

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

Роботу виконав:
Студент 3 курсу
Групи ФБ-06
Кононець В. М.

Київ – 2022

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

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 $(10) H$, $(20) H$, $(30) H$.

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

Починав я свій шлях виконання роботи з фільтрації тексту від усього зайвого та приведення його до «робочого» формату.

Для цього я зробив 3 функції:

```
def without_punctuation(string, value): # прибираємо усі зайві
знаки
    text_punctuation = ""
    if value == 1:
        text_punctuation = alphabet_with_gap
    elif value == 0:
        text_punctuation = alphabet
    for p in string:
        if p not in text_punctuation:
            # банальна заміна символу у строчці
            string = string.replace(p, '')
    return string

def stripped_lines(text, value): # робимо єдиний текст якщо
текст починається з нової строчки
    with open(text, "r", encoding="utf-8") as file:
        newline_breaks = ""
        for line in file:
            if value == 1: # ставимо пробіл?
                stripped_line = line.strip() + " "
            elif value == 0: # не ставимо пробіл?
                stripped_line = line.strip()
            newline_breaks += stripped_line.lower()
    return newline_breaks

def pretty_text(text, value): # приводимо текст до
потрібного нам за завданням
    if value == 1: # ставимо пробіл?
        newline_breaks = stripped_lines(text, value)
        newline_breaks = without_punctuation(newline_breaks, value)
        return newline_breaks
    elif value == 0: # не ставимо пробіл?
        newline_breaks = stripped_lines(text, value)
```

```
newline_breaks = without_punctuation(newline_breaks, value)
return newline_breaks
```

Перші дві є частинами останньої головної, основні аспекти їх роботи прописані у коментарях. Тепер наш текст відфільтровано.

Другий етап цієї роботи я назвав теоретично-обчислювальний (бо для цього потрібно знати теорію та вміти обчислювати). На цьому етапі створюю функції для обчислення частот вживаності літер та біграм H1, H2, для обчислення ентропії за формулою (1) та надлишковості за формулою (2)

Формула (1)

Ентропія на символ стаціонарного джерела визначається як

$$H_{\infty} = \lim_{n \rightarrow \infty} H_n, \text{ де } H_n = \frac{1}{n} H(x_1, x_2, \dots, x_n),$$

Формула(2)

Надлишковість джерела відкритого тексту (мови) дорівнює $R = 1 - \frac{H_{\infty}}{H_0}$

Коментарі до відповідних функцій надані у файлі з кодом.

Етап третій, та найдовший для мене, бо я мучився з матрицею, але потім вирішив робити по-своєму склеївши символи біграм та помістити їх та їх частоту у два стовпчики таблиці (вважаю його не гіршим, бо коли шукаєш перетин потрібних символів можна не туди глянути, а тут із вмінням робити пошук по екселю можна дуже швидко знайти потрібну біграму). Додавши до цього функцію додавання нотатки у таблицю, помістив біля частот ентропію та надлишковість. Але головне, що я створюю під кожен експеримент окремий лист ексель, а не окремий файл, таким чином усі необхідні дані знаходяться у нас у межах одного файлу. Опис цих функцій надано у коментарях коду.

І саме останній етап - це оформити усе все у два блоки експериментів, а саме: алфавіт із пробілом та без. І ось тут хочу додати, що результати експерименту можна побачити і без таблиці, для цього я закоментував виводи експериментів, тому, розкоментувавши їх, можна побачити усе необхідне.

Перейдемо до самої таблиці.

Частота букв, ентропія та надлишковість (алфавіт з пробілом):

	A	B	C	D	E	F	
1	а	0,071594293					
2	б	0,014468494		Ентропія:	4,38589186		
3	в	0,033431704		Надлишковість:	0,137901937		
4	г	0,014436375					
5	д	0,026431735					
6	е	0,067840193					
7	ё	0,000022672					
8	ж	0,008641798					
9	з	0,014409924					
10	и	0,055805157					
11	й	0,009756503					
12	к	0,029562355					
13	л	0,043656762					
14	м	0,028132132					
15	н	0,057395973					
16	о	0,094489201					
17	п	0,022186409					
18	р	0,035266244					
19	с	0,042623298					
20	т	0,053356585					
21	у	0,022541603					
22	ф	0,001570034					
23	х	0,006646665					
24	ц	0,002930352					
25	ч	0,013580508					
26	ш	0,006100648					
27	щ	0,002656399					
28	ъ	0,00013981					
29	ы	0,016992639					

