

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
групи ФБ-21
КАЮН Вероніка
РУДЮК 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.

Постановка задачі

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 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

Довільний текст **text1**

Спочатку підрахуємо частоту букв та біграм у нашому тексті **text2** – очищений текст з пробілами.

Частота букв	Частота біграм з <i>перетином</i>	Частота біграм <i>без</i> <i>перетину</i>
--------------	--------------------------------------	--

Буква	Частота	Біграма	Частота	Біграма	Частота
	0,15862	од	0,00452	од	0,00467
о	0,09315	дн	0,00194	на	0,00925
а	0,0729	на	0,00914	жд	0,00072
е	0,06827	аж	0,00107	ы	0,00373
и	0,05709	жд	0,00071	ве	0,00569
н	0,05396	ды	0,00049	сн	0,00082
т	0,05069	ы	0,00383	ою	0,00043
л	0,04441	в	0,01626	ч	0,00527
с	0,0426	ве	0,00558	ас	0,00386
р	0,0402	ес	0,00405	н	0,01514
в	0,03935	сн	0,00086	еб	0,0009
к	0,03085	но	0,00971	ыв	0,00086
у	0,02552	ою	0,00041	ал	0,00875
м	0,02528	ю	0,0026	о	0,02075
п	0,02399	ч	0,00515	жа	0,00134
д	0,02362	ча	0,00221	рк	0,00041
г	0,01615	ас	0,00406	ог	0,00441
я	0,01608	с	0,00291	за	0,00564
ь	0,01527	н	0,01519	ка	0,00706
з	0,015	не	0,00883	та	0,00541
ы	0,01455	еб	0,00087	в	0,01613
ч	0,01327	бы	0,00274	м	0,00599
б	0,01302	ыв	0,00085	ос	0,00545
й	0,00978	ва	0,00623	кв	0,00053
ж	0,00756	ал	0,00875	е	0,01589
ш	0,00738	ло	0,00622	п	0,01694
х	0,00685	о	0,02097	ат	0,00492
ю	0,00448	ж	0,00181	ри	0,00559
щ	0,00296	жа	0,00133	ар	0,00443
ц	0,00276	ар	0,00441	ши	0,00177
э	0,00258	рк	0,00041	х	0,00289
ф	0,00181	ко	0,00837	пр	0,00655
с	3E-06	ог	0,00433	уд	0,00189

Всі значення можна переглянути у файлі **frequency_data.xlsx**

Далі обчислюємо ентропію та надлишковість тексту.

```

Ентропія H1 (монограми): 4.37573
Надлишковість R1 (монограми): 0.13256
Ентропія H2 (біграми з перетином): 3.98572
Надлишковість R2 (біграми з перетином): 0.20987
Ентропія H2 (біграми без перетину): 3.98463
Надлишковість R2 (біграми без перетину): 0.21009

```

Текст без пробілів

Видаляємо пробіли із тексту та зберігаємо у **text3**.

Частота букв			Частота біграм з перетином			Частота біграм без перетину		
	Буква	Частота		Біграма	Частота		Біграма	Частота
	о	0,11071		од	0,00622		од	0,00628
	а	0,08664		дн	0,00253		на	0,01073
	е	0,08114		на	0,01096		жд	0,00079
	и	0,06785		аж	0,00147		ыв	0,0014
	н	0,06414		жд	0,00086		ес	0,0066
	т	0,06025		ды	0,00059		но	0,01176
	л	0,05278		ыв	0,00147		юч	0,00022
	с	0,05063		ве	0,00677		ас	0,00676
	р	0,04778		ес	0,00661		не	0,01049
	в	0,04677		сн	0,00142		бы	0,00306
	к	0,03666		но	0,01182		ва	0,00755
	у	0,03034		ою	0,0005		ло	0,00804
	м	0,03004		юч	0,00027		жа	0,0016
	п	0,02851		ча	0,00265		рк	0,00059
	д	0,02807		ас	0,00681		ог	0,00569
	г	0,0192		не	0,01055		оз	0,00249
	я	0,01911		еб	0,002		ак	0,00641
	ь	0,01815		бы	0,00325		ат	0,00696
	з	0,01783		ва	0,00748		ав	0,00507
	ы	0,01729		ал	0,01082		мо	0,00412
	ч	0,01577		ло	0,0079		ск	0,00475
	б	0,01547		ож	0,00232		ве	0,00665
	й	0,01162		жа	0,00158		па	0,002
	ж	0,00898		ар	0,00573		тр	0,00353
	ш	0,00877		рк	0,00059		иа	0,00053
	х	0,00814		ко	0,0102		рш	0,00052
	ю	0,00533		ог	0,00569		их	0,00185
	щ	0,00351		го	0,00913		пр	0,00795
	ц	0,00328		оз	0,0025		уд	0,00256
	э	0,00306		за	0,00658		ах	0,00136
	ф	0,00215		ак	0,0064		по	0,01089
	с	3,6E-06		ка	0,00856		яв	0,00206

