



МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра Інформаційної Безпеки

Криптографія

Комп'ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Мета:

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.

Перевірів:

Виконали:

студенти II курсу

групи ФБ-01

Приходько І.Ю.

та Сахній Н.Р.

Київ 2022

ФБ-01 (Приходько Ігор та Сахній Назар)

Постановка задачі:

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 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:

1. Таблиці частот букв і біграм тексту, а також значення H_1 та H_2 та оцінки надлишковості R російської мови у різних моделях відкритого тексту.

1.1. Отримані результати для тексту, у якому присутні пробіли

Таблиця частот букв у тексті			
Номер по частоті використання	Буква	Частота	У відсотках
1	" "	0,15856	15,856%
2	о	0,09311	9,311%
3	а	0,07269	7,269%
4	е	0,06840	6,840%
5	и	0,05757	5,757%
6	н	0,05395	5,395%
7	т	0,05074	5,074%
8	л	0,04421	4,421%
9	с	0,04264	4,264%
10	р	0,03994	3,994%
11	в	0,03963	3,963%
12	к	0,03078	3,078%
13	у	0,02529	2,529%
14	м	0,02525	2,525%

15	п	0,02392	2,392%
16	д	0,02365	2,365%
17	г	0,01612	1,612%
18	я	0,01604	1,604%
19	з	0,01496	1,496%
20	ь	0,01496	1,496%
21	ы	0,01456	1,456%
22	ч	0,01338	1,338%
23	б	0,01303	1,303%
24	й	0,00986	0,986%
25	ж	0,00758	0,758%
26	ш	0,00735	0,735%
27	х	0,00692	0,692%
28	ю	0,00444	0,444%
29	щ	0,00298	0,298%
30	ц	0,00279	0,279%
31	э	0,00259	0,259%
32	ф	0,00182	0,182%
33	ъ	0,00027	0,027%

$H_1 \rightarrow 4.377865166149289$

$R \rightarrow 0.1321326084834017$

Таблиця Топ-10 найчастіших біграм в тексті			
Номер по частоті використання	Буква	Частота	У відсотках
1	"о "	0,02064	2,064%
2	"а "	0,01786	1,786%
3	" п"	0,01703	1,703%
4	"н "	0,01670	1,670%
5	" в"	0,01626	1,626%
6	"е "	0,01606	1,606%
7	" н"	0,01532	1,532%
8	" с"	0,01479	1,479%
9	"то"	0,01340	1,340%
10	"но"	0,00995	0,995%

$H_2 \rightarrow 3.987135744624616$

$R \rightarrow 0.20959075554316509$

Таблиця Топ-10 найчастіших перехресних біграм в тексті			
Номер по частоті використання	Буква	Частота	У відсотках
1	"о "	0,02091	2,091%
2	"а "	0,01788	1,788%
3	" п"	0,01693	1,693%
4	" в"	0,01630	1,630%
5	"н "	0,01620	1,620%
6	"е "	0,01604	1,604%
7	" н"	0,01516	1,516%
8	" с"	0,01475	1,475%
9	"то"	0,01325	1,325%
10	"и"	0,01109	1,109%

$H_2 \rightarrow 3.9875214560036625$

$R \rightarrow 0.20951429217215967$

1.2. Отримані результати для тексту, у якому пробіли відсутні

Таблиця частот букв у тексті			
Номер по частоті	Буква	Частота	У відсотках
1	о	0,10662	10,662%
2	а	0,08638	8,638%
3	е	0,08129	8,129%
4	и	0,06842	6,842%
5	н	0,06412	6,412%
6	т	0,06030	6,030%
7	л	0,05255	5,255%
8	с	0,05068	5,068%
9	р	0,04747	4,747%
10	в	0,04710	4,710%
11	к	0,03658	3,658%
12	у	0,03006	3,006%
13	м	0,03001	3,001%
14	п	0,02843	2,843%

15	д	0,02811	2,811%
16	г	0,01915	1,915%
17	я	0,01907	1,907%
18	з	0,01778	1,778%
19	ы	0,01777	1,777%
20	ъ	0,01730	1,730%
21	ч	0,01591	1,591%
22	б	0,01549	1,549%
23	й	0,01172	1,172%
24	ж	0,00901	0,901%
25	ш	0,00873	0,873%
26	х	0,00823	0,823%
27	ю	0,00528	0,528%
28	щ	0,00354	0,354%
29	ц	0,00332	0,332%
30	э	0,00307	0,307%
31	ф	0,00216	0,216%
32	ь	0,00032	0,032%

$H_1 \rightarrow 4.4531050182728436$

$R \rightarrow 0.10937899634543125$

Таблиця Топ-10 найчастіших біграм в тексті			
Номер по частоті використання	Буква	Частота	У відсотках
1	'то'	0,01604	1,604%
2	'но'	0,01201	1,201%
3	'ст'	0,01120	1,120%
4	'ов'	0,01102	1,102%
5	'на'	0,01088	1,088%
6	'по'	0,01072	1,072%
7	'ал'	0,01055	1,055%
8	'не'	0,01038	1,038%
9	'ко'	0,00994	0,994%
10	'ро'	0,00941	0,941%

