

НТУУ «Київський політехнічний інститут ім. Ігоря Сікорського»

Навчально-науковий Фізико-технічний інститут

Криптографія

Комп'ютерний практикум №1

*Експериментальна оцінка ентропії на символ джерела
відкритого тексту*

Варіант №6

Виконали:

Студенти 3 курсу НН ФТІ

групи ФБ-31

Гаврилюк Володимир

Гек Роман

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

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 видалено 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.

Виконання роботи

Було проаналізовано текст розміром $>1\text{MB}$, текст був попередньо відфільтрований як і було вказано у методичних вказівках

Був написаний скрипт мовою Python, який обчислює:

- частоти появи кожного символу в тексті;
- ентропію H_1 на основі частот символів;
- частоти появи біграм двома методами: з перетином (крок 1 символ) та - без перетину (крок 2 символи)
- ентропію H_2 для обох методів підрахунку біграм
- усі розрахунки були проведені для двох варіантів тексту: з пробілами та з видаленими пробілами

Результати програмного аналізу:

Таблиця частот символів (з пробілами)

Таблиця частот символів:		
Символ	Кількість	Ймовірність

—	122779	0.145642
о	73463	0.087143
е	60012	0.071187
а	59924	0.071083
н	47699	0.056581
и	46785	0.055497
т	43940	0.052122
с	36680	0.043510
в	32767	0.038869
р	32606	0.038678
л	32570	0.038635
д	24672	0.029266
к	23715	0.028131
м	23344	0.027691
у	22256	0.026400
п	19054	0.022602
ь	16090	0.019086
я	15453	0.018331
ы	13743	0.016302
б	13402	0.015898
г	13320	0.015800
з	12345	0.014644
й	10931	0.012966
ч	10770	0.012775
ж	7971	0.009455
х	7116	0.008441
ш	6673	0.007916
ю	5568	0.006605
ц	3532	0.004190
щ	1748	0.002073
ф	1109	0.001316
э	983	0.001166

Таблиця частот символів (без пробілів)

Таблиця частот символів:		
Символ	Кількість	Ймовірність

о	73463	0.101998
е	60012	0.083322
а	59924	0.083200
н	47699	0.066226
и	46785	0.064957
т	43940	0.061007
с	36680	0.050927
в	32767	0.045494
р	32606	0.045271
л	32570	0.045221
д	24672	0.034255
к	23715	0.032926
м	23344	0.032411
у	22256	0.030901
п	19054	0.026455
ь	16090	0.022340
я	15453	0.021455
ы	13743	0.019081
б	13402	0.018608
г	13320	0.018494
з	12345	0.017140
й	10931	0.015177
ч	10770	0.014953
ж	7971	0.011067
х	7116	0.009880
ш	6673	0.009265
ю	5568	0.007731
ц	3532	0.004904
щ	1748	0.002427
ф	1109	0.001540
э	983	0.001365

Підсумкові значення ентропій:

```
З ПРОБІЛАМИ:  
H1 = 4.4356 біт  
H2 (перетин) = 4.0686 біт  
H2 (без перетину) = 4.0690 біт  
  
БЕЗ ПРОБІЛІВ:  
H1 = 4.4908 біт  
H2 (перетин) = 4.1854 біт  
H2 (без перетину) = 4.1841 біт
```

Надлишковість:

```
Модель H1:  
R (з пробілами) = 0.1207 (12.07%)  
R (без пробілів) = 0.1018 (10.18%)  
  
Модель H2 (з перетином):  
R (з пробілами) = 0.1934 (19.34%)  
R (без пробілів) = 0.1629 (16.29%)  
  
Модель H2 (без перетину):  
R (з пробілами) = 0.1934 (19.34%)  
R (без пробілів) = 0.1632 (16.32%)
```

Далі було проведено експерименти з CoolPinkProgram.exe, для кожної програми по 50 експериментів. Під час експериментів навмисно підроблювати результати не намагалися, як виходило так і вгадували символи.