Частота біграм H1, ентропія та надлишковість (алфавіт з пробілом):

	A	B	C	D	E	F	G
1	аа	0,000026451					
2	аб	0,000903102		Ентропія:	3,984815		
3	ав	0,00238434		Надлишковість:	0,216738		
4	аг	0,00053846					
5	ад	0,002293652					
6	ае	0,001394329					
7	ає	0,000001889					
8	аж	0,001407554					
9	аз	0,00383157					
10	аи	0,000130364					
11	ай	0,001273411					
12	ак	0,004513619					
13	ал	0,009439113					
14	ам	0,002981369					
15	ан	0,00405829					
16	ао	0,000109581					
17	ап	0,000984343					
18	ар	0,002303098					
19	ас	0,003927926					
20	ат	0,004606197					
21	ау	0,000249392					
22	аф	0,000173819					
23	ах	0,000869094					
24	ац	0,000236167					
25	ач	0,00087854					
26	аш	0,000723615					
27	ащ	0,000241835					
28	аъ	0					
29	аы	0					
<div> <div> <div></div> <div></div> </div> <div>experiment_1</div> <div>experiment_H1_1</div> <div>experiment_H2_1</div> <div>experiment_2</div> <div>experiment_3</div> </div>							

Частота біграм H2, ентропія та надлишковість (алфавіт з пробілом):

	A	B	C	D	E	F	
1	аа	0,000037787					
2	аб	0,000891764		Ентропія:	3,985365786		
3	ав	0,002346549		Надлишковість:	0,216629996		
4	аг	0,000578135					
5	ад	0,002297426					
6	ае	0,001526579					
7	аё	0					
8	аж	0,001432113					
9	аз	0,003789997					
10	аи	0,000128474					
11	ай	0,001322531					
12	ак	0,004538172					
13	ал	0,009269056					
14	ам	0,003056937					
15	ан	0,003971373					
16	ао	0,000094467					
17	ап	0,001039132					
18	ар	0,002350327					
19	ас	0,003967594					
20	ат	0,004496607					
21	ау	0,00025317					
22	аф	0,000162482					
23	ах	0,000944665					
24	ац	0,000219162					
25	ач	0,000921993					
26	аш	0,00078974					
27	ащ	0,000249392					
28	аъ	0					
29	аы	0					
<div> <div> <div></div> <div></div> </div> <div> <div>experiment_1</div> <div>experiment_H1_1</div> <div>experiment_H2_1</div> <div>experiment_2</div> <div>experiment_3</div> </div> </div>							

Частота букв, ентропія та надлишковість (алфавіт без пробілу):

	A	B	C	D	E	F	G
1	а	0,085210395					
2	б	0,017220172		Ентропія:	4,465661008		
3	в	0,039789886		Надлишковість:	0,114727973		
4	г	0,017181945					
5	д	0,031458633					
6	е	0,080742325					
7	ё	0,000026984					
8	ж	0,010285331					
9	з	0,017150464					
10	и	0,066418416					
11	й	0,011612036					
12	к	0,035184648					
13	л	0,051959587					
14	м	0,033482419					
15	н	0,06831178					
16	о	0,112459552					
17	п	0,026405913					
18	р	0,041973326					
19	с	0,050729575					
20	т	0,063504163					
21	у	0,02682866					
22	ф	0,001868629					
23	х	0,007910756					
24	ц	0,003487658					
25	ч	0,016163306					
26	ш	0,007260895					
27	щ	0,003161604					
28	ъ	0,0001664					
29	ы	0,02022437					
<div> ◀ ▶ experiment_1 experiment_H1_1 experiment_H2_1 experiment_2 exper </div>							

Частота біграм H1, ентропія та надлишковість (алфавіт без пробілу):

