

**Міністерство освіти і науки України 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**

Виконав:  
Дворніков Дмитро  
Група:  
ФБ-03

Київ - 2022

## 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  $(10) H$ ,  $(20) H$ ,  $(30) H$ .
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

## Хід роботи

Після написання програми, код якої буде разом з роботою ми отримали такі дані.

### Частота букв з пробілами

```
к : 0.15872
о : 0.09504
е : 0.07071
а : 0.06699
н : 0.05671
и : 0.05656
т : 0.05604
с : 0.04454
л : 0.0406
р : 0.03849
в : 0.03524
м : 0.02777
к : 0.02723
д : 0.02405
у : 0.02376
п : 0.02351
я : 0.01723
ь : 0.0167
ы : 0.0163
г : 0.01496
з : 0.0145
б : 0.01353
ч : 0.01269
й : 0.00935
х : 0.00838
ж : 0.00824
ш : 0.00744
ю : 0.00445
щ : 0.00358
```

Тут ми бачимо, що пробіл є найчастішою “буквою” у російській мові, за ним ідуть всі нижче наведені.

Усі файли з даними будуть прикріплені разом з кодом та протоколом.

### *Частота букв без пробілів*

```
о : 0.11297
е : 0.08405
а : 0.07963
н : 0.0674
и : 0.06723
т : 0.06661
с : 0.05294
л : 0.04826
р : 0.04575
в : 0.04189
м : 0.033
к : 0.03237
д : 0.02859
у : 0.02824
п : 0.02795
я : 0.02048
ь : 0.01985
ы : 0.01938
г : 0.01778
з : 0.01723
б : 0.01608
ч : 0.01509
й : 0.01111
х : 0.00996
ж : 0.0098
ш : 0.00884
ю : 0.00529
щ : 0.00426
```

Нижче наведені значення про ентропію та надлишковість російської мови основане на частоті букв.

### *Ентропія та надлишковість російської мови*

```
4.379441435294986 - Ентропія для тексту з пробілом
4.455388438097443 - Ентропія для тексту без пробілів
0.1391698432103622 - Надлишковість для тексту з пробілом
0.11676440565986546 - Надлишковість для тексту без пробілів
```

Далі буде наведені таблиці з частотами біграм з пробілами та без. Під час підрахунку значень були труднощі з написанням функції, так як я хотів зробити універсальну функцію. Але завдяки вкладці я зміг зрозуміти свою проблеми та вирішити її(проблема була у використанні мною неправильних методів для типу даних list)

Частота  
перехресних  
біграм

```

р* : 0.02271
и* : 0.01841
*н : 0.01697
е* : 0.01683
*п : 0.01635
*с : 0.01601
а* : 0.01391
то : 0.01377
*в : 0.01337
*о : 0.01106
ст : 0.01062
ь* : 0.01026
*и : 0.01016
на : 0.01002
я* : 0.00994
но : 0.00984
по : 0.00953
не : 0.00952
м* : 0.00896
*т : 0.00885
*к : 0.00816
ал : 0.00814
ни : 0.00809
ко : 0.00771

```

Частота  
не перехресних  
біграм

```

о* : 0.02284
и* : 0.01843
е* : 0.01703
*н : 0.0169
*п : 0.01611
*с : 0.01596
а* : 0.01397
то : 0.01362
*в : 0.01338
*о : 0.01108
ст : 0.01078
ь* : 0.0103
*и : 0.0101
я* : 0.01002
на : 0.00995
но : 0.00985
по : 0.00957
не : 0.00954
м* : 0.00896
*т : 0.00881
*к : 0.00818
ал : 0.00816
ни : 0.00808
от : 0.00775

```

Частота  
перехресних  
біграм без пробілу

```

то : 0.01679
ст : 0.01296
на : 0.01199
но : 0.01198
не : 0.01138
по : 0.01133
от : 0.01077
ов : 0.01015
он : 0.01005
ос : 0.01003
ни : 0.0099
ал : 0.00989
ко : 0.00949
го : 0.00895
ен : 0.00892
ро : 0.00886
ра : 0.00866
ли : 0.0086
та : 0.00851
ем : 0.00833
ер : 0.00796
те : 0.00793
ло : 0.00787
ан : 0.00758
пр : 0.00738

```

Частота  
не перехресних  
біграм без пробілу

```

то : 0.0169
ст : 0.01299
но : 0.01208
на : 0.01203
по : 0.01131
не : 0.01125
от : 0.0107
он : 0.01024
ов : 0.01015
ос : 0.01006
ал : 0.00989
ни : 0.00989
ко : 0.00945
ен : 0.00915
го : 0.00907
ро : 0.00897
ра : 0.00873
ли : 0.0085
та : 0.00844
ем : 0.00839
те : 0.00796
ер : 0.00789
ло : 0.00785
ан : 0.00756
ре : 0.00729
пр : 0.00728

```

Також для всіх цих значень треба порахувати ентропію та надлишковість.

