Міністерство освіти і науки України

Національний технічний університет України

"Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

Криптографія

Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

Студенти 3 курсу

Гончаров Д. К. та Сергеев А. А.

Вступ

Метою цієї роботи є експериментальна оцінка ентропії на символ та надлишковості російської мови, використовуючи різні моделі джерела. Це дозволило засвоїти теоретичні поняття ентропії та набути практичних навичок їх розрахунку.

Задача

Визначити надлишковість російської мови в різних моделях джерела. Для цього за допомогою власноруч написаної програми було розраховано частоту входження буквених символів у текст розміром понад 1 МБ, а також ентропію

H1 та **H2** для уніграм і біграм (що перетинаються та не перетинаються) з пробілами та без них. За допомогою програми CoolPinkProgram було додатково оцінено умовну ентропію

Хід роботи

Для виконання завдань був написаний програмний код на Python, який працює з текстом розміром понад 1 МБ.

1. Підготовка тексту

Програма виконує попередню фільтрацію тексту:

- Переводить усі літери у нижній регістр.
- Видаляє зайві символи, залишаючи лише літери та пробіли.
- Замінює послідовності пробілів на один пробіл.
- Замінює букву «ё» на «е», а «ъ» на «ь»

2. Обчислення частот та ентропії

Скрипт CryptoLab1.py розраховує:

- **Частоти символів** (уніграм), **біграм**, які перетинаються та **біграм**, які не перетинаються(з кроком 2).
- **Ентропію H1** (для уніграм) та **H2** (для біграм).
- **Надлишковість R** з попередніх розрахунків
  Розрахунки частот виконуються для двох варіантів тексту: з пробілами та без них. Частоти зберігаються у таблицях Excel.

Частоти символів(з пробілами та без):

Буква	Кількість	Частота
	367707	0.163474
о	208453	0.092673
е	160929	0.071545
а	156356	0.069512
т	119069	0.052935
н	118772	0.052803
и	117991	0.052456
с	102628	0.045626
л	93302	0.04148
р	86049	0.038255
в	83348	0.037055
м	63576	0.028264
к	63092	0.028049
д	60770	0.027017
у	53296	0.023694
п	48848	0.021717
я	38739	0.017222
ь	38376	0.017061
ы	35569	0.015813
г	35352	0.015717
б	32439	0.014422
з	31832	0.014152
ч	28118	0.012501
й	19905	0.008849
ж	19848	0.008824
ш	15031	0.006682
х	12178	0.005414
ю	11213	0.004985
э	7870	0.003499
ц	6883	0.00306
ф	5948	0.002644
щ	5848	0.0026

Буква	Кількість	Частота
о	208453	0.110783
е	160929	0.085526
а	156356	0.083096
т	119069	0.06328
н	118772	0.063122
и	117991	0.062707
с	102628	0.054542
л	93302	0.049586
р	86049	0.045731
в	83348	0.044296
м	63576	0.033788
к	63092	0.033531
д	60770	0.032297
у	53296	0.028324
п	48848	0.02596
я	38739	0.020588
ь	38376	0.020395
ы	35569	0.018903
г	35352	0.018788
б	32439	0.01724
з	31832	0.016917
ч	28118	0.014943
й	19905	0.010579
ж	19848	0.010548
ш	15031	0.007988
х	12178	0.006472
ю	11213	0.005959
э	7870	0.004183
ц	6883	0.003658
ф	5948	0.003161
щ	5848	0.003108

Таблиці: Unigram.xlsx(з пробілами) UnigramNoSpace.xlsx(без пробілів)

Таблиці для біграм виглядають наступним чином:

