# НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ "КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО" ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

# КРИПТОГРАФІЯ КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали студенти 3 курсу групи ФБ-21 ДЗИСЮК Владислав ТЕЛУХ Анастасія

#### Варіант 8

**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.

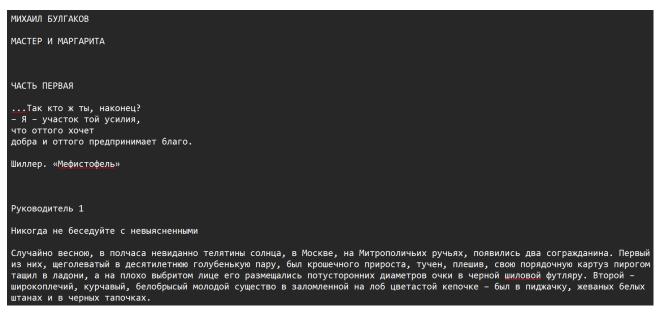
#### Порядок виконання роботи:

- 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 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

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку Н1 та Н2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення Н1 та Н2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 на тому ж тексті, в якому вилучено всі пробіли.

Наш початковий текст збережений у файлі text.txt. Це частина роману М. Булгакова «Майстер і Маргарита» російською.



### Частоти букв:

Отримана нами статистика по літерам збережена у файлі letters\_results.txt.

Наведемо скріншоти з результатами:

```
Частоти літер без пробілів:
Частоти літер з пробілами:
  ': 0.144859 (108668)
                                       'o': 0.107447 (68927)
                                        'a': 0.086784 (55672)
'o': 0.091882 (68927)
'a': 0.074213 (55672)
                                        'e': 0.080903 (51899)
                                        'u': 0.070295 (45094)
 e': 0.069183 (51899)
'u': 0.060112 (45094)
                                        'h': 0.065116 (41772)
                                       'T': 0.058801 (37721)
'h': 0.055684 (41772)
'T': 0.050284 (37721)
                                       'л': 0.051790 (33223)
'л': 0.044288 (33223)
                                       'c': 0.051430 (32992)
                                       'в': 0.046956 (30122)
   : 0.043980 (32992)
                                        'p': 0.043174 (27696)
'в': 0.040154 (30122)
'p': 0.036920 (27696)
                                        'k': 0.039679 (25454)
                                       'y': 0.030474 (19549)
  ': 0.033931 (25454)
                                       'n': 0.028674 (18394)
'y': 0.026060 (19549)
'π': 0.024520 (18394)
                                       'm': 0.028115 (18036)
'm': 0.024043 (18036)
                                        'д': 0.027486 (17632)
 д': 0.023504 (17632)
                                        's': 0.019199 (12316)
 з': 0.016418 (12316)
                                        'ы': 0.018090 (11605)
'ы': 0.015470 (11605)
                                        'ь': 0.018070 (11592)
'ь': 0.015453 (11592)
                                        'я': 0.017774 (11402)
'я': 0.015199 (11402)
                                       'r': 0.017099 (10969)
'r': 0.014622 (10969)
                                       '4': 0.016901 (10842)
'4': 0.014453 (10842)
                                       '6': 0.016401 (10521)
'6': 0.014025 (10521)
                                       'й': 0.014712 (9438)
'й': 0.012581 (9438)
                                       'ш': 0.009362 (6006)
'ш': 0.008006 (6006)
                                        'ж': 0.008817 (5656)
'ж': 0.007540 (5656)
                                        'x': 0.007411 (4754)
'x': 0.006337 (4754)
                                        'ю': 0.005414 (3473)
'ю': 0.004630 (3473)
                                       'ц': 0.004352 (2792)
'ц': 0.003722 (2792)
                                       'щ': 0.004137 (2654)
'щ': 0.003538 (2654)
                                       '∍': 0.002940 (1886)
 'э': 0.002514 (1886)
                                       'ф': 0.002196 (1409)
'¢': 0.001878 (1409)
                                       Ентропія Н1: 4.464745
Ентропія Н1: 4.414811
                                       Надлишковість: 0.107051
Надлишковість: 0.117038
```

#### Бачимо, що:

- 1) Для літер з пробілами: ентропія H1 становить 4.414811, надлишковість 0.117038.
- 2) Для літер без пробілів: ентропія Н1 становить 4.464745, надлишковість 0.107051.

