

КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

ФБ-12 Юрченко Вікторія

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

Текст завантажую з ia801806.us.archive.org/29/items/Elektrosudorozhnaya-terapiya-Nelson-2005/Nelson_djvu.txt. Фільтрую його функцією `clear()`. Частота рахується шляхом ділення кількості елемента/елементів у тексті на загальну кількість зустрічей цих символів. Таблиці отриманих частот:

Частота букв у тексті:

Символ	Кількість	Частота
	148122	0.1453898
о	89579	0.0879267
е	79559	0.0780915
и	74957	0.0735744
а	65198	0.0639954
н	61250	0.0601202
т	56760	0.0557130
с	51411	0.0504627
р	42882	0.0420910
в	36626	0.0359504
л	32220	0.0316257
п	30460	0.0298982
к	27344	0.0268396
м	27028	0.0265295
д	23808	0.0233689
у	17824	0.0174952
я	17239	0.0169210
ы	14895	0.0146203
э	14273	0.0140097
г	13913	0.0136564
ч	13314	0.0130684
ь	11865	0.0116461
б	10982	0.0107794
й	10314	0.0101238
х	9627	0.0094494
э	7997	0.0078495
ж	6513	0.0063929
ц	5694	0.0055890
ю	4958	0.0048665
ш	4416	0.0043345
ф	3898	0.0038261
щ	3094	0.0030369
ё	580	0.0005693
ъ	192	0.0001885

Частота біграм у тексті з перетинами:

Символ	Кількість	Частота
п	18822	0.0184748
и	17114	0.0167983
е	16846	0.0165353
ст	15865	0.0155724
о	15364	0.0150806
а	14483	0.0142159
с	13774	0.0135199
в	13122	0.0128800
ен	12723	0.0124883
ни	12188	0.0119632
но	12179	0.0119544
н	11939	0.0117188
я	11257	0.0110494
т	9875	0.0096929
то	9815	0.0096340
пр	9797	0.0096163
ро	9133	0.0089645
и	8956	0.0087908
ра	8776	0.0086141
на	8413	0.0082578
ов	8315	0.0081616
в	8306	0.0081528
о	8076	0.0079270
по	8039	0.0078907
й	8038	0.0078897
ко	7819	0.0076748
ре	7549	0.0074098
к	7348	0.0072125
ри	7318	0.0071830
э	7246	0.0071124
м	7237	0.0071035
ос	7233	0.0070996
не	7214	0.0070809
ле	7155	0.0070230
м	7065	0.0069347
ан	6981	0.0068522
д	6558	0.0064370
ти	6305	0.0061887
т	6226	0.0061112
х	6218	0.0061033
ны	6207	0.0060925
ол	6191	0.0060768

...

ьр	1	0.0000010
фм	1	0.0000010
гю	1	0.0000010
оь	1	0.0000010
йз	1	0.0000010
кё	1	0.0000010
ёл	1	0.0000010
фб	1	0.0000010
гь	1	0.0000010
мц	1	0.0000010
дф	1	0.0000010
дю	1	0.0000010
чб	1	0.0000010
гщ	1	0.0000010
рь	1	0.0000010
ън	1	0.0000010
дщ	1	0.0000010
ьь	1	0.0000010
щю	1	0.0000010
ьу	1	0.0000010
цм	1	0.0000010
уё	1	0.0000010
ёк	1	0.0000010
лр	1	0.0000010
ьч	1	0.0000010
ьы	1	0.0000010
щв	1	0.0000010
ёч	1	0.0000010
ьй	1	0.0000010
цл	1	0.0000010
фэ	1	0.0000010
щц	1	0.0000010
щы	1	0.0000010
жц	1	0.0000010
рё	1	0.0000010
бй	1	0.0000010
лй	1	0.0000010
гй	1	0.0000010
вй	1	0.0000010
ёр	1	0.0000010
кя	1	0.0000010
жм	1	0.0000010
юз	1	0.0000010
пч	1	0.0000010
фд	1	0.0000010