	A	B	C	D	E	F	G	H
1	аа	0,000213622						
2	аб	0,001866385		Ентропія:	4,146777213			
3	ав	0,004650221		Надлишковість:	0,177943453			
4	аг	0,001124328						
5	ад	0,003696791						
6	ае	0,002295878						
7	ає	0,000002249						
8	аж	0,001814666						
9	аз	0,005084212						
10	аи	0,001250253						
11	ай	0,00152234						
12	ак	0,00667851						
13	ал	0,011490634						
14	ам	0,004357896						
15	ан	0,007368847						
16	ао	0,001897866						
17	ап	0,003271795						
18	ар	0,003274044						
19	ас	0,006750467						
20	ат	0,006428909						
21	ау	0,000890468						
22	аф	0,000344044						
23	ах	0,001196285						
24	ац	0,000328304						
25	ач	0,001735963						
26	аш	0,000969171						
27	ащ	0,000290077						
28	аъ	0						
29	аы	0						
<div> <div> <div></div> <div></div> </div> <div>experiment_1</div> <div>experiment_H1_1</div> <div>experiment_H2_1</div> <div>experiment_2</div> <div>experiment_H1_2</div> </div>								

Частота біграм H2, ентропія та надлишковість (алфавіт без пробілу):

Ентропія: $2,50123750 < H(10) < 3,12665092$

The screenshot shows a software interface for a laboratory experiment on entropy. The interface is divided into several sections:

- Top Section:** Contains a text input field with the text "Произвольная часть текста:" and a sample text "м_конечно_не_имелось_в_виду_что_время_от_времени_нам_не_будут_попадаться_ин". Below it is a field for "Использованные буквы:".
- Left Panel:** A list of "Порядок n-граммы:" (Order of n-grams) from 5 to 50. The value 37 is selected and highlighted in blue.
- Center Section:** Includes a "Введенный символ:" (Entered symbol) field with the value "ь", a "Символ по счету:" (Symbol by number) field with the value "1", and a "Номер эксперимента:" (Experiment number) field with the value "52". Below these is a "Поле ввода символов:" (Symbol input field) containing "ь". At the bottom of this section are two buttons: "Продолжить" (Continue) and "Другой" (Other).
- Right Panel:** Displays "Вероятности:" (Probabilities) for 32 symbols, labeled q[1] through q[32]. The values are: q[1] = 0.5384615, q[2] = 0.0961538, q[3] = 0.0961538, q[4] = 0.0576923, q[5] = 0.0192307, q[6] = 0.0384615, q[7] = 0.0576923, q[8] = 0, q[9] = 0, q[10] = 0, q[11] = 0, q[12] = 0, q[13] = 0.019230, q[14] = 0, q[15] = 0, q[16] = 0, q[17] = 0, q[18] = 0, q[19] = 0, q[20] = 0, q[21] = 0.019230, q[22] = 0, q[23] = 0, q[24] = 0, q[25] = 0, q[26] = 0, q[27] = 0.019230, q[28] = 0.019230, q[29] = 0, q[30] = 0.019230, q[31] = 0, q[32] = 0.
- Bottom Section:** A "Строка состояния:" (Status bar) with the text "Вы угадали. Для продолжения опыта нажмите 'Продолжить', или 'Другой' для выбора другого порядка" (You guessed. To continue the experiment, click 'Continue', or 'Other' to choose another order).

Ентропія: $1.67213167 < H(20) < 2.44399753$

[illegible]

Ентропія: $1.68334109 < H(30) < 2.47040885$

	H	R
H(10)	2.81394421	0.43721116
H(20)	2.05806460	0.58838708
H(30)	2.07687497	0.58462501

Висновки

У ході даної лабораторної роботи, я опанував поняття ентропія та надлишковість. Написав програму мовою Python, яка фільтрує текст, рахує частоту букв та біграм, ентропію, надлишковість, а також заносить усі дані до таблиці. Я побачив залежність надлишковості від ентропії на прикладі, можна зробити висновок, що чим більше ентропія, тим менше надлишковість мови. Попрацював з програмою CoolPinkProgram, та виявив ентропію для кожного з дослідів, за отриманими значеннями порахував надлишковість: $H(10) - R = 0.43721116$, $H(20) - R = 0.58838708$, $H(30) - R = 0.58462501$.