$H_2 \rightarrow 4.147634299400495$

$R \rightarrow 0.1704731401199011$

Таблиця Топ-10 найчастіших перехресних біграм в тексті			
Номер по частоті використання	Буква	Частота	У відсотках
1	'то'	0,01623	1,623%
2	'но'	0,01182	1,182%
3	'ст'	0,01118	1,118%
4	'на'	0,01097	1,097%
5	'по'	0,01090	1,090%
6	'ов'	0,01082	1,082%
7	'ал'	0,01069	1,069%
8	'не'	0,01051	1,051%
9	'ко'	0,01025	1,025%
10	'ро'	0,00955	0,955%

$H_2 \rightarrow 4.147794075186039$

$R \rightarrow 0.17044118496279226$

Results			
With spaces			
<i>H1</i>	<i>4,37787</i>	<i>R</i>	<i>0,13213</i>
<i>H2</i>	<i>3,98714</i>	<i>R</i>	<i>0,20959</i>
<i>Cross H2</i>	<i>3,98752</i>	<i>R</i>	<i>0,20951</i>
Without spaces			
<i>H1</i>	<i>4,45311</i>	<i>R</i>	<i>0,10938</i>
<i>H2</i>	<i>4,14763</i>	<i>R</i>	<i>0,17047</i>
<i>Cross H2</i>	<i>4,14779</i>	<i>R</i>	<i>0,17044</i>

2. Оцінки для значень $H^{(10)}$, $(+H^{(15)})$, $H^{(20)}$, $H^{(30)}$, з використанням програми CoolPinkProgram

Лабораторная работа №1

Произвольная часть текста:
е_удается_как_следует_соблюдать_естественный_закон_и_как_только_кто_нибудь_

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: _ (пробел)

Символ по счету: 1

Номер эксперимента: 50

Неравенство для энтропии:
 $1,98937778394808 < H < 2,7074741436677$

Двоичная таблица угаданных символов:

10000000000000000000000000000000	▲
10000000000000000000000000000000	
00100000000000000000000000000000	
10000000000000000000000000000000	
01000000000000000000000000000000	▼

Вероятности:

q[1] = 0,4
q[2] = 0,14
q[3] = 0,12
q[4] = 0,1
q[5] = 0,1
q[6] = 0,04
q[7] = 0
q[8] = 0
q[9] = 0,02
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0,02
q[15] = 0
q[16] = 0,02
q[17] = 0
q[18] = 0,02
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] = 0,02
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0

Поле ввода символов:

Продолжить Другой

Умовна ентропія джерела:
 $1,98937778394808 < H^{(10)} < 2,7074741436677$

Надлишковість джерела відкритого тексту:
 $0,45850517126646 < R^{(10)} < 0,602124443210384$

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

[illegible]

Лабораторная работа №1

Произвольная часть текста:
е_место_я_его_первый_занял_оставьте_его_в_покое_он_не_делает_вам_ничего_пло

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: й

Символ по счету: 1

Номер эксперимента: 50

Неравенство для энтропии:
 $1,12939483347442 < H < 1,78786885835739$

Двоичная таблица угаданных символов:

10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000

Вероятности:

q[1] = 0,62
q[2] = 0,2
q[3] = 0,04
q[4] = 0,04
q[5] = 0,04
q[6] = 0
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0
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] = 0,02
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

Поле ввода символов:
й

Продолжить Другой

Умовна ентропія джерела:
 $1,12939483347442 < H^{(20)} < 1,78786885835739$

Надлишковість джерела відкритого тексту:
 $0,642426228328522 < R^{(20)} < 0,774121033305116$

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

Лабораторная работа №1

Произвольная часть текста:
лений_их_как_ошибаются_скажем_при_сложении_чисел_но_понятие_о_добре_и_зле_н

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: _ (пробел)

Символ по счету: 1

Номер эксперимента: 50

Неравенство для энтропии:
 $1,20779146758356 < H < 1,77335807791325$

Двоичная таблица угаданных символов:

10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Вероятности:

q[1] = 0,7
q[2] = 0,08
q[3] = 0,04
q[4] = 0
q[5] = 0
q[6] = 0
q[7] = 0,04
q[8] = 0
q[9] = 0,02
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0,02
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0,04
q[19] = 0
q[20] = 0
q[21] = 0,02
q[22] = 0,02
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

Поле ввода символов:

Продолжить Другой

Умовна ентропія джерела:
 $1,20779146758356 < H^{(30)} < 1,77335807791325$

Надлишковість джерела відкритого тексту:
 $0,64532838441735 < R^{(30)} < 0,758441706483288$

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

Висновки:

У ході виконання лабораторної роботи було експериментально досліджено, яке значення може приймати ентропія та надлишковість джерела відкритого тексту. Для цього було проведено декілька дослідів, та у результаті яких, знайдено частоту букв та [перехресних] біграм тексту, що дозволило за наявними теоретичними формулами обрахувати кожне значення ентропії та надлишковості. Усі проведені експерименти та обрахунки були необхідними, щоб вивчити та порівняти різні моделі джерела відкритого тексту, а також надлишковість російського письмового тексту згідно наданих даних.