

# КРИПТОГРАФІЯ

## КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

*ФБ-12 Приходько Юрій*

### **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:**

- Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- 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.
- За допомогою програми CoolPinkProgram оцінити значення  $H(10)$ ,  $H(20)$ ,  $H(30)$ .
- Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

### **Хід роботи:**

Напишемо програму мовою python що задовольнить поставлену задачу. Програма має підраховувати  $H_1$ ,  $H_2$ ,  $R_1$ ,  $R_2$  для текстів з пробілами і без. Для частотного аналізу біграм в тексті візьмемо їх з перетинами і без. Текст довільний, тому перед виконанням треба його спарсити. Крім функцій безпосереднього обрахування заданих значень з тексту, програма має мати презентабельний аутпут, напишемо відповідні функції.

### **Результати виконання програми:**

```

getinfo@dedsec: /Documents/work/cyber/loose/robo-15-04/cyber/loose/robo-15-04 $ ./entropy_assessment.py
[*] Counting the frequency for characters one by one...
[!] Frequency for each char:
| б:0.000246027695 | ё:0.000669964667 | ф:0.001505817714 | ц:0.002389754349 | щ:0.002424214254
| э:0.002945119796 | ю:0.004927766429 | х:0.005992016522 | ш:0.006726894033 | й:0.008098077698
| ж:0.009709679307 | 6:0.013782359253 | ы:0.014284832753 | з:0.014293648078 | р:0.015466086244
| ч:0.015738559913 | ь:0.018558662379 | я:0.019669393273 | у:0.021614374429 | н:0.022309983676
| д:0.025301584274 | м:0.025748761647 | к:0.026760119792 | п:0.032156701209 | л:0.038227254259
| в:0.038814675432 | с:0.043811361671 | т:0.051923063053 | и:0.052570588711 | н:0.055211179114
| а:0.065519499097 | е:0.073611165650 | о:0.092721186511 | :0.176269627120 |
[!] Entropy H1: 4.349367425456624808349713018627857431932176824808464486680182842260178072840703933366057754028588534
[!] Redundancy R1: 0.1450930333699103334624704760854187449754309124431588521674530753218999181230899474558517228208555034
|
[*] Same text without spaces and special chars...
| б:0.000298675031 | ё:0.000813330051 | ф:0.001828046850 | ц:0.002901136619 | щ:0.002942970501
| э:0.003575344425 | ю:0.005902256563 | х:0.007274244970 | ш:0.008166378531 | й:0.009830981065
| ж:0.011787448449 | 6:0.016731639025 | ы:0.017341636564 | з:0.017352338275 | р:0.01877565865
| ч:0.019186446039 | ь:0.022530020732 | я:0.023878436344 | у:0.026239622988 | н:0.027084085291
| д:0.030715856920 | м:0.031258725543 | к:0.032486503683 | п:0.039037896705 | л:0.044407484193
| в:0.047120607312 | с:0.053186531799 | т:0.063834051072 | и:0.063820141204 | н:0.067025790151
| а:0.079539982079 | е:0.089363180002 | о:0.112562544205 |
[!] H1 w/o spaces: 4.464397021682714272574624245341287651514282285597005988425275169962573411364115116839457186870276929
[!] R1 w/o spaces: 0.1149785452825607550625247546633613878157909019974695879989963664660830761134923311736189487018735658
|
[*] Counting frequency for bigrams with overlay...
[!] Frequency for bigrams:
> Show bigram frequencies table? [y/n] n
[!] Entropy H2: 3.949797797799579554765868547470816683473225392641756572829873967319345200756026525307218477678361520
[!] Redundancy R2: 0.2236212978749863629407498657022792096101365463541335886299976583438861335077737324971523466185541211
[!] Entropy H2' for text without overlapping: 3.949111384017690422896503214001087876120210399671443966081075680162905697259093269757546096343503451
[!] Redundancy H2' for text without overlapping: 0.2237562204882615953990830227107979745855308997516251492320762570097859700141302827095717449531151171
|
[*] Counting frequency for bigrams without spaces...
[!] Frequency for bigrams:
> Show bigram frequencies table? [y/n] n
[!] H2 w/o spaces: 4.133196459438742576090010326976047139282374820954705949336823975866886174469782590235045205417918626
[!] R2 w/o spaces: 0.180635699423818421363609162412772611595153062248539953714065238829840907104184530001718787212966837
[!] H2' w/o spaces, w/o overlapping: 4.13269989260173820628005101000552178917186758017629357488882014189114288887629336533002799569658238
[!] R2' w/o spaces, w/o overlapping: 0.1807341394264865533517074961391532132967012271384546799066739663737161834472662464473964631436663808
> Show bigram frequencies matrix? [y/n] n

```

Таким чином

*	H	R
Монограми з пробілом	4.349	0.145
Монограми без пробілу	4.464	0.114
Біграми з перетином з пробілом	3.950	0.223
Біграми з перетином без пробілу	3.949	0.224
Біграми без перетину з пробілом	4.133	0.181
Бігрми без перетину без пробілу	4.132	0.181