## Частоти біграм:

Статистика щодо частот біграм збережена у файлі bigrams\_results.txt. Отримані результати (список біграм наведений частково):

• Біграми з пробілами (перетинаються):

```
Частоти біграм з пробілами (перетинаються):
'o ': 0.019202 (14405)
'a ': 0.016858 (12646)
' п': 0.015441 (11583)
'и ': 0.014997 (11250)
'в': 0.014861 (11148)
'e ': 0.014778 (11086)
' н': 0.014225 (10671)
' c': 0.013434 (10078)
'то': 0.012312 (9236)
'и': 0.010115 (7588)
'по': 0.010044 (7535)
'но': 0.009750 (7314)
' o': 0.009473 (7106)
'CT': 0.009347 (7012)
'на': 0.009250 (6939)
```

Отримані результати: Надлишковість: 0.193548

Ентропія Н2: 4.032262

Біграми з пробілами (не перетинаються):

```
Частоти біграм з пробілами (не перетинаються):
'o ': 0.019262 (7225)
'a ': 0.016943 (6355)
' п': 0.015231 (5713)
'e ': 0.014922 (5597)
'и ': 0.014869 (5577)
'в': 0.014813 (5556)
 н': 0.014034 (5264)
 c': 0.013309 (4992)
'то': 0.012416 (4657)
'по': 0.010246 (3843)
'и': 0.010136 (3802)
'но': 0.009590 (3597)
' o': 0.009547 (3581)
'на': 0.009419 (3533)
'cT': 0.009265 (3475)
```

Ентропія Н2: 4.031067 Отримані результати: Надлишковість: 0.193787

Біграми без пробілів (перетинаються):

```
Частоти біграм без пробілів (перетинаються):
'то': 0.014828 (9512)
'по': 0.011746 (7535)
'но': 0.011652 (7475)
'ст': 0.011090 (7114)
'на': 0.010884 (6982)
'κο': 0.010639 (6825)
'не': 0.010538 (6760)
'ал': 0.010315 (6617)
'ка<mark>': 0.</mark>010089 (6472)
'ов': 0.009824 (6302)
'он': 0.009679 (6209)
'ла': 0.009412 (6038)
'oc': 0.009330 (5985)
'ен': 0.009083 (5827)
'ли': 0.008538 (5477)
```

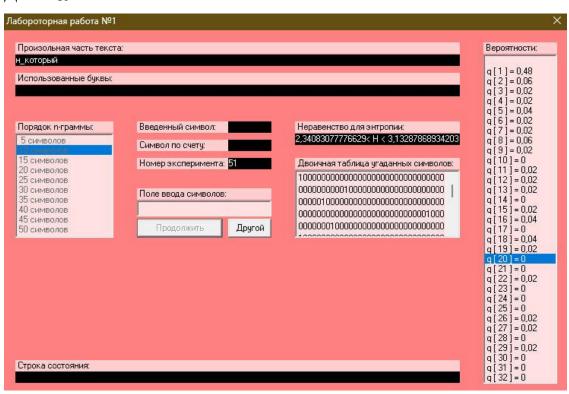
Ентропія Н2: 4.157232 Отримані результати: Надлишковість: 0.168554

Біграми без пробілів (не перетинаються):

```
Частоти біграм без пробілів (не перетинаються):
'то': 0.014619 (4689)
'по': 0.011953 (3834)
'но': 0.011592 (3718)
'<u>ст':</u> 0.0<mark>11164 (3581)</mark>
'ко': 0.010666 (3421)
'не': 0.010585 (3395)
на': 0.010444 (3350)
'ал': 0.010242 (3285)
'ка': 0.009945 (3190)
'ов': 0.009911 (3179)
'он': 0.009768 (3133)
'ла': 0.009359 (3002)
'oc': 0.009241 (2964)
'ен': 0.009048 (2902)
'po': 0.008527 (2735)
```

