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.

Спочатку я зробила програму, що поділить текст на літери та біграми, а потім підрахує частоти. Щоб не винаходити велосипед я пошукала у Інтернеті, як це роблять на Python. Там я знайшла бібліотеку collections. Вона дозволяє підрахувати частоти в одну строку, але виводить результат у вигляді collections. Та це не проблема, бо за ще одну строку усе це можна перетворити на словник.

За допомогою циклу рахую H1, H2. Я бачила, що можно за допомогою якоїсь бібліотеки це зробити, але там треба було і для частот використовувати її. А так попрактикувавшись у math за формулою, що є у теоритичних відомостях, підрахувала H1, H2.

$$H(Z) = -\sum_{i=1}^{n} p_i \log p_i \ .$$

та

$$\text{, де } H_n = \frac{1}{n} H(x_1, x_2, \ldots, x_n),$$

Отримала щось таке:

```
Counter({'о': 25052, 'е': 18536, 'а': 18166, 'и': 15690, 'н': 15584, 'т': 13914, 'с': 11960,
With space:  Counter({' ': 41620, 'о': 25052, 'е': 18536, 'а': 18166, 'и': 15690, 'н': 1558
Counter({'то': 3852, 'ст': 2964, 'но': 2830, 'ен': 2629, 'на': 2611, 'по': 2460, 'не': 2437,
With space:  Counter({'о ': 5452, 'и ': 4468, ' п': 4343, 'е ': 4185, ' с': 4099, ' н': 403
226135
4.469026386584125
4.155944245731662
For case with space:
4.91844461322267
4.608150620555682
```

H1 без пробілів = 4.469

H2 без пробілів = 4.156

H1 з пробілами = 4.918

H2 з пробілами = 4.608

В Інтернеті знайшла, що я мала отримати десь H1=4.358 H2=3.52. Мій результат може бути не таким самим, але він близький до цих значень. (Можна подивитись таблиці частот в csv файлах)

Таблиця, що демонструє кількість букв:

буквы	кількість
‘ ‘	41620
о	25052
е	18536
а	18166
и	15690
н	15584
т	13914
с	11960
л	11328
р	10368
в	9992
к	7590
м	7188
д	7047
у	6325
п	6249
я	4661
ы	4392
ь	4123
б	3935
г	3861
з	3769
ч	3390
й	2452
ж	2136
х	1968
ш	1671
ю	1263
ц	1136
э	857
щ	697
ф	437
ё	343
ъ	55

Тепер перейдемо до частини з експерементами.

Номер експеременту	H(10)	H(20)	H(30)
1	2<H<0	0<H<0	0<H<0
2	2<H<0	0<H<0	1.95<H<1
3	1.33<H<0.91	0.66<H<0.91	1.3<H<0.91
4	2.39<H<1.5	0.5<H<0.81	0.97<H<0.81
5	2.6<H<1.9	0.95<H<1.37	1.94<H<1.37
6	2.5<H<1.79	1.39<H<1.79	1.61<H<1.25
7	2.68<H<1.84	2.02<H<2.12	2.16<H<1.66
8	2.59<H<1.75	2.11<H<2.15	1.89<H<1.54
9	2.3<H<1.8	2.1<H<2.19	1.68<H<1.44
10	2.4<H<2.12	2.09<H<2.17	1.71<H<1.77
11	2.4<H<2.1	2.08<H<2.11	1.56<H<1.67
12	2.4<H<2.1	2.35<H<2.35	1.43<H<1.58
13	2.25<H<2.13	2.17<H<2.31	1.32<H<1.5
14	2.32<H<2.15	2.02<H<2.26	1.37<H<1.62
15	2.16<H<2.14	1.88<H<2.2	1.27<H<1.55
16	2.2<H<2.35	1.97<H<2.4	1.19<H<1.49
17	2.33<H<2.5	1.85<H<2.34	1.12<H<1.43
18	2.38<H<2.52	1.75<H<2.28	1.42<H<1.66
19	2.44<H<2.54	1.66<H<2.22	1.59<H<1.87
20	2.58<H<2.7	1.57<H<2.16	1.61<H<1.93
21	2.46<H<2.67	1.5<H<2.1	1.53<H<1.87
23	2.44<H<2.64	1.64<H<2.27	1.61<H<2.05
24	2.5<H<2.7	1.56<H<2.21	1.54<H<1.99
25	2.4<H<2.67	1.72<H<2.37	1.48<H<1.94
26	2.3<H<2.6	1.79<H<2.44	1.42<H<1.89
27	2.23<H<2.6	1.72<H<2.38	1.49<H<1.97
28	2.22<H<2.59	1.66<H<2.33	1.43<H<1.93
29	2.14<H<2.55	1.6<H<2.28	1.54<H<2.08
30	2.26<H<2.6	1.67<H<2.32	1.58<H<2.22
31	2.18<H<2.58	1.61<H<2.28	1.53<H<2.17
32	2.1<H<2.55	1.56<H<2.23	1.55<H<2.21
33	2.15<H<2.56	1.51<H<2.19	1.5<H<2.16
34	2.08<H<2.52	1.53<H<2.21	1.51<H<2.18
35	2.02<H<2.49	1.54<H<2.22	1.47<H<2.14
36	1.96<H<2.46	1.6<H<2.25	1.43<H<2.1
37	1.9<H<2.43	1.7<H<2.37	1.39<H<2.06
38	1.86<H<2.39	1.65<H<2.33	1.4<H<2.08
39	1.86<H<2.39	1.61<H<2.3	1.37<H<2.04
40	1.94<H<2.5	1.62<H<2.3	1.38<H<2.06
41	1.94<H<2.49	1.58<H<2.27	1.35<H<2.02
42	2.02<H<2.55	1.54<H<2.24	1.38<H<2.09
43	2.05<H<2.56	1.5<H<2.2	1.43<H<2.2
44	2.09<H<2.66	1.51<H<2.21	1.52<H<2.31
45	2.04<H<2.63	1.48<H<2.18	1.58<H<2.42
46	2<H<2.6	1.44<H<2.15	1.55<H<2.38
47	1.95<H<2.57	1.41<H<2.12	1.62<H<2.44
48	1.97<H<2.62	1.38<H<2.09	1.65<H<2.48
49	1.93<H<2.6	1.35<H<2.07	1.62<H<2.44

50	0<H<0	1.33<H<2.04	1.65<H<2.47
Среднее	2.06<H<2.12	1.52<H<1.92	1.44<H<1.75
Избыточность языка	(1-1.28/5)*100%=75**%<**H<72**%**=(1-1.41/5)*100%		

Як я зрозуміла, треба було самим вгадувати букву. Десь в Інтернеті я читала, що надлишковість російської мови десь 72-76%, але в нашому експерименті було одне салабке звено – тобто я. Звідки я взяла 1.28 та 1.41? Я намалювала графік залежності H від n. Графік був схожим на гіперболу. Тож по трьом точкам легко було знайти рівняння функції та її lim. Ліміти функцій це і є 1.28 та 1.41. Надлишковість мови по формулі $R = 1 - \frac{H_\infty}{H_0}$.

Надлишковість для H1 і H2 десь 10%. Це можна подивитись у программі crypto_lab1.py.

Висновок: Засвоїла понять ентропії на символ джерела та його надлишковості, вивчила та порівняла різних моделей джерела відкритого тексту для наближеного визначення ентропії, набула практичних навичок щодо оцінки ентропії на символ джерела.