Частота біграм у тексті без перетинів:

Символ	Кількість	Частота
п	9445	0.0185416
и	8508	0.0167021
е	8465	0.0166177
ст	7981	0.0156676
о	7835	0.0153810
а	7158	0.0140519
с	6844	0.0134355
в	6546	0.0128505
ен	6421	0.0126051
но	6072	0.0119200
ни	6011	0.0118002
н	5996	0.0117708
я	5489	0.0107755
т	4901	0.0096212
то	4900	0.0096192
пр	4847	0.0095152
ро	4632	0.0090931
и	4466	0.0087672
ра	4446	0.0087280
ов	4174	0.0081940
в	4144	0.0081351
на	4140	0.0081273
й	4063	0.0079761
о	4058	0.0079663
по	3991	0.0078348
ко	3917	0.0076895
ре	3792	0.0074441
э	3692	0.0072478
ри	3655	0.0071752
к	3646	0.0071575
ос	3606	0.0070790
м	3598	0.0070633
м	3582	0.0070319
не	3570	0.0070083
ле	3547	0.0069631
ан	3496	0.0068630
д	3342	0.0065607
х	3135	0.0061543
ти	3131	0.0061465
т	3106	0.0060974
ол	3087	0.0060601
ес	3085	0.0060562

...

дф	1	0.0000020
дю	1	0.0000020
чб	1	0.0000020
эх	1	0.0000020
юи	1	0.0000020
ън	1	0.0000020
аъ	1	0.0000020
дщ	1	0.0000020
чп	1	0.0000020
тх	1	0.0000020
ьь	1	0.0000020
уё	1	0.0000020
иы	1	0.0000020
лр	1	0.0000020
шш	1	0.0000020
ьч	1	0.0000020
гы	1	0.0000020
ээ	1	0.0000020
ьы	1	0.0000020
щр	1	0.0000020
кд	1	0.0000020
ёч	1	0.0000020
ьй	1	0.0000020
гш	1	0.0000020
фэ	1	0.0000020
щц	1	0.0000020
шы	1	0.0000020
ний	1	0.0000020
йф	1	0.0000020
жц	1	0.0000020
рё	1	0.0000020
цг	1	0.0000020
пм	1	0.0000020
бй	1	0.0000020
що	1	0.0000020
лй	1	0.0000020
гй	1	0.0000020
вй	1	0.0000020
лф	1	0.0000020
хп	1	0.0000020
кя	1	0.0000020
жм	1	0.0000020
хж	1	0.0000020
фд	1	0.0000020
тф	1	0.0000020

Для обрахунку Н1 використовую формулу: $H1 = -\sum_{i=1}^n q_i \log_2 q_i$, де q – частота, n – кількість символів. Для Н2: $H2 = -\frac{1}{2} \sum_{i,j} q_{ij} \log_2 q_{ij}$.

Отримані значення ентропії:

H1 з пробілами: 4.402461438763894
 H1 без пробілів: 4.451479170192825
 H2 з пробілами та перетинами: 3.9970844782541084
 H2 з пробілами та без перетинів: 3.997104842702902
 H2 без пробілів та з перетинами: 4.1145792748163235
 H2 без пробілів та перетинів: 4.1148592017574375

Надлишковість рахується за формулою $R = 1 - \frac{H_\infty}{H_0}$; H_∞ дорівнюватиме значенню, порахованому вище, а $H_0 = \log_2 m$, де m – кількість символів у алфавіті (для алфавіту з пробілами: $m = 34$; без: $m = 33$).

