КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав

ФБ-12 Сущенко Олександр

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 1 Н та 2 Н за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1 Н та 2 Н на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1 Н та 2 Н 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

Створюємо програму, яка спочатку буде фільтрувати за критеріями текст, створюючи два нових з пробілами та без відповідно. Далі обчислюємо значення частот символів та їх ентропію, використовуючи формулу:

$$H_1 = -\sum_{i=1}^{n} p(i) \log_2 p(i)$$

Далі рахуємо частоти біграм з перетинами та без; з пробілами та без. Наступний етап – обчислення ентропії, використовуємо наступну формулу:

$$H_2 = -\sum_i i, j p(i, j) \log_2 p(i, j)/2$$

Потім рахуємо надлишковість за допомгою іншої формули:

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

Результати виконання:

Частота символів без пробілів:

- o 0.11223058354016847
- e 0.08425898160360601
- a 0.07951647652602405
- и 0.07499076263817471
- н 0.06438223609826496
- т 0.05865863220583453
- c 0.05533634678361787
- л 0.04760465515563471
- p 0.04738786396590207
- в 0.04736729253913912
- м 0.032533711425607785
- к 0.030935469807870788
- л 0.03041722809518874
- п 0.02819630521351163
- y 0.02523560217708992
- я 0.019903646601877222

- ы 0.01914487743781297
- г 0.01809256983801584
- ь 0.01760281163931326
- 6 0.016867778736898573
- з 0.016728526001887824
- ч 0.012957625234494485
- й 0.011432174819149467
- x 0.010305493599517046
- ж 0.009618724429123126
- ш 0.007281493865363177
- ю 0.006091515947998598
- ц 0.005409494029934591
- ш 0.0037978018639295064
- э 0.0035580656212689564
- Φ 0.0018625053307687622
- ъ 0.0002650549217534135
- ë 1.582417443303961e-06

```
Частота символів з пробілами:
```

- 0.15449078722822107
- o 0.09489191725865267
- e 0.07124168883757118
- a 0.06723186027547
- и 0.06340533051650793
- н 0.054435730689905926
- т 0.04959637469754025
- c 0.04678735398812037
- л 0.04025014265864879
- p 0.0400668438555097
- в 0.04004945053842351
- м 0.027507530971811453
- к 0.02615620402896121
- д 0.02571802623313601
- п 0.02384021696156144
- v 0.021336917248618067
- я 0.016828703254624114

- ы 0.01618715744363729
- г 0.0152974223773052
- ь 0.014883327635907032
- 6 0.014261851036942736
- з 0.014144111659743903
- ч 0.01095578284309823
- й 0.009666001483783743
- x 0.00871338288644773
- ж 0.008132713685262576
- ш 0.006156565274777617
- ю 0.005150428778714861
- ц 0.004573773419934213
- ш 0.0032110739236045375
- э 0.003008374882177001
- Ф 0.001574764170034392
- ъ 0.00022410620091823334
- ë 1.3379474681685574e-06

Частоти біграм без пробілу без перетину: .

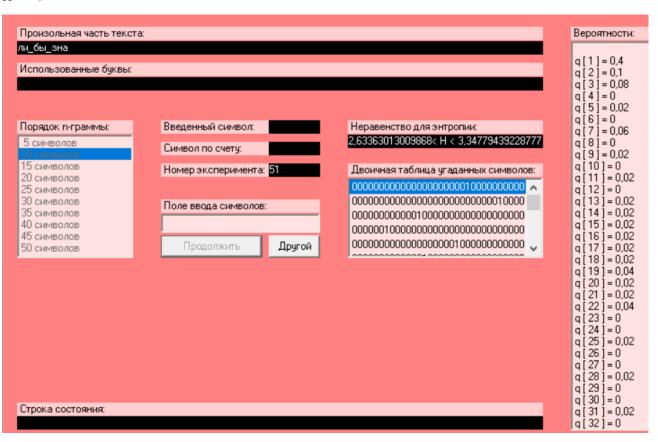
{'cт': 18879, 'то': 17823, 'на': 15387, 'но': 14631, 'eн': 14198, 'ов': 14034, 'по': 13551, 'ни': 13066, 'ра': 13068, 'ос': 12884, 'ко': 12155, 'не': 12103, 'во': 11788, 'го': {'cт': 0.01430427379641234, 'то': 0.014101724203410429, 'на': 0.012174338232501613, 'но': 0.011576183965667845, 'ен': 0.011233590318129455, 'ов': 0.011103831985112604, 'по': 0.0111038319851

{' o': 49447, ' a': 51875, ' e': 26799, ' ' ': 23665, 'o ': 22656, 'H ': 21995, 'H ': 21442, 'T ': 20384, 'oo': 19889, ' p': 15897, 'e ': 15255, ' T': 14115, 'a ': 15765, 'aH'
{' o': 0.027058016824711924, ' a': 0.021322228354489655, ' e': 0.01792183031458532, ' ': 0.01518312845985316, 'o ': 0.015156289197732176, 'H ': 0.014714096967872493, 'H ': 0.1

```
Ентропія Н1 без пробілів: 4.457100157745117
Ентропія Н1 з пробілами: 4.389477048616659
Ентропія Н2 без пробілів без перетину: 4.150057053053292
Ентропія Н2 без пробілів з перетином: 4.3666412328157636
Ентропія Н2 з пробілами без перетину: 3.9983739015880295
Ентропія Н2 з пробілами з перетином: 4.2698211786626095

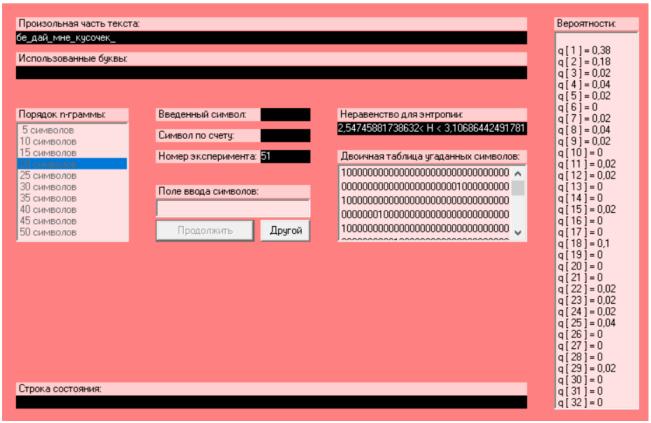
Надлишковість Н1 без пробілів: 0.11642507459112428
Надлишковість Н1 з пробілами: 0.13719722667539658
Надлишковість Н2 без пробілів без перетину: 0.17729325765269577
Надлишковість Н2 без пробілів з перетином: 0.13435763949167523
Надлишковість Н2 з пробілами без перетину: 0.2140730996267376
Надлишковість Н2 з пробілами з перетином: 0.16071697977980293
```

$H^{(10)}$:



0.330441121542446 < R < 0.473273973980264

Лабороторная работа №1



0.378627115016438 < R < 0.490508236522736

 $H^{(30)}$:

Лабороторная работа №1



0,432443812140248 < R < 0.574843682564354

Висновки

Під час виконання цієї лаборатоної роботи, я набув навичок оцінки частот букв та біграм. За допомогою практичної частини я навчився розраховувати ентропію та надлишковість. Також цікавою виявилася можливість застосування цих знань при роботі з CoolPinkProgram. Я використовував ці дані для передбачення наступного символу, вибираючи його на основі ймовірності, хоча в деяких випадках всеодно виникали складності.