

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

КРИПТОГРАФІЯ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела
відкритого текст

Виконали:
студентки групи ФБ-23
Гуз Вікторія
Шукалович Марія

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

Для початку ми залишили у тексті тільки російські літери(очистили від англійських, цифр, розділових знаків і подвійних пробілів):

```
def clean_text():
    with open('lab1.txt', 'r', encoding='utf-8') as file:
        content = file.read()
    content = content.lower()
    cleaned = re.sub(r'^a-я\s+', '', content)
    cleaned = re.sub(r'\s+', ' ', cleaned)
    cleaned = cleaned.strip()
    return cleaned
```

Також ми врахували, що моментами потрібно буде працювати з текстом без пробілів і очистили від них:

```
def remove_spaces(text):
    return text.replace(" ", "")
```

З цим етапом труднощів не виникало.

Далі слід було обчислити частоту кожної літери, щоб використати ці дані у формулі обчислення ентропії H_1 :

```
def letter_frequencies(letter_counts, total_count):
    frequencies = {}
    for letter, count in letter_counts.items():
        frequencies[letter] = count / total_count
    sorted_frequencies = dict(sorted(frequencies.items(), key=lambda item:
item[1], reverse=True))
    return sorted_frequencies
```

Також відсортували вивід частот у порядку спадання.

H_1 обраховували за наступною формулою:

$$H_1 = - \sum_{i=1}^n p(i) \log_2 p(i)$$

$p(i)$ – частота літери

n – кількість літер в алфавіті

Реалізація формули у коді:

```
def entropy_H1(letter_frequencies):
```

```

entropy_value = 0
for p_i in letter_frequencies.values():
    if p_i > 0:
        entropy_value -= p_i * math.log2(p_i)

return entropy_value

```

Також обчислили частоту біграм(з/без перетин(-y)), щоб використати у формулі обчислення ентропії H2:

```

def bigram_frequencies(bigram_counts, total_bigrams):
    frequencies = {}
    for bigram, count in bigram_counts.items():
        frequencies[bigram] = count / total_bigrams
    return frequencies

```

Формула для обчислення H2:

$$H_2 = -\frac{1}{2} \sum_{i,j} p(i,j) \log_2 p(i,j)$$

$p(i,j)$ – частота біграми

Реалізація формули у коді:

```

def entropy_H2(bigram_frequencies):
    entropy = 0
    for frequency in bigram_frequencies.values():
        if frequency > 0:
            entropy -= frequency * math.log2(frequency)
    return entropy / 2

```

Під час обчислення ентропії проблем не виникло. Усі необхідні кроки виконано правильно, а результати відповідали очікуванням. Формули для обчислення ентропії використовувалися без помилок, дані оброблялися належним чином, і процес обчислення був успішним

Приклад виконання:

<pre> ♥Меню♥ 1. Обробка тексту 2. Аналіз літер 3. Аналіз біграм 4. Ентропія 5. Залис усіх даних у файли 6. Вийти Введіть номер опції: 1 -♥- Обробка тексту -♥- 1. Вивести текст з пробілами 2. Вивести текст без пробілів 3. Повернутись до головного меню Введіть номер опції: 1 Текст з пробілами: леди и бродяга й полнометражный мультфильм студии уолта диснея года п яемый с помощью дочерней компании ранее распространением мультфильмов источниках впервых на рассказе уорда грина счастливый дэн циничный </pre>	<pre> -♥- Обробка тексту -♥- 1. Вивести текст з пробілами 2. Вивести текст без пробілів 3. Повернутись до головного меню Введіть номер опції: 2 Текст без пробілів: ледибродягайполнометражныймультфильмстуди нейкомпанииранеераспространениеммультфильм гринасчастливыйдэнциничныйпсавоторыхнадав </pre>
--	--

♥- Аналіз літер -♥-

1. Вивести кількість літер без пробілів
 2. Вивести частоту літер без пробілів
 3. Вивести кількість літер з пробілами
 4. Вивести частоту літер з пробілами
 5. Повернутись до головного меню
- Введіть номер опції: 2

Частота літер:

Частота літери 'o': 0.11344491868895597
Частота літери 'e': 0.08495437590367301
Частота літери 'a': 0.08020190203099696
Частота літери 'и': 0.06598440258209647
Частота літери 'т': 0.06507487142457921
Частота літери 'н': 0.06407334570341994
Частота літери 'с': 0.05271114922038947
Частота літери 'в': 0.04608580489168074
Частота літери 'л': 0.04588619404604241
Частота літери 'р': 0.04364187384247406
Частота літери 'к': 0.033298560545440975
Частота літери 'д': 0.03173812445649438
Частота літери 'м': 0.03132501679334723
Частота літери 'у': 0.02956670560524612
Частота літери 'я': 0.027624405289860984