Всі значення можна переглянути у файлі **frequency_data_spaces_del.xlsx**

```

Ентропія H1 (монограми): 4.45070
Надлишковість R1 (монограми): 0.10986
Ентропія H2 (біграми з перетином): 4.14648
Надлишковість R2 (біграми з перетином): 0.17070
Ентропія H2 (біграми без перетину): 4.14791
Надлишковість R2 (біграми без перетину): 0.17042

```

За допомогою програми CoolPinkProgram оцінимо значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$

$H^{(10)}$

Лабораторная работа №1

Произвольная часть текста:
_учить_ем

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: 51

Поле ввода символов:
Продолжить Другой

Неравенство для энтропии:
2,49602844092468< H < 3,13622304540502

Двоичная таблица угаданных символов:
00000100000000000000000000000000
10000000000000000000000000000000
00000100000000000000000000000000
00001000000000000000000000000000
00100000000000000000000000000000

Вероятности:
q[1] = 0,38
q[2] = 0,14
q[3] = 0,04
q[4] = 0,06
q[5] = 0,06
q[6] = 0,06
q[7] = 0
q[8] = 0,02
q[9] = 0
q[10] = 0,02
q[11] = 0
q[12] = 0,02
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0,06
q[17] = 0
q[18] = 0
q[19] = 0
q[20] = 0
q[21] = 0,02
q[22] = 0
q[23] = 0,04
q[24] = 0,02
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0,04
q[32] = 0,02

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

Надлишковість при мінімальному значенні ентропії складає: 0.5008

Надлишковість при максимальному значенні ентропії складає: 0.3728

$H^{(20)}$

Лабораторная работа №1

Произвольная часть текста:
_что_случилось_нечто_непредвиденное_освобождающее_его_от_необходимости_выпо

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: o
Символ по счету: 1
Номер эксперимента: 50

Поле ввода символов:
o
Продолжить Другой

Неравенство для энтропии:
1,50123513905468< H < 2,25243127287761

Двоичная таблица угаданных символов:
00000000100000000000000000000000
00010000000000000000000000000000
00010000000000000000000000000000
00000000000000000000000000000010
00000000000000000000100000000000

Вероятности:
q[1] = 0,6
q[2] = 0,1
q[3] = 0,04
q[4] = 0,06
q[5] = 0,02
q[6] = 0,02
q[7] = 0
q[8] = 0
q[9] = 0,04
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] = 0,02
q[20] = 0,02
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0,02
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0,02
q[32] = 0

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

Надлишковість при максимальному значенні ентропії складає: 0.5495

$$H \quad (30)$$

[illegible]

Надлишковість при максимальному значенні ентропії складає: 0.5973

Висновок: під час виконання даної лабораторної роботи ми навчилися експериментально визначати частоти літер і біграм у тексті і на основі цих значень обчислювати ентропію і надлишковість. За допомогою програми CoolPinkProgram ми здійснили серію експериментів, щоб оцінити значення ентропії $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.