

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 III курсу
групи ФБ-95

Пашинський М. О.

Бурчак Б. Ю.

Перевірила:

Селюх П. В.

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.

Завдання:

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 $(10) H$, $(20) H$, $(30) H$.
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

За основу тексту було взято науково-фантастичний роман «Дюна». Програма була розроблена на мові Python, вона фільтрує текст від усього лишнього і підраховує частоти монограм та біграм. Інформація записується в окремі файли. За допомогою отриманих даних було підраховано значення ентропій і надлишковості за даними формулами:

Ентропія(**H**):

$$-0,5 \times \sum_{i=0}^n (x(i) \times \log_2(x(i)))$$

Надлишковість(**R**):

$$1 - \left(\frac{h}{\log_2(31)} \right)$$

Монограми і їх частоти:

Символи з пробілом	Символи без пробілу
а --> 0.06932653754842562	а --> 0.08179723672781458
б --> 0.014045151669829198	б --> 0.016571642499996302
в --> 0.036196250027695664	в --> 0.04270735762775485
г --> 0.015252473693297566	г --> 0.01799614181660163
д --> 0.02611210118763197	д --> 0.030809236950762535
е --> 0.07135448738770714	е --> 0.08418998125165546
ж --> 0.009208338715698657	ж --> 0.010864766775230482
з --> 0.014073578919689257	з --> 0.01660518334263105
и --> 0.058503280212000394	и --> 0.06902705414231873
й --> 0.00925516006840934	й --> 0.01092001051604066
к --> 0.02652680459735518	к --> 0.03129853865508127
л --> 0.043232412415852205	л --> 0.05100920942754167
м --> 0.026674793515744307	м --> 0.031473148335856294
н --> 0.05660994176176209	н --> 0.06679313537330711
о --> 0.09319163185373679	о --> 0.10995526736737433
п --> 0.02546663539669182	п --> 0.030047662523879352
р --> 0.0376268095721239	р --> 0.04439525120858014
с --> 0.04626911157737375	с --> 0.05459216061723043
т --> 0.05319240106167417	т --> 0.06276083554184984
у --> 0.023863004066350873	у --> 0.02815556440113072
ф --> 0.0017980235536487002	ф --> 0.0021214582966480365
х --> 0.007820001998268445	х --> 0.009226691210671512
ц --> 0.003035863065937424	ц --> 0.0035819646943171454
ч --> 0.011567800454083512	ч --> 0.013648656713914863
ш --> 0.006299144130755316	ш --> 0.007432256129712303
щ --> 0.0034004007406134686	щ --> 0.0040120766763392535
ы --> 0.016297593173446773	ы --> 0.019229261031114556
ь --> 0.01554803348228376	ь --> 0.018344867930465886
э --> 0.0032310913848292966	э --> 0.00381231136358816
ю --> 0.005216818397112795	ю --> 0.006155237871162718
я --> 0.017345638988140403	я --> 0.020465832979428118
_ --> 0.15245868538183022	

Перші 10 бірам і їх частоти (за найбільшою частотою появи):

Частоти з кроком 1 без пробілів	Частоти з кроком 2 без пробілів	Частоти з кроком 1 з пробілами	Частоти з кроком 2 з пробілами
то : 0.015386087842834668 ст : 0.013523693845120848 на : 0.012524167953897485 по : 0.011953362694200826 но : 0.011815224861257814 ен : 0.011452613049782408 ни : 0.0106089855700233 ов : 0.010382044844474067 ал : 0.01029225525306111 не : 0.010067287925125347	то : 0.015386300385853898 ст : 0.0135238806610776 на : 0.012524340962429444 по : 0.011953527817643701 но : 0.011815388076468241 ен : 0.011452771255882656 ни : 0.01060913212227538 ов : 0.010382188261772837 ал : 0.010292397430008787 не : 0.01006742699438018	о_ : 0.019159271372424746 _п : 0.01755221369686394 а_ : 0.016984893908691196 е_ : 0.015535866358354631 и_ : 0.015523742354038484 _с : 0.015308854829262664 _в : 0.013682984043556112 _н : 0.013370268346022427 то : 0.012564649162670274 _о : 0.011779933435035613	о_ : 0.019159471622008057 _п : 0.017552397149742216 а_ : 0.016985071432035467 е_ : 0.015536028736698992 и_ : 0.015523904605664728 _с : 0.015309014834919504 _в : 0.013683127055876357 _н : 0.013370408089889143 то : 0.012564780486336522 _о : 0.011780056556980908

	Ентропія:	Надлишковість:
Монограми без пробілу.	2,229394	0,549999
Монограми з пробілом.	2,197483	0,560503
Біграми без перетину та без пробілів.	4,14984	0,162359
Біграми без перетину з пробілами.	4,003918	0,199216
Біграми з перетином та без пробілів.	4,149933	0,16234
Біграми з перетином та пробілами.	4,00399	0,199202

Ентропія для 10 символів:

$$H(10) = 2.57834128284852 \text{ (середнє значення)}$$

Лабораторная работа №1

X

[illegible]

Ентропія для 20 символів:

$$H(20) = 1.736765224107085 \text{ (середнє значення)}$$

Лабораторная работа №1

X

Произвольная часть текста:	
амое_замечательное_	
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:	
Символ по счету:	
Номер эксперимента:	50
Поле ввода символов:	
<button>Продолжить</button>	<button>Другой</button>
Неравенство для энтропии:	
1,39126673932998 < H < 2,08226370888419	
Двоичная таблица угаданных символов:	
1000000000000000000000000000000000 1000000000000000000000000000000000 1000000000000000000000000000000000 0100000000000000000000000000000000 1000000000000000000000000000000000	
Вероятности:	
q[1] = 0,6326530	
q[2] = 0,1224489	
q[3] = 0,0408163	
q[4] = 0,0204081	
q[5] = 0,0204081	
q[6] = 0	
q[7] = 0	
q[8] = 0	
q[9] = 0	
q[10] = 0	
q[11] = 0,040816	
q[12] = 0,020408	
q[13] = 0	
q[14] = 0,020408	
q[15] = 0	
q[16] = 0	
q[17] = 0	
q[18] = 0,020408	
q[19] = 0	
q[20] = 0	
q[21] = 0	
q[22] = 0,020408	
q[23] = 0	
q[24] = 0	
q[25] = 0	
q[26] = 0,020408	
q[27] = 0	
q[28] = 0	
q[29] = 0	
q[30] = 0	
q[31] = 0	
q[32] = 0,020408	
Строка состояния:	

$$H(30) = 2.102131777073025 \text{ (середнє значення)}$$

×

[illegible]

	Ентропія:	Надлишковість:
H(10)	2.57834128284852	0.484332
H(20)	1.736765224107085	0.652647
H(30)	2.102131777073025	0.579574

В ході цієї лабораторної роботи ми засвоїли поняття ентропії та надлишковості, навчилися знаходити частоти біграм та монограм в тексті. А також вдосконалили свої практичні навички щодо мови програмування Python.