Програма створює таблиці частот для монограм, та матрицю для біграм.  
Для зручності запису частоти біграм в матриці вказані в %

Частота символів з пробілом:

```

[!] Frequency for each char:
| ь:0.000246027695 | ё:0.000669964667 | ф:0.001505817714 | ц:0.002389754349 | щ:0.002424214254
| э:0.002945119796 | ю:0.004927766429 | х:0.005992016522 | ш:0.006726894033 | й:0.008098077698
| ж:0.009709679307 | 6:0.013782359253 | ы:0.014284832753 | з:0.014293648078 | р:0.015466086244
| ч:0.015738559913 | ь:0.018558662379 | я:0.019669393273 | у:0.021614374429 | н:0.022309983676
| д:0.025301584274 | м:0.025748761647 | к:0.026760119792 | п:0.032156701209 | л:0.038227254259
| в:0.038814675432 | с:0.043811361671 | т:0.051923063053 | и:0.052570588711 | н:0.055211179114
| а:0.065519499097 | е:0.073611165650 | о:0.092721186511 | :0.176269627120 |

```

Частота символів без пробілу:

Матриця частот в % для біграм з перетином:

### Оцінка значень $H(10)$ , $H(20)$ , $H(30)$ за допомогою CoolPinkProgram

[illegible]

$$0.5124988142987461 < R < 0.6688909897862052$$

H (20)

Произвольная часть текста:  
ываем\_на\_себе\_такое\_сильное\_давление\_того\_закона\_или\_правила\_что\_не\_в\_состо

Использованные буквы:

Порядок n-граммы:

5	
10	
15	
20	
25	
30	
35	
40	
45	
50	

Введенный символ: (пробел)

Символ по счету: 1

Номер эксперимента: 50

Неравенство для энтропии:  
 $1.83195837413134 < H < 2.75489025085716$

Двоичная таблица угаданных символов:

00000000010000000000000000000000
00010000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000
00000100000000000000000000000000

Вероятности:

q[1]	= 0.46
q[2]	= 0.14
q[3]	= 0.08
q[4]	= 0.06
q[5]	= 0.04
q[6]	= 0.04
q[7]	= 0.04
q[8]	= 0.04
q[9]	= 0.02
q[10]	= 0.02
q[11]	= 0
q[12]	= 0.02
q[13]	= 0.02
q[14]	= 0
q[15]	= 0
q[16]	= 0
q[17]	= 0
q[18]	= 0.02
q[19]	= 0
q[20]	= 0
q[21]	= 0
q[22]	= 0
q[23]	= 0
q[24]	= 0
q[25]	= 0
q[26]	= 0
q[27]	= 0
q[28]	= 0
q[29]	= 0
q[30]	= 0
q[31]	= 0
q[32]	= 0

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

$$0.4584942756692265 < R < 0.6399072717981559$$

H (30)

Произвольная часть текста:  
сорятся\_между\_собой\_иногда\_это\_выглядит\_смешно\_иногда\_просто\_неприятно\_но\_к

Использованные буквы:

Порядок n-граммы:

5	
10	
15	
20	
25	
30	
35	
40	
45	
50	

Введенный символ: o

Символ по счету: 1

Номер эксперимента: 50

Неравенство для энтропии:  
 $1.8142892431584 < H < 2.54250114215797$

Двоичная таблица угаданных символов:

00100000000000000000000000000000
10000000000000000000000000000000
00001000000000000000000000000000
00100000000000000000000000000000
10000000000000000000000000000000

Вероятности:

q[1]	= 0.4
q[2]	= 0.24
q[3]	= 0.1
q[4]	= 0.04
q[5]	= 0.08
q[6]	= 0.02
q[7]	= 0.04
q[8]	= 0
q[9]	= 0.04
q[10]	= 0
q[11]	= 0
q[12]	= 0
q[13]	= 0
q[14]	= 0.02
q[15]	= 0
q[16]	= 0
q[17]	= 0
q[18]	= 0.02
q[19]	= 0
q[20]	= 0
q[21]	= 0
q[22]	= 0
q[23]	= 0
q[24]	= 0
q[25]	= 0
q[26]	= 0
q[27]	= 0
q[28]	= 0
q[29]	= 0
q[30]	= 0
q[31]	= 0
q[32]	= 0

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

$$0.500241833830253 < R < 0.6433803928179445$$

## **Висновки:**

При виконанні комп'ютерного практикуму я отримав навички аналізу частот монограм на біграм. Закріпив на практиці формули обрахування ентропії та надлишковості. Зміг порівняти та оцінити отримані значення для різних наборів вхідних алфавітів, спробував з двома різними за змістом текстами. При роботі з CoolPinkProgram застосував отримані дані для передбачення наступного символу, як найбільш ймовірного, у ситуаціях де знання слів російської мови не допомогло б ( наприклад нове слово після пробілу ).