♥- Аналіз літер -♥-

1. Вивести кількість літер без пробілів
 2. Вивести частоту літер без пробілів
 3. Вивести кількість літер з пробілами
 4. Вивести частоту літер з пробілами
 5. Повернутись до головного меню
- Введіть номер опції: 4

Частота літер:

Частота символу ' ': 0.16664713906114165
Частота символу 'o': 0.09453964754841761
Частота символу 'e': 0.07079697220860111
Частота символу 'a': 0.06683648451026936
Частота символу 'и': 0.05498829066913148
Частота символу 'т': 0.054230330276901446
Частота символу 'н': 0.053395705951869515
Частота символу 'с': 0.043926987006186635
Частота символу 'в': 0.03840573735515218
Частота символу 'л': 0.03823939108586505
Частота символу 'р': 0.036369080423358485
Частота символу 'к': 0.02774945069568903
Частота символу 'д': 0.026449056816653143
Частота символу 'м': 0.0261047923636937
Частота символу 'у': 0.024639498704668833

♥- Аналіз біграм -♥-

-----3 пробілами-----

1. Вивести кількість біграм з перетином
2. Вивести кількість біграм без перетину
3. Вивести частоту біграм з перетином
4. Вивести частоту біграм без перетину

-----Без пробілів-----

5. Вивести кількість біграм з перетином
6. Вивести кількість біграм без перетину
7. Вивести частоту біграм з перетином
8. Вивести частоту біграм без перетину
9. Повернутись до головного меню

Введіть номер опції: 7

	а	б	в	г	д	е
а	0.00036	0.00139	0.00527	0.00125	0.00326	0.00229
б	0.00082	0.00001	0.00007	0.00000	0.00002	0.00246
в	0.00676	0.00024	0.00048	0.00030	0.00135	0.00635
г	0.00115	0.00003	0.00014	0.00002	0.00129	0.00042
д	0.00619	0.00007	0.00127	0.00002	0.00016	0.00496

♥- Аналіз біграм -♥-

-----3 пробілами-----

1. Вивести кількість біграм з перетином
2. Вивести кількість біграм без перетину
3. Вивести частоту біграм з перетином
4. Вивести частоту біграм без перетину

-----Без пробілів-----

5. Вивести кількість біграм з перетином
6. Вивести кількість біграм без перетину
7. Вивести частоту біграм з перетином
8. Вивести частоту біграм без перетину
9. Повернутись до головного меню

Введіть номер опції: 3

	а	б	в	г	д	е	ё
а	0.00006	0.00057	0.00278	0.00072	0.00184	0.00147	0.00000
б	0.00068	0.00000	0.00003	0.00000	0.00001	0.00204	0.00000
в	0.00556	0.00000	0.00003	0.00002	0.00086	0.00516	0.00000
г	0.00094	0.00000	0.00002	0.00000	0.00102	0.00033	0.00000

♥- Ентропія -♥-

1. H1 з пробілами
 2. H1 без пробілів
 3. H2 з пробілами, з перетинами
 4. H2 без пробілів, з перетинами
 5. H2 з пробілами, без перетинів
 6. H2 без пробілами, без перетинів
 7. Повернутись до головного меню
- Введіть номер опції: 2

Ентропія H1: 4.45793

♥- Ентропія -♥-

1. H1 з пробілами
 2. H1 без пробілів
 3. H2 з пробілами, з перетинами
 4. H2 без пробілів, з перетинами
 5. H2 з пробілами, без перетинів
 6. H2 без пробілами, без перетинів
 7. Повернутись до головного меню
- Введіть номер опції: 3

Ентропія H2: 3.96032

♥Меню♥

1. Обробка тексту
 2. Аналіз літер
 3. Аналіз біграм
 4. Ентропія
 5. Запис усіх даних у файли
 6. Вийти
- Введіть номер опції: 5
Дані успішно записані у файли CSV.

♥Меню♥

1. Обробка тексту
 2. Аналіз літер
 3. Аналіз біграм
 4. Ентропія
 5. Запис усіх даних у файли
 6. Вийти
- Введіть номер опції: 6
/}___/}⚡
(. .)
/ > > Byeee

На наведених скриншотах лише частина даних для прикладу виконання, решта даних у прикріплених файлах:

- bigram_counts_no_overlap_with_spaces.csv – кількість і частота біграм без перетину, з пробілами;
- bigram_counts_no_overlap_without_spaces.csv – кількість і частота біграм без перетину, без пробілів;
- bigram_counts_with_spaces.csv – кількість і частота біграм з перетином, з пробілами;
- bigram_counts_without_spaces.csv – кількість і частота біграм з перетином, без пробілів;
- letter_counts_and_frequencies_with_spaces.csv – кількість і частота літер з пробілами;
- letter_counts_and_frequencies_without_spaces.csv – кількість і частота літер без пробілів;

До файлів не увійшли біграми та літери з частотою 0, тому що неможливо взяти логарифм від 0.