Ентропія та надлишковість для біграм

```

3.976565682256298 - Ентропія для перехресних біграм з пробілом
3.9765024430760287 - Ентропія для не перехресних біграм без пробілів
4.136699739436335 - Ентропія для перехресних біграм з пробілом
4.13635500411463 - Ентропія для не перехресних біграм без пробілів
0.21835975881467418 - Надлишковість для перехресних біграм з пробілом
0.21837218921116897 - Надлишковість для не перехресних біграм з пробілом
0.17994120967644744 - Надлишковість для перехресних біграм без пробілів
0.1800095499594523 - Надлишковість для не перехресних біграм без пробілів

```

10 символів

Лабораторная работа №1

<b>Произвольная часть текста:</b>	<div>едует_ста</div>	
<b>Использованные буквы:</b>		
<b>Порядок n-граммы:</b>	<b>Введенный символ:</b>	<b>Неравенство для энтропии:</b>
<input type="radio"/> 5 символов <input checked="" type="radio"/> 10 символов <input type="radio"/> 15 символов <input type="radio"/> 20 символов <input type="radio"/> 25 символов <input type="radio"/> 30 символов <input type="radio"/> 35 символов <input type="radio"/> 40 символов <input type="radio"/> 45 символов <input type="radio"/> 50 символов	<b>Символ по счету:</b>  <b>Номер эксперимента:</b> 51	$2,62212919465741 < H < 3,37998720177651$
<b>Поле ввода символов:</b>	<b>Двоичная таблица угаданных символов:</b>	
<div>Продолжить Другой</div>	<pre> 10000000000000000000000000000000 00000000000001000000000000000000 01000000000000000000000000000000 00000010000000000000000000000000 00001000000000000000000000000000 .....           </pre>	
<b>Строка состояния:</b>	Вероятности:	
	q[1] = 0,4 q[2] = 0,08 q[3] = 0,08 q[4] = 0,06 q[5] = 0,04 q[6] = 0,02 q[7] = 0,02 q[8] = 0,02 q[9] = 0 q[10] = 0,02 q[11] = 0 q[12] = 0 q[13] = 0 q[14] = 0,02 q[15] = 0 q[16] = 0,02 q[17] = 0,02 q[18] = 0 q[19] = 0 q[20] = 0 q[21] = 0 q[22] = 0 q[23] = 0,02 q[24] = 0 q[25] = 0,02 q[26] = 0,02 q[27] = 0,04 q[28] = 0,02 q[29] = 0,02 q[30] = 0,04 q[31] = 0,02 q[32] = 0	

20 символів

Лабораторная работа №1

Произвольная часть текста:		Вероятности:	
<u>собираюсь_говорить_</u>		q[1] = 0,48 q[2] = 0,14 q[3] = 0,1 q[4] = 0,02 q[5] = 0,04 q[6] = 0 q[7] = 0,02 q[8] = 0,02 q[9] = 0 q[10] = 0 q[11] = 0 q[12] = 0,02 q[13] = 0 q[14] = 0 q[15] = 0,02 q[16] = 0 q[17] = 0 q[18] = 0 q[19] = 0 q[20] = 0,02 q[21] = 0,02 q[22] = 0,02 q[23] = 0 q[24] = 0 q[25] = 0,02 q[26] = 0 q[27] = 0,02 q[28] = 0 q[29] = 0,02 q[30] = 0,02 q[31] = 0 q[32] = 0	
Использованные буквы:			
Порядок n-граммы:	Введенный символ:	Неравенство для энтропии:	
<input type="radio"/> 5 символов	<input type="text"/>	<input type="text"/>	
<input type="radio"/> 10 символов	Символ по счету:	<input type="text"/>	
<input type="radio"/> 15 символов	Номер эксперимента:	<input type="text"/>	
<input checked="" type="radio"/> 20 символов		Двоичная таблица угаданных символов:	
<input type="radio"/> 25 символов	Поле ввода символов:	<input type="text"/>	
<input type="radio"/> 30 символов	<input type="text"/>	<input type="text"/>	
<input type="radio"/> 35 символов	<input type="button" value="Продолжить"/>	<input type="button" value="Другой"/>	
<input type="radio"/> 40 символов			
<input type="radio"/> 45 символов			
<input type="radio"/> 50 символов			
Строка состояния:			

**Лабораторная работа №1**

---

Произвольная часть текста:  
**знают\_этот\_закон\_\_человеческо**

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:

Неравенство для энтропии:  
 **$1,71007668351909 < H < 2,35715672729159$**

Двоичная таблица угаданных символов:

10000000000000000000000000000000
10000000000000000000000000000000
00100000000000000000000000000000
10000000000000000000000000000000
00000001000000000000000000000000
00000000000000000000000000000000

Вероятности:

$q[1] = 0,62$
$q[2] = 0,04$
$q[3] = 0,06$
$q[4] = 0,02$
$q[5] = 0,02$
$q[6] = 0$
$q[7] = 0$
$q[8] = 0,02$
$q[9] = 0$
$q[10] = 0$
$q[11] = 0,02$
$q[12] = 0,02$
$q[13] = 0$
$q[14] = 0$
$q[15] = 0$
$q[16] = 0,04$
$q[17] = 0$
$q[18] = 0,04$
$q[19] = 0$
$q[20] = 0$
$q[21] = 0,02$
$q[22] = 0,02$
$q[23] = 0,02$
$q[24] = 0$
$q[25] = 0$
$q[26] = 0$
$q[27] = 0$
$q[28] = 0$
$q[29] = 0,02$
$q[30] = 0$
$q[31] = 0$
$q[32] = 0,02$

Строка состояния:

---

Кіл-сть символів	H	R
10	3,001	0,90621875
20	2,38115	0,9255890625
30	2,0336	0,93645

Під час роботи на цією лабораторною дізнався багато нового про мову, а саме про методи її аналізу такі як ентропія і надлишковість. Також практично порахував значення для вибраного тексту.