Отримані значення надлишковості:

```
R для букв з пробілами: 0.13464499375450822
R для букв без пробілів: 0.117539378394374
R для біграм з пробілами та перетинами: 0.21432655078189222
R для біграм з пробілами та без перетинів: 0.21432254791259786
R для біграм без пробілів та з перетинами: 0.18432636755598786
R для біграм без пробілів та перетинів: 0.18427087487748361
```

Отже:

	Ентропія	Надлишковість
Букви з пробілами	4,402461	0,134645
Букви без пробілів	4,451479	0,117539
Біграми з пробілами та перетинами	3,997085	0,214327
Біграми з пробілами та без перетинів	3,997105	0,214323
Біграми без пробілів та з перетинами	4,114579	0,184326
Біграми без пробілів та перетинів	4,114859	0,184271

CoolPinkProgram

H(10)

The screenshot shows a software window titled "Лабораторная работа №1" (Laboratory work №1). The interface is divided into several sections:

- Top Section:** Contains input fields for "Произвольная часть текста:" (Arbitrary part of the text) with the value "_to_или_n", "Использованные буквы:" (Used letters), and "Порядок n-граммы:" (Order of n-grams) with a dropdown menu showing options from 5 to 50 symbols.
- Input Section:** Includes fields for "Введенный символ:" (Entered symbol), "Символ по счету:" (Symbol by count), and "Номер эксперимента:" (Experiment number) set to 52. There is also a "Поле ввода символов:" (Symbol input field).
- Buttons:** Two buttons are present: "Продолжить" (Continue) and "Другой" (Other).
- Entropy Calculation:** A section titled "Неравенство для энтропии:" (Entropy inequality) displays the formula $2.63534517429844 < H < 3.32441669721305$.
- Binary Table:** A section titled "Двоичная таблица угаданных символов:" (Binary table of guessed symbols) shows a grid of 0s and 1s.
- Probability List:** A vertical list on the right side, titled "Вероятности:" (Probabilities), shows values for $q[1]$ through $q[32]$, mostly around 0.0196078.
- Status Bar:** At the bottom, a "Строка состояния:" (Status bar) is visible.

The screenshot displays a software interface for calculating entropy. It is titled "Лабораторная работа №1" (Laboratory Work No. 1).

Top Section:

- Произвольная часть текста:** A text input field containing "к_делающий_их_не_пр".
- Использованные буквы:** An empty text input field.

Left Panel:

- Порядок n-граммы:** A list of values from 5 to 50. The value 15 is selected and highlighted in blue.
- Полное сообщение:** A text input field containing "к_делающий_их_не_пр".
- Строка состояния:** An empty text input field.

Center Section:

- Введенный символ:** An empty text input field.
- Символ по счету:** An empty text input field.
- Номер эксперимента:** A text input field containing "52".
- Поле ввода символов:** A large text input field.
- Buttons:** "Продолжить" (Continue) and "Другой" (Other).

Right Section:

