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

Криптографія

Комп'ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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 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.

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: {'H': 0.07327
H1: 4.420639465250134
Surplus: 0.13107189119760154

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

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

Частота букв:

0.11416701202533162	'д'	0.01797463411960941
0.09745165264110899	'y'	0.01745288565716902
0.07830451158293003	'й'	0.016727917881130403
0.07720050553658461	'ч'	0.015079329741603217
0.07327312104906833	'б'	0.014218501862653428
0.07129344526379122	'ь'	0.011787861870234194
0.05633285040204315	'r'	0.010630196791646545
0.048107034927463266	'ц'	0.009039834410509864
0.04450160463340035	'x'	0.007753159318233453
0.042140607214742304	П	0.0072976283849430475
0.02501766752178614	'ф'	0.0046580606712402425
0.024322383465711306	'ж'	0.004534759065236825
0.02209382110535323	'ю'	0.004043836004297289
0.021629156719766274	'щ'	0.0022959215710451294
0.020662152457869096	'э'	0.0019933759637219273
0.01886857076313419	'ë'	0.0004178554425671394
0.01872814393407474		
	0.09745165264110899 0.07830451158293003 0.07720050553658461 0.07327312104906833 0.07129344526379122 0.05633285040204315 0.048107034927463266 0.04450160463340035 0.042140607214742304 0.02501766752178614 0.024322383465711306 0.02209382110535323 0.021629156719766274 0.020662152457869096 0.01886857076313419	0.09745165264110899 'y' 0.07830451158293003 'й' 0.07720050553658461 'ч' 0.07327312104906833 '6' 0.07129344526379122 'ь' 0.05633285040204315 'r' 0.048107034927463266 'ц' 0.04450160463340035 'x' 0.042140607214742304 'ш' 0.02501766752178614 'ф' 0.024322383465711306 'ж' 0.02209382110535323 'ю' 0.021629156719766274 'щ' 0.020662152457869096 'э' 0.01886857076313419 'ë'

H1: 4.420639465250134

Надлишковість: 0.13107189119760154

Частота біграм (у тексті з пробілами):

Частота біграм у тексті з пробілами		
'ст'	0.018575158608107538	
'ни'	0.018285171497692092	
¹ n¹	0.016030350461981435	
'и '	0.015274557284441962	
'pa'	0.015151255678438546	
'но'	0.015006262123230822	
'й '	0.014330386653286158	
'ен'	0.01422535195187584	
¹ o¹	0.014112325479706039	
'OB'	0.0140335494536483	

H2: 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
'pa'	0,017210936
'ен'	0,017172272
'но'	0,016927395
'ов'	0,015973667
'на'	0,015659194
'ти'	0,013133102
'po'	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: {'H': 0.08271663173957536, 'a': H1: 4.4119841397811665
Surplus: 0.13277319609928806

Frequency of bigrams in rtext {'Ha': 0.015699147572760478, 'a': H2: 3.8634871834695197
Surplus: 0.24058665310664062

