

**"Київський політехнічний інститут імені Ігоря
Сікорського"
Фізико-технічний інститут**

Криптографія

Комп'ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:
Студенти ФБ-05
Береза О.А
Ковбель Д.О

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

Постановка задачі

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.

Хід роботи

Результати для тексту з пробілами:

Повні таблиці даних можна знайти у `lab1_result.xlsx`

```
Some calculations for rtext w/ spaces
```

```
Frequency of letters in rtext: {'н': 0.07327  
H1: 4.420639465250134  
Surplus: 0.13107189119760154
```

```
Frequency of bigrams in rtext {'на': 0.013906  
H2: 3.8090749049653234  
Surplus: 0.2512820193829324
```

```
Frequency of cross bigrams in rtext {'на': 0.  
H2: 2.1549096334190376  
Surplus: 0.5764274451409206
```

Частота букв:

' '	0.11416701202533162	'д'	0.01797463411960941
'о'	0.09745165264110899	'у'	0.01745288565716902
'и'	0.07830451158293003	'й'	0.016727917881130403
'а'	0.07720050553658461	'ч'	0.015079329741603217
'н'	0.07327312104906833	'б'	0.014218501862653428
'е'	0.07129344526379122	'ь'	0.011787861870234194
'т'	0.05633285040204315	'г'	0.010630196791646545
'с'	0.048107034927463266	'ц'	0.009039834410509864
'р'	0.04450160463340035	'х'	0.007753159318233453
'в'	0.042140607214742304	'ш'	0.0072976283849430475
'л'	0.02501766752178614	'ф'	0.0046580606712402425
'п'	0.024322383465711306	'ж'	0.004534759065236825
'м'	0.02209382110535323	'ю'	0.004043836004297289
'з'	0.021629156719766274	'щ'	0.0022959215710451294
'к'	0.020662152457869096	'э'	0.0019933759637219273
'ы'	0.01886857076313419	'ё'	0.0004178554425671394
'я'	0.01872814393407474		

Н1: 4.420639465250134

Надлишковість: 0.13107189119760154

Частота біграм (у тексті з пробілами):

Частота біграм у тексті з пробілами	
'ст'	0.018575158608107538
'ни'	0.018285171497692092
' п'	0.016030350461981435
'и '	0.015274557284441962
'ра'	0.015151255678438546
'но'	0.015006262123230822
'й '	0.014330386653286158
'ен'	0.01422535195187584
' о'	0.014112325479706039
'ов'	0.0140335494536483

Н2: 3.8090749049653234

Надлишковість: 0.2512820193829324

Частота перехресних біграм (у тексті з пробілами):

Частота перехресних біграм у тексті з пробілами	
'ст'	0,020902124
'ни'	0,020616006
'ра'	0,017210936
'ен'	0,017172272
'но'	0,016927395
'ов'	0,015973667
'на'	0,015659194
'ти'	0,013133102
'ро'	0,012973288
'то'	0,012367541

H2: 3.8098192668380726

Надлишковість: 0.2511357063982529

Результати для тексту без пробілів

```
Doing calculations for rtext w/o spaces
```

```
Frequency of letters in rtext: {'н': 0.08271663173957536, 'а':
```

```
H1: 4.4119841397811665
```

```
Surplus: 0.13277319609928806
```

```
Frequency of bigrams in rtext {'на': 0.015699147572760478, 'ач':
```

```
H2: 3.8634871834695197
```

```
Surplus: 0.24058665310664062
```

```
Frequency of cross bigrams in rtext {'на': 0.007829597036739177,
```

```
H2: 2.1820053424603176
```

```
Surplus: 0.5711014683452609
```

Частота букв (у тексті без пробілів):

'н'	0,082716632
'а'	0,087150181
'ч'	0,017022768
'и'	0,088396473
'м'	0,024941294
'о'	0,110011316
'с'	0,054307116
'т'	0,063593083
'ь'	0,013307093
'э'	0,002250284
'х'	0,008752394
'п'	0,027457076
'р'	0,050237014
'б'	0,016050996
'л'	0,028241969
'е'	0,080481813
'к'	0,023325111
'в'	0,047571729
'д'	0,020291222
'й'	0,018883828
'ы'	0,02130037
'ф'	0,005258396
'ю'	0,004565009
'з'	0,024416743
'ц'	0,010204897
'я'	0,021141845
'у'	0,01970223
'ж'	0,005119203
'щ'	0,002591822
'г'	0,012000227
'ш'	0,008238154
'ё'	0,000471709

Н1: 4.4119841397811665

Надлишковість: 0.13277319609928806

Частота біграм (у тексті без пробілів):

Частота біграм у тексті без пробілів	
'ст'	0,020969143
'ни'	0,020641782
'ра'	0,017103964
'ен'	0,017098809
'но'	0,016940284
'ов'	0,016071617
'на'	0,015699148
'ти'	0,013185944
'ро'	0,012943645
'то'	0,012381718

H2: 3.8634871834695197

Надлишковість: 0.24058665310664062

Частота перехресних біграм (у тексті без пробілів):

Частота перехресних біграм у тексті без пробілів	
'ст'	0,020902124
'ни'	0,020616006
'ра'	0,017210936
'ен'	0,017172272
'но'	0,016927395
'ов'	0,015973667
'на'	0,015659194
'ти'	0,013133102
'ро'	0,012973288
'то'	0,012367541

H2: 3.864010684920642

Надлишковість: 0.2404837528069318

$$0.17585 < R < 0.2800282$$

H(20):

[illegible]

$$2,6684744 < H(20) < 3,3136899$$

$$0,33726202 < R < 0,46630512$$

H(30):

Лабораторная работа №1

Произвольная часть текста:
_греции_и_риме_то_его_поразит_факт_насколько_эти_учения_были_похожи_друг_на_

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: _ (пробел)

Символ по счету: 1

Номер эксперимента: 51

Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:
3,60241468941491 < H < 4,13677023782443

Двоичная таблица угаданных символов:

0000000000000000010000000000000000
0000000000000100000000000000000000
0000000001000000000000000000000000
0000000000000000000000000000000000
0000000000000000000000000000000000
0000010000000000000000000000000000
~~~~~

Вероятности:

- q[1] = 0,2156862
- q[2] = 0,0784313
- q[3] = 0,0784313
- q[4] = 0,0588235
- q[5] = 0,0392156
- q[6] = 0,0196078
- q[7] = 0,0196078
- q[8] = 0,0392156
- q[9] = 0,0392156
- q[10] = 0
- q[11] = 0
- q[12] = 0
- q[13] = 0,019607
- q[14] = 0,039215
- q[15] = 0,019607
- q[16] = 0,019607
- q[17] = 0
- q[18] = 0,058823
- q[19] = 0
- q[20] = 0,058823
- q[21] = 0,019607
- q[22] = 0,019607
- q[23] = 0
- q[24] = 0,019607
- q[25] = 0
- q[26] = 0
- q[27] = 0,019607
- q[28] = 0,019607
- q[29] = 0,019607
- q[30] = 0,019607
- q[31] = 0,019607
- q[32] = 0,039215

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

$$3,602,414 < H(30) < 4,136,770$$

$$0.1726458 < R < 0.2795172$$



## **Висновки**

Під час виконання роботи було засвоєно поняття ентропії та надлишковості.

Також було побудовано алгоритм визначення ентропії та надлишковості для тексту з пробілами та без.

Завдяки виконанню роботи було набуто навичок оцінки ентропії на символ джерела