Ентропія Н2: 4.157152 Отримані результати: Надлишковість: 0.168570

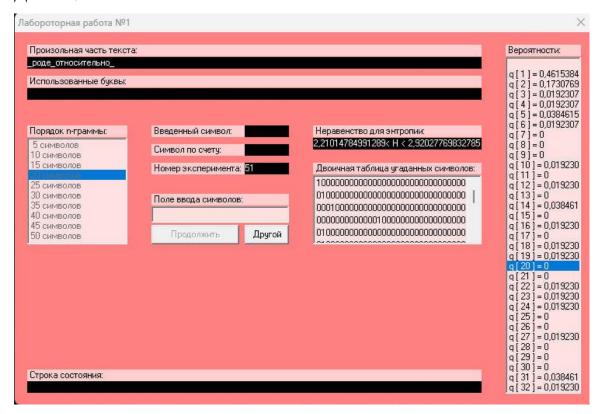
2. За допомогою програми CoolPinkProgram оцінити значення  $H_{10}$ ,  $H_{20}$ ,  $H_{30}$ **Для Н**<sub>10</sub>:



Для мінімальної ентропії H=2.3408: R=0.53184

Для максимальної ентропії H=3.1328: R=0.37344

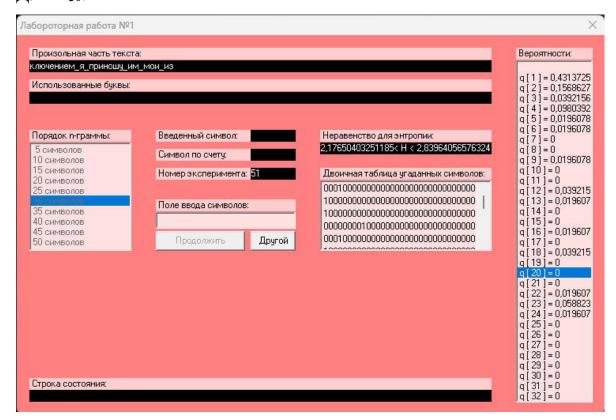
#### Для Н20:



Для мінімальної ентропії H=2.2101: R=0.55798

Для максимальної ентропії H=2.9202: R=0.41596

#### Для Нзо:



Для мінімальної ентропії H=2.1765: R=0.5647

Для максимальної ентропії H=2.8396: R=0.43208

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

Проведемо аналіз для результатів, які ми отримали у вибраному нами тексті. Бачимо, що ентропія літер з пробілами є трохи нижчою за ентропію літер без пробілів (0.414811 < 4.464745). Причиною цього може бути те, що пробіл — дуже поширений символ, тому його наявність впливає на ентропію тексту, зменшуючи  $\ddot{\text{п}}$ .

Поглянемо на ентропію біграм. Візьмемо біграми, що перетинаються і не перетинаються як два окремі випадки. Помічаємо тут схожу закономірність — ентропія для біграм, які містять пробіли,  $\epsilon$  нижчою, ніж у тих, де її немає. Можемо припустити, що причиною цього також  $\epsilon$  те, що пробіл дуже поширений символ, отже, його частота буде високою, знижуючи ентропію.

Також, при аналізі результатів було помічено, що надлишковість у біграмах з пробілами і без теж відрізняються — ті експерименти, які включають пробіли, показують вищий показник надлишковості ніж ті, які їх не включають. Тому можна зробити висновок, що пробіли крім зниження ентропії текстів також підвищують їх надлишковість.

**Висновок:** У цій роботі ми розібралися з поняттями ентропії та надлишковості джерела тексту. Ми проаналізували частоти символів і біграм у текстах, обчислили їх ентропію та надлишковість. Це допомогло зрозуміти, як різні моделі тексту впливають на кількість інформації, що міститься в ньому, і дало практичний досвід роботи з оцінкою цих характеристик.