

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
1	а	0,070311181			
2	б	0,014209189		Ентропія:	4,353515751
3	в	0,032832542		Надлишковість:	0,144265838
4	г	0,014177646			
5	д	0,025958026			
6	е	0,066624362			
7	ё	0,000022266			
8	ж	0,00848692			
9	з	0,01415167			
10	и	0,054805018			
11	й	0,009581647			
12	к	0,029032539			
13	л	0,042874346			
14	м	0,027627949			
15	н	0,056367324			
16	о	0,092795767			
17	п	0,021788784			
18	р	0,034634203			
19	с	0,041859404			
20	т	0,052400329			
21	у	0,022137613			
22	ф	0,001541896			
23	х	0,006527544			
24	ц	0,002877834			
25	ч	0,013337118			
26	ш	0,005991313			
27	щ	0,002608791			
28	ъ	0,000137305			
29	ы	0,016688097			
30	ь	0,016368957			
31	э	0,006614751			
32	ю	0,004781546			
33	я	0,014994053			
34		0,174852073			
Σ					

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

	A	B	C	D	E
1	аа	0,000009277			
2	аб	0,000877639		Ентропія:	3,967963559
3	ав	0,002308209		Надлишковість:	0,220050606
4	аг	0,0004917			
5	ад	0,002208013			
6	ае	0,00135264			
7	аё	0,000001855			
8	аж	0,001380472			
9	аз	0,003761045			
10	аи	0,000102051			
11	ай	0,001246878			
12	ак	0,004401183			
13	ал	0,009264379			
14	ам	0,002903816			
15	ан	0,00391876			
16	ао	0,000022266			
17	ап	0,000931448			
18	ар	0,002252545			
19	ас	0,003838975			
20	ат	0,004490246			
21	ау	0,000230079			
22	аф	0,000170704			
23	ах	0,000853518			
24	ац	0,000230079			
25	ач	0,000849807			
26	аш	0,000710646			
27	ащ	0,000237501			
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>experimer</div> </div>					
Готово					

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

	A	B	C	D	E	
1	аа	0,000007422				
2	аб	0,000820116		Ентропія:	3,966745804	
3	ав	0,002341599		Надлишковість:	0,22028997	
4	аг	0,000474999				
5	ад	0,002282224				
6	ае	0,001347069				
7	аё	0				
8	аж	0,001406444				
9	аз	0,004048628				
10	аи	0,000103906				
11	ай	0,001235741				
12	ак	0,004323237				
13	ал	0,009088075				
14	ам	0,002894528				
15	ан	0,003774019				
16	ао	0,000025977				
17	ап	0,000894335				
18	ар	0,002174607				
19	ас	0,003941011				
20	ат	0,004490229				
21	ау	0,000222656				
22	аф	0,000178125				
23	ах	0,000872069				
24	ац	0,000189258				
25	ач	0,000797851				
26	аш	0,000738476				
27	ащ	0,000256054				
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>experime</div> </div>						
Готово						

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

	A	B	C	D	E	F
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				
30	ь	0,019837602				
31	э	0,008016442				
32	ю	0,005794775				
33	я	0,018171352				
34						

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

	A	B	C	D	E	F	G	H
1	aa	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						
		experiment_1	experiment_H1_1	experiment_H2_1	experiment_2	experiment_H1_2		

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

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

[illegible]

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

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: $H(10) - R = 0.43721116$, $H(20) - R = 0.58838708$, $H(30) - R = 0.58462501$.