частота
то	10170	0.017412
не	7910	0.013542
но	7871	0.013476
ст	7280	0.012464
на	7158	0.012255
по	6919	0.011846
он	6613	0.011322
ос	6274	0.010742
ко	6167	0.010558
от	6104	0.01045
ни	5927	0.010147
ал	5873	0.010055
ен	5601	0.009589
ов	5592	0.009574
ка	5339	0.009141

З пробілом з перетином

Біграма	Кількість	Частота
о_	17500	0.024849
е_	13987	0.019861
_н	12930	0.01836
а_	12465	0.0177
и_	12359	0.017549
_с	11790	0.016741
_п	11118	0.015787
_в	10810	0.01535
то	9834	0.013964
_о	9017	0.012804
ь_	8158	0.011584
я_	7913	0.011236
не	7848	0.011144
но	7723	0.010966
_к	7413	0.010526

## 2.3 Надлишковість


Для обчислення надлишковості була використана формула:


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


де  $H_0 = \log_2 33$  для тексту з пробілами і  $H_0 = \log_2 32$  для тексту без пробілу.

	Ентропія	Надлишковість
$H_1$ без пробілу	4.447267003541618	0.11054659929167643
$H_1$ з пробілом	4.347590209022435	0.13813431184172376
$H_2$ без пробілу без перет.	4.11350950910577	0.17729809817884612
$H_2$ без пробілу з перет.	4.113794730173205	0.17724105396535905
$H_2$ з пробілом без перет.	3.9219529030559217	0.22251259313681138
$H_2$ з пробілом з перет.	3.9227061279249646	0.22236327394183564

## 3. CoolPinkProgram

$H_{10}$  

$H_{20}$  

$H_{30}$  

Таблиця ентропій і надлишковостей

	Ентропія	Надлишковість
$H_{10}$	2.1383065<H<2.8515107	0.439502< R<0.579691
$H_{20}$	2.0543214<H<2.6213152	0.48475 <R< 0.596199

$H_{30}$	$2.1466696 < H < 2.7079764$	$0.467716 < R < 0.578047$
----------	-----------------------------	---------------------------

Висновки: в даній лабораторній роботі я вдосконалив навички фільтрації текстів, а також їхній аналіз на кількість окремих символів і біграм. Також я помітив, що значення  $H_1$  з пробілом та без нього майже не відрізняються. Теж саме можна сказати і для  $H_2$ . Ознайомився з поняття ентропії та надлишковості.