Frequency of cross bigrams in rtext {'Ha': 0.007829597036739177, H2: 2.1820053424603176
Surplus: 0.5711014683452609
```

Частота букв (у тексті без пробілів):

'н'	0,082716632
'a'	0,087150181
'ч'	0,017022768
'и'	0,088396473
'M'	0,024941294
'o'	0,110011316
'c'	0,054307116
'T'	0,063593083
'ь'	0,013307093
'э'	0,002250284
'x'	0,008752394
'n'	0,027457076
'p'	0,050237014
'ნ'	0,016050996
'л'	0,028241969
'e'	0,080481813
'к'	0,023325111
'в'	0,047571729
'д'	0,020291222
'й'	0,018883828
'ы'	0,02130037
'φ'	0,005258396
'ю'	0,004565009
'3'	0,024416743
'ц'	0,010204897
'я'	0,021141845
'y'	0,01970223
'ж'	0,005119203
'щ'	0,002591822
'r'	0,012000227
'ш'	0,008238154
'ë'	0,000471709

H1: 4.4119841397811665

Надлишковість: 0.13277319609928806

Частота біграм (у тексті без пробілів):

Частота біграм у тексті без пробілів	
'ст'	0,020969143
'ни'	0,020641782
'pa'	0,017103964
'ен'	0,017098809
'но'	0,016940284
'ов'	0,016071617
'на'	0,015699148
¹TИ¹	0,013185944
'po'	0,012943645
'то'	0,012381718

H2: 3.8634871834695197

Надлишковість: 0.24058665310664062

Частота перехресних біграм (у тексті без пробілів):

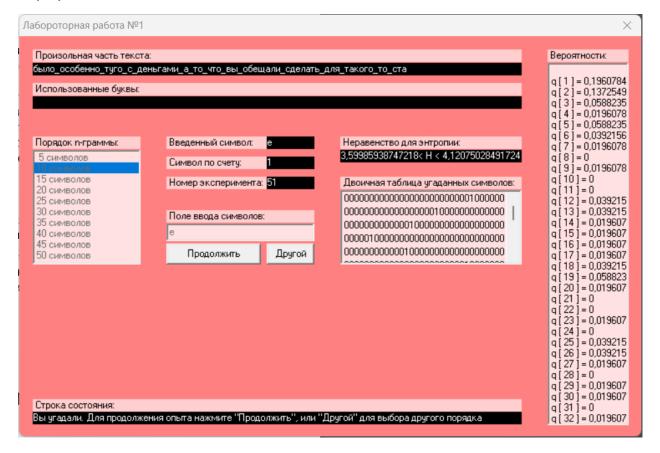
Частота перехресни	х біграм у тексті без пробілів
'ст'	0,020902124
'ни'	0,020616006
'pa'	0,017210936
'ен'	0,017172272
'но'	0,016927395
'ов'	0,015973667
'на'	0,015659194
'ти'	0,013133102
'po'	0,012973288
'то'	0,012367541

H2: 3.864010684920642

Надлишковість: 0.2404837528069318

CoolPinkProgram

H(10):



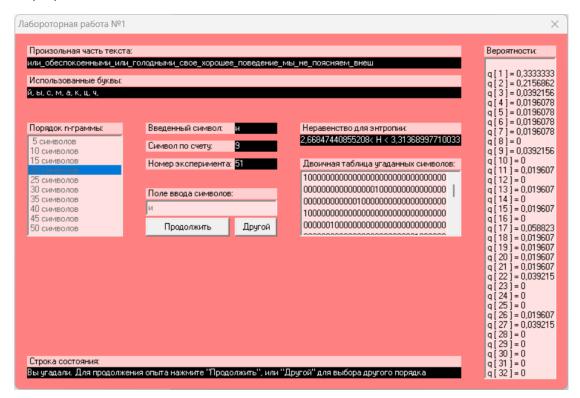
3,599859 < H10 < 4,120750

Для знаходження R використовуємо формулу:

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

0.17585 < R < 0.2800282

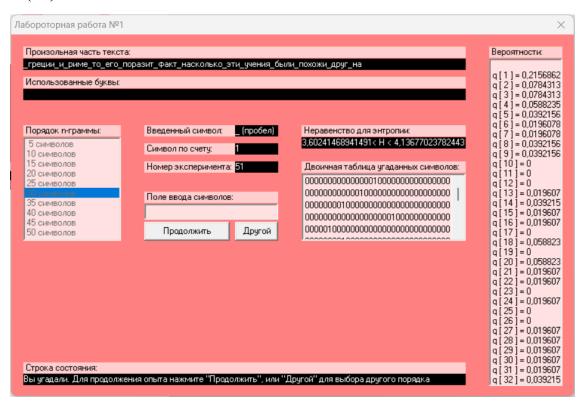
H(20):



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

0,33726202 < R < 0,46630512

H(30):



3,602414 < H(30) < 4,136770

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