Таблиця з отриманими значеннями ентропії:

	Без пробілів	З пробілами
H1	4.46598	4.37207
H2 з перетинами	4.14440	3.96426
H2 без перетинів	4.14597	3.96331

Таблиці з оцінками надлишковості російської мови:

$$R = 1 - \frac{H_{\infty}}{H_0}$$

Для тексту з пробілами $H_0 = \log_2 34 = 5.087$ (33 літери і пробіл)

Для тексту без пробілів $H_0 = \log_2 33 = 5.044$ (33 літери)

	Без пробілів	З пробілами
H1	0,11459	0,14054
H2 з перетинами	0,17835	0,2207
H2 без перетинів	0,17803	0,22089

Для тексту, який використовує CoolPinkProgram, $H_0 = \log_2 32 = 5$ (32 літери)

	Надлишковість
$2,401809 < H^{(10)} < 3,12449$	$0,51963 > R^{(10)} > 0,375101$
$1,93969 < H^{(20)} < 2,76849$	$0,612062 > R^{(20)} > 0,446302$
$1,18636 < H^{(30)} < 1,783405$	$0,76272 > R^{(30)} > 0,643319$

CoolPinkProgram

Для 10 символов:

Лабораторная работа №1

Произвольная часть текста:
о_то_что_я_не_пытаюсь_показаться_лучше_других_я_просто_стараюсь_обратить_ва

Использованные буквы:
п, а, р, л, м, б, в, г, д, ж, з, к, н, о, с, т, у, ф, х, ц, ч, ш, щ, ю,

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: я

Символ по счету: 25

Номер эксперимента: 53

Поле ввода символов:
я

Продолжить Другой

Неравенство для энтропии:
2.40180988833205< H < 3.12449029246985

Двоичная таблица угаданных символов:
00100000000000000000000000000000
00001000000000000000000000000000
00000000000010000000000000000000
00000010000000000000000000000000
10000000000000000000000000000000

Вероятности:
q[1] = 0.4150943
q[2] = 0.0566037
q[3] = 0.1132075
q[4] = 0
q[5] = 0.0943396
q[6] = 0.0188679
q[7] = 0.0188679
q[8] = 0.0188679
q[9] = 0.0188679
q[10] = 0
q[11] = 0.037735
q[12] = 0.018867
q[13] = 0.018867
q[14] = 0.018867
q[15] = 0.037735
q[16] = 0.037735
q[17] = 0
q[18] = 0.018867
q[19] = 0
q[20] = 0
q[21] = 0.018867
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0.037735
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0

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

Для 20 символов:

Лабораторная работа №1

Произвольная часть текста:
ые_причины_он_делае

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: 53

Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:
1.93969028983975< H < 2.76849870427632

Двоичная таблица угаданных символов:
10000000000000000000000000000000
10000000000000000000000000000000
00000000000100000000000000000000
00000010000000000000000000000000
0000000000000000000001000000000000

Вероятности:
q[1] = 0.4807692
q[2] = 0.0961538
q[3] = 0.0769230
q[4] = 0.0384615
q[5] = 0.0576923
q[6] = 0
q[7] = 0.0384615
q[8] = 0.0384615
q[9] = 0.0192307
q[10] = 0
q[11] = 0.038461
q[12] = 0.038461
q[13] = 0.019230
q[14] = 0
q[15] = 0.038461
q[16] = 0
q[17] = 0
q[18] = 0
q[19] = 0
q[20] = 0.019230
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
q[31] = 0
q[32] = 0

Строка состояния:

Для 30 символів:

Лабораторная работа №1

Произвольная часть текста:
к_же_как_их_не_могут_нарушить

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: 53

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:
 $1,18636016374571 < H < 1,78340506510595$

Двоичная таблица угаданных символов:
10000000000000000000000000000000
00010000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
01000000000000000000000000000000

Вероятности:
q[1] = 0,5961538
q[2] = 0,25
q[3] = 0,0192307
q[4] = 0,0384615
q[5] = 0
q[6] = 0,0192307
q[7] = 0,0192307
q[8] = 0
q[9] = 0
q[10] = 0
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
q[20] = 0
q[21] = 0
q[22] = 0,019230
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0,019230
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0,019230

Строка состояния:

Висновки: у ході виконання лабораторної роботи ми навчались визначати частоту літер та біграм в тексті, а також обраховувати значення його ентропії та надлишковості. Під час роботи з різними текстами ми помітили, що на ентропію суттєво впливають як довжина тексту, так і його структура (наявність повторюваних символів тощо). У текстах з високим рівнем впорядкованості (наприклад, з частими повтореннями тих самих слів чи символів) ентропія була нижчою, що вказує на меншу кількість інформації, необхідної для його передачі. Окрім цього, ми дослідили, що висока надлишковість свідчить про те, що текст містить багато передбачуваних елементів.