	а	б	в	г	д	е	ж	з	и	й	к
а	2.67E-06	0.001115	0.006105	0.000978	0.005465	0.000201	0.001446	0.005559	0.000306	0	0.00778
б	0.000724	0.000209	6.22E-06	8.89E-07	4.53E-05	0.001047	3.25E-05	0.000174	0.000602	4.45E-07	1.33E-06
в	0.00254	6.58E-05	3.07E-05	5.78E-06	0.001003	0.001121	2.22E-06	0.000816	0.001767	0	8.71E-05
г	0.000736	2.67E-06	1.16E-05	0	7.11E-06	0.003317	1.02E-05	0.00038	0.000355	0	0
д	0.002197	1.24E-05	0.000247	0.001122	3.47E-05	0.003016	0.000873	0.0006	0.0015	0.000235	0
е	0.001435	0.002247	0.005102	0.000247	0.004979	0.001139	0.003545	0.000381	0.001913	1.78E-06	0.00044
ж	0.001123	8E-06	0	0	0.000104	0.000719	7.11E-06	9.25E-05	0.000384	0	1.87E-05
з	0.00432	1.78E-06	0.00061	0	4E-05	0.001142	0	5.78E-06	0.00186	2.22E-06	2.67E-06
и	0.000108	0.000713	0.003366	0.000524	0.00235	9.87E-05	0.001357	0.000331	0.000535	0	0.002098
й	0.000765	0	0	0	0	0.001711	0	0	0.000943	0	0
к	0.00436	0.000131	0.000103	8.76E-05	0.000202	0.001047	6.36E-05	6.4E-05	0.001816	3.65E-05	5.11E-05
л	0.008757	0.000819	0.000683	0.001656	0.000642	0.005892	6.67E-06	0.000202	0.005537	2.8E-05	0.0007
м	0.003141	4.4E-05	0.000142	8E-06	0.000181	0.004053	2.67E-06	0.000259	0.002843	4.09E-05	1.33E-06
н	0.005103	0.000248	0.000723	0.000225	0.001801	0.007147	0.000793	0.00173	0.003673	0.000392	0.000498
о	4.8E-05	0.002031	0.006123	0.008026	0.003853	0.000204	4.31E-05	0.000566	0.000181	4.27E-05	0.007763
п	0.000722	0	0.000216	0	0.000115	0.000723	0	0	0.000225	4.45E-07	4.45E-07
р	0.003253	0.001055	0.000694	0.001442	0.00132	0.006136	5.78E-06	0.000303	0.000616	2.22E-06	0.002033
с	0.003972	0.000125	0.002251	1.02E-05	0.000299	0.00519	3.02E-05	4.36E-05	0.002775	0.000371	0.000221
т	0.004303	3.56E-06	0.000252	7.56E-06	9.74E-05	0.005276	0	1.47E-05	0.004217	0.000452	0.000503
у	5.38E-05	0.001217	0.000612	0.000498	0.001515	0.000108	0.00024	0.000365	6E-05	0	0.001312
ф	0.000882	0	4.45E-07	0	0	1.51E-05	0	0	4.22E-05	0	4.45E-07
х	0.000682	4.49E-05	3.91E-05	0	3.87E-05	0.000474	0	0	0.001257	4.45E-07	1.78E-06

Повні таблиці: Bigram.xlsx(Перетинаються, з пробілами), BigramNoSpace(Перетинаються, без пробілів), BigramStep2(Не перетинаються, з пробілами), BigramStep2NoSpace(Не перетинаються, без пробілів)

Значення ентропії на надлишковості:

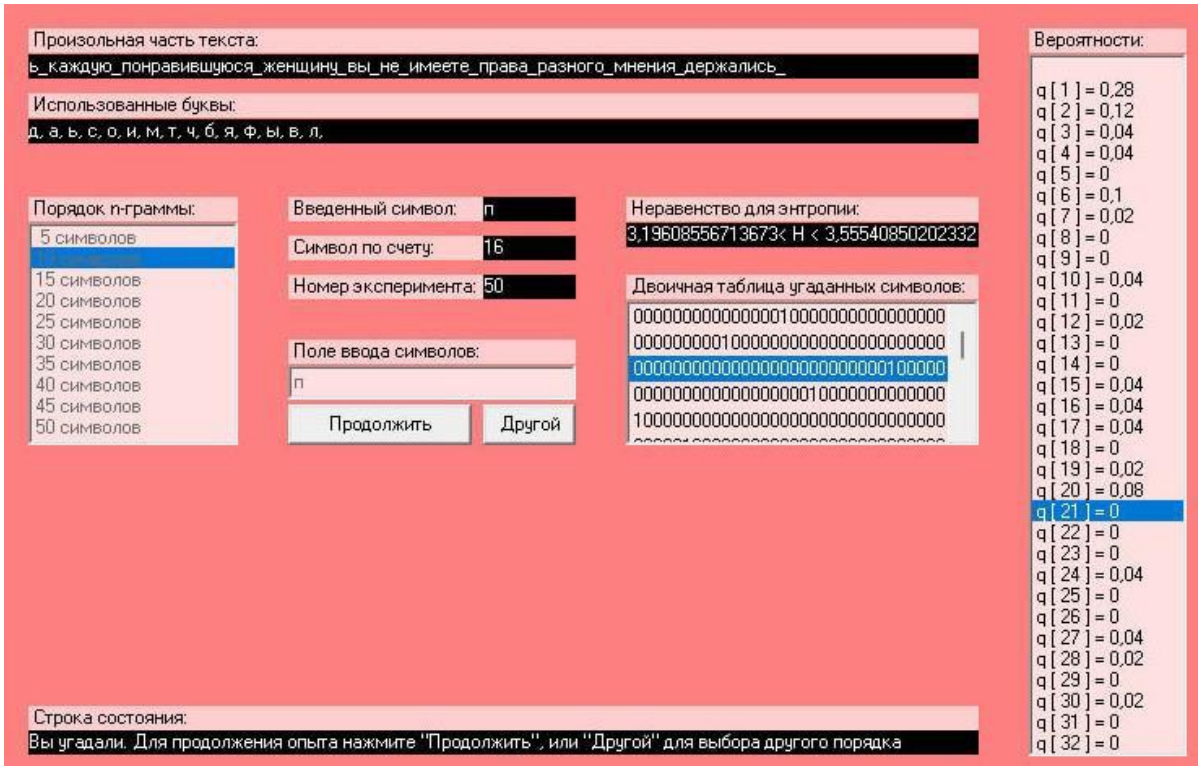
```
--- Значення ентропії ---
H1 (з пробілами): 4.37587
H1 (без пробілів): 4.46287
H2 (з пробілами): 3.96864
H2 (без пробілів): 4.14559
H2 з кроком 2 (з пробілами): 3.96848
H2 з кроком 2 (без пробілів): 4.14520

--- Значення надлишковості ---
R для H1 (з пробілами): 0.12483
R для H1 (без пробілів): 0.09917
R для H2 (з пробілами): 0.20627
R для H2 (без пробілів): 0.16322
R для H2 з кроком 2 (з пробілами): 0.20630
R для H2 з кроком 2 (без пробілів): 0.16329
```