- Неравенство для энтропии:** A text input field containing the inequality $1,9904432597629 < H < 2,65054861997267$.
- Двоичная таблица угаданных символов:** A table with 32 rows and 2 columns. The first row contains "10000000000000000000000000000000". The second row contains "10000000000000000000000000000000". The third row contains "10000000000000000000000000000000". The fourth row contains "10000000000000000000000000000000". The fifth row contains "10000000000000000000000000000000". The sixth row contains "10000000000000000000000000000000". The seventh row contains "10000000000000000000000000000000". The eighth row contains "10000000000000000000000000000000". The ninth row contains "10000000000000000000000000000000". The tenth row contains "10000000000000000000000000000000". The eleventh row contains "10000000000000000000000000000000". The twelfth row contains "10000000000000000000000000000000". The thirteenth row contains "10000000000000000000000000000000". The fourteenth row contains "10000000000000000000000000000000". The fifteenth row contains "10000000000000000000000000000000". The sixteenth row contains "10000000000000000000000000000000". The seventeenth row contains "10000000000000000000000000000000". The eighteenth row contains "10000000000000000000000000000000". The nineteenth row contains "10000000000000000000000000000000". The twentieth row contains "10000000000000000000000000000000". The twenty-first row contains "10000000000000000000000000000000". The twenty-second row contains "10000000000000000000000000000000". The twenty-third row contains "10000000000000000000000000000000". The twenty-fourth row contains "10000000000000000000000000000000". The twenty-fifth row contains "10000000000000000000000000000000". The twenty-sixth row contains "10000000000000000000000000000000". The twenty-seventh row contains "10000000000000000000000000000000". The twenty-eighth row contains "10000000000000000000000000000000". The twenty-ninth row contains "10000000000000000000000000000000". The thirtieth row contains "10000000000000000000000000000000". The thirty-first row contains "10000000000000000000000000000000". The thirty-second row contains "10000000000000000000000000000000".
- Вероятности:** A list of probabilities for each symbol in the alphabet. The values are: $q[1] = 0,5490196$, $q[2] = 0,1176470$, $q[3] = 0,0196078$, $q[4] = 0,0196078$, $q[5] = 0,0196078$, $q[6] = 0$, $q[7] = 0$, $q[8] = 0$, $q[9] = 0,0196078$, $q[10] = 0$, $q[11] = 0,019607$, $q[12] = 0,019607$, $q[13] = 0$, $q[14] = 0$, $q[15] = 0,019607$, $q[16] = 0$, $q[17] = 0$, $q[18] = 0,019607$, $q[19] = 0,019607$, $q[20] = 0,019607$, $q[21] = 0$, $q[22] = 0,019607$, $q[23] = 0$, $q[24] = 0$, $q[25] = 0,019607$, $q[26] = 0,039215$, $q[27] = 0$, $q[28] = 0$, $q[29] = 0$, $q[30] = 0,019607$, $q[31] = 0$, $q[32] = 0,039215$.

The screenshot shows a software interface for a laboratory experiment on entropy. The interface is divided into several sections:

- Top Section:** Contains three input fields. The first is labeled "Произвольная часть текста:" (Arbitrary part of the text) and contains the text "ытываем_на_себе_такое_сильное". The second is labeled "Использованные буквы:" (Used letters) and is empty.
- Left Panel:** A vertical list of options for "Порядок n-граммы:" (Order of n-grams). The options are 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 symbols. The option "30 символов" (30 symbols) is currently selected and highlighted in blue.
- Center Section:** Contains three input fields. The first is labeled "Введенный символ:" (Entered symbol) and is empty. The second is labeled "Символ по счету:" (Symbol by count) and is empty. The third is labeled "Номер эксперимента:" (Experiment number) and contains the value "52". Below these is a section labeled "Поле ввода символов:" (Symbol input field) with a text input area and two buttons: "Продолжить" (Continue) and "Другой" (Other).
- Right Section:** Contains two main areas. The top area is labeled "Вероятности:" (Probabilities) and displays a list of probabilities $q[1]$ through $q[32]$. The bottom area is labeled "Неравенство для энтропии:" (Entropy inequality) and displays the inequality $1.29197733279381 < H < 1.83479688884579$. Below this is a section labeled "Двоичная таблица угаданных символов:" (Binary table of guessed symbols) which displays a grid of 0s and 1s.

$R = 1 - \frac{H_{\infty}}{H_0}$; H_{∞} дорівнюватиме значенню, що видає програма, а $H_0 = \log_2 m$, де m – кількість символів у алфавіті (у даному випадку $m=32$), тому $H_0 = 5$.

	Ентропія	Надлишковість
H(10)	2,635345<H<3,324417	0,335117<R<0,472931
H(20)	1,990443<H<2,650549	0,469890<R<0,601911
H(30)	1,291977<H<1,834797	0,633041<R<0,741605

Висновок:

Під час виконання комп'ютерного практикуму, я практичним шляхом навчилася підраховувати частоти букв та біграм на довільно обраному тексті, які у подальшому використовувала для визначення значення ентропії на символ джерела та його надлишковості. За допомогою програми CoolPinkProgram отримала значення умовних ентропій джерела для 10-грам, 20-грам, 30-грам, які були використані для оцінки надлишковості.