Эксперименты из CoolPinkProgram

N10

Лабораторная работа №1

Произвольная часть текста:
сегда_он_старается_показать_что_то_что_он_сделал_на_самом_деле_не_идет_враз

Использованные буквы:
п, о, р, д, м,

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ: с

Символ по счету: 6

Номер эксперимента: 50

Неравенство для энтропии:
 $2,0055003267734 < H < 2,6966396833792$

Двоичная таблица угаданных символов:
00100000000000000000000000000000
00000000000001000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Поле ввода символов:
с

Продолжить Другой

Вероятности:
q[1] = 0,56
q[2] = 0,04
q[3] = 0,06
q[4] = 0
q[5] = 0,02
q[6] = 0,04
q[7] = 0
q[8] = 0,02
q[9] = 0,02
q[10] = 0
q[11] = 0,04
q[12] = 0
q[13] = 0
q[14] = 0,02
q[15] = 0
q[16] = 0,02
q[17] = 0
q[18] = 0
q[19] = 0,02
q[20] = 0,02
q[21] = 0,04
q[22] = 0,02
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0,02
q[27] = 0,02
q[28] = 0
q[29] = 0
q[30] = 0,02
q[31] = 0
q[32] = 0

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

N20

Лабораторная работа №1

Произвольная часть текста:
ы_чтоб_вели_себя_другие_люди_для_этого_может_быть_сколько_угодно_пояснений_

Использованные буквы:

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ: у

Символ по счету: 1

Номер эксперимента: 50

Неравенство для энтропии:
 $1,65123384644615 < H < 2,40934912384719$

Двоичная таблица угаданных символов:
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Поле ввода символов:
у

Продолжить Другой

Вероятности:
q[1] = 0,56
q[2] = 0,1
q[3] = 0,04
q[4] = 0,04
q[5] = 0
q[6] = 0,02
q[7] = 0,04
q[8] = 0,06
q[9] = 0,06
q[10] = 0,02
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
q[19] = 0
q[20] = 0
q[21] = 0,02
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0,02
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

Лабораторная работа №1

×

Произвольная часть текста:
в следующую минуту они перечеркнут собственное утверждение заявив что дого

Использованные буквы:

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ: e

Символ по счету: 1

Номер эксперимента: 50

Поле ввода символов:
e

Продолжить Другой

Неравенство для энтропии:
1,43655937717164 < H < 2,18563043834684

Двоичная таблица угаданных символов:
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
00000000100000000000000000000000
.....

Вероятности:

q[1] = 0,54
q[2] = 0,18
q[3] = 0,08
q[4] = 0,04
q[5] = 0
q[6] = 0,04
q[7] = 0,04
q[8] = 0
q[9] = 0,02
q[10] = 0,04
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0
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] = 0,02
q[31] = 0
q[32] = 0

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

$$H_0 = \log_2 32 = 5$$

$$R = 1 - \frac{H}{H_0}$$

$$0,4606606332416 < R^{(10)} < 0,59889993464532$$

$$0,518130175230562 < R^{(20)} < 0,66975323071077$$

$$0,562873912330632 < R^{(30)} < 0,712688124565672$$

Підсумок та висновки

В ході виконання роботи було експериментально досліджено статистичні властивості російської мови. Також ми провели оцінку ентропії та надлишковості за допомогою різних моделей. Отримані результати дозволяють зробити декілька висновків

1. Підтверджено теоретичну залежність ентропій від моделі джерела. Чітко видно ієрархію $H_0 > H_1 > H_2 > H_n$. Врахування зв'язків між сусідніми символами (модель біграм, або n-грам) значно зменшує невизначеність у порівнянні з моделлю, що враховує тільки частоти окремо одного символу.
2. Доведено вплив довжини контексту на точність оцінки ентропій. Оцінка умовної ентропії в експериментах показала, що зі збільшенням кількості символів (n) ентропія значно зменшується. Так середнє значення H_{10} складає близько ~ 2.35 біт, тоді як H_{30} складає ~ 1.81 біт. Це свідчить, що російська мова має залежності, які дозволяють людині ефективно передбачити наступні символи.
3. Встановлено значну надлишковість російської мови. Найбільш точні оцінки, які були отримані, показують, що надлишковість російської мови може сягати 60-70%. Це означає, що більше половини символів у тексті є передбачуваними і визначаються структурними закономірностями мови. Простіші статистичні моделі (H_1 , H_2) недооцінюють реальну надлишковість, показуючи значення в діапазоні 10-20%.
4. Також проаналізовано вплив алфавіту на ентропію. Видалення пробілів з тексту призвело до незначного зростання ентропії (H_1 з та без – 4.43 біт, 4.49 біт). Це пояснюється тим, що пробіл є дуже частим символом, структурно передбачуваним. Його видалення усуває частину передбачуваності з тексту, роблячи його в середньому трішки більш хаотичним для аналізу.