3. Робота з програмою CoolPinkProgram

Звіт також включає результати, отримані за допомогою програми CoolPinkProgram. Це дозволило оцінити умовну ентропію

H(10), H(20) і H(30). Для цього було проведено понад 50 експериментів. Завдяки цій програмі, було продемонстровано, що ентропія зменшується зі збільшенням порядку n-грами.

H10


Произольная часть текста:
ь_каждую_понравившуюся_женщину_вы_не_имеете_права_разного_мнения_держались_
Использованные буквы:
д, а, ь, с, о, и, м, т, ч, б, я, ф, ы, в, л,
Порядок n-граммы:
5 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ: п
Символ по счету: 16
Номер эксперимента: 50
Поле ввода символов:
п
Продолжить
Другой
Неравенство для энтропии:
3,19608556713673< H < 3,55540850202332
Двоичная таблица угаданных символов:
Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка
Вероятности:
q[1] = 0,28
q[2] = 0,12
q[3] = 0,04
q[4] = 0,04
q[5] = 0
q[6] = 0,1
q[7] = 0,02
q[8] = 0
q[9] = 0
q[10] = 0,04
q[11] = 0
q[12] = 0,02
q[13] = 0
q[14] = 0
q[15] = 0,04
q[16] = 0,04
q[17] = 0,04
q[18] = 0
q[19] = 0,02
q[20] = 0,08
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0,04
q[25] = 0
q[26] = 0
q[27] = 0,04
q[28] = 0,02
q[29] = 0
q[30] = 0,02
q[31] = 0
q[32] = 0

H20


Произольная часть текста:
людьми_которые_убегают_с_поля_битвы_или_где_человек_гордится_тем_что_обману
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ: а
Символ по счету: 1
Номер эксперимента: 50
Поле ввода символов:
а
Продолжить
Другой
Неравенство для энтропии:
1,9335652141565< H < 2,77979250081389
Двоичная таблица угаданных символов:
Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка
Вероятности:
q[1] = 0,46
q[2] = 0,12
q[3] = 0,14
q[4] = 0,02
q[5] = 0
q[6] = 0,04
q[7] = 0,02
q[8] = 0,02
q[9] = 0,02
q[10] = 0,04
q[11] = 0,02
q[12] = 0,02
q[13] = 0,02
q[14] = 0
q[15] = 0,02
q[16] = 0
q[17] = 0,02
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] = 0,02
q[30] = 0
q[31] = 0
q[32] = 0

H30

Висновки

Отримані результати підтвердили, що ентропія джерела зменшується зі зростанням порядку **n-грами**.

Зокрема, значення **H2** виявилося меншим за **H1**, що пов'язано з урахуванням залежності між символами в біграмах. Також було встановлено, що присутність пробілів зменшує загальну ентропію тексту, оскільки вони є частиною алфавіту.

Експериментально оцінена надлишковість російської мови знаходиться в діапазоні **0.11–0.22**, що показано у підсумкових розрахунках. Висновки підтверджуються наданими таблицями частот та скріншотами з програми CoolPinkProgram.