

Міністерство освіти і науки України Національний технічний
університет України "Київський політехнічний інститут імені Ігоря
Сікорського"

Фізико-технічний інститут

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

**Експериментальна оцінка ентропії на символ джерела
відкритого тексту**

Виконали: Савченко Є. ФБ-25
Заєць М. ФБ-25

Київ 2023

Порядок виконання роботи

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.

Код

```
import re
from collections import Counter, defaultdict
import math
import pandas as pd

def filter_text(text, no_space=False):
    text = re.sub(r'\s+', ' ', text).lower()
    if no_space:
        return re.sub(r'^a-яА-Я', '', text)

    return re.sub(r'^a-яА-Я ', '', text)

def letter_freq(text):
    freq = Counter(text)
    return freq

def bigrams_freq(text, overlp=False):
    step = 1
    if not overlp:
        step = 2

    bigram = defaultdict(int)
```

```

for i in range(0, len(text) - 1, step):
    bigram[text[i:i+2]] += 1

return bigram

def entropy(values):
    total = sum(values)
    return -sum((count / total) * math.log2(count / total) for count in values)

def calcBigrams(text, overlp = False, space = False):
    bigrams_freq = bigrams_freq(text, overlp)

    ALPHABET = ' абвгдеёжзийклмнопрстуфхцчшщъыьэюя'
    if not space:
        ALPHABET = ALPHABET[1:]
    bmatrix = pd.DataFrame(0.0, index=list(ALPHABET),
        columns=list(ALPHABET))
    values = bigrams_freq.values()
    sum_bigram_overlap_spaces = sum(values)
    alphabet_size = len(values)
    max_entropy = math.log2(alphabet_size) if alphabet_size > 0 else 0
    ev = entropy(values)

    redundancy = max_entropy - ev
    for letter1 in ALPHABET:
        for letter2 in ALPHABET:
            freq = bigrams_freq.get(letter1 + letter2, 0)
            bmatrix.at[letter1, letter2] = freq / sum_bigram_overlap_spaces
    return bmatrix, ev, redundancy

def calcForText(text, spaces = False):
    text_len = len(text)
    letters_freq_items = letter_freq(text)

    letters_freq = defaultdict(float)

```

```

for letter, count in letters_freq_items.items():
    letters_freq[letter] = count / text_len

bigrams_overlap, bigrams_overlap_entropy, r_overlap = calcBigrams(text,
True, spaces)
bigrams_no_overlap, bigrams_no_overlap_entropy, r_overlap_no =
calcBigrams(text, False, spaces)

bigrams_overlap.loc["Entropy"] = [str(bigrams_overlap_entropy)] + [""] *
(len(bigrams_overlap.columns) - 1)
bigrams_overlap.loc["R"] = [str(r_overlap)] + [""] *
(len(bigrams_overlap.columns) - 1)

bigrams_no_overlap.loc["Entropy"] = [str(bigrams_no_overlap_entropy)] + [""]
* (len(bigrams_no_overlap.columns) - 1)
bigrams_no_overlap.loc["R"] = [str(r_overlap_no)] + [""] *
(len(bigrams_no_overlap.columns) - 1)
letters_entropy = entropy(letters_freq_items.values())
alphabet_size = len(letters_freq)
max_entropy = math.log2(alphabet_size) if alphabet_size > 0 else 0

redundancy = max_entropy - letters_entropy

return letters_freq, bigrams_overlap, bigrams_no_overlap, letters_entropy,
redundancy

def main():
    with open('text.txt', "r")
    as file:
        text = file.read()

text_with_spaces = filter_text(text)
text_wo_spaces = filter_text(text, True)

```

```

letters_freq_spaces, bigrams_overlap_spaces, bigrams_no_overlap_spaces,
letters_entropy_spaces, redundancy_spaces = calcForText(text_with_spaces,
True)
letters_freq_wo_spaces, bigrams_overlap_wo_spaces,
bigrams_no_overlap_wo_spaces, letters_entropy_wo_spaces,
redundancy_wo_spaces = calcForText(text_wo_spaces)

letters_stats = "WITH SPACES:\n"
for letter, freq in sorted(letters_freq_spaces.items(), key = lambda x: x[1],
reverse=True):
letters_stats += f"{letter}: {freq}\n"
letters_stats += f"Entropy: {letters_entropy_spaces}\n"
letters_stats += f"R: {redundancy_spaces}\n"

letters_stats += "\n\nWITHOUT SPACES:\n"
for letter, freq in sorted(letters_freq_wo_spaces.items(), key = lambda x: x[1],
reverse=True):
letters_stats += f"{letter}: {freq}\n"
letters_stats += f"Entropy: {letters_entropy_wo_spaces}\n"
letters_stats += f"R: {redundancy_wo_spaces}\n"

bigrams_overlap_spaces.to_csv("bigrams_overlap_spaces.csv")
bigrams_no_overlap_spaces.to_csv("bigrams_no_overlap_spaces.csv")
bigrams_overlap_wo_spaces.to_csv("bigrams_overlap_wo_spaces.csv")
bigrams_no_overlap_wo_spaces.to_csv("bigrams_no_overlap_wo_spaces.csv")
with open('result_letters.txt', 'w') as file:
file.write(letters_stats)

if __name__ == "__main__":
main()

```

Приклад виконання

Без пробілів

	а	б	в	г	д	е	
	0.00745048155551517	0.00345266218426313	0.00502756072445333	0.0129020534254043	0.00321037010115694	0.00817735780483373	0.00
а	0.0139923677993822	0.000181719062329638	0.00102974135320128	0.0023623478102853	0.000545157186988915	0.00193833666484948	0.000
б	6.05730207765461E-05	0.000848022290871646	0	0	0	0	0.00
в	0.00557271791144224	0.00611787509843116	0	0	0	0.000302865103882731	0
г	0.000908595311648192	0.000969168332424738	0	0	0	0.000908595311648192	6.057
д	0.000484584166212369	0.00448240353746441	0	0.00109031437397783	0.000242292083106185	0	0.00
е	0.0172027379005391	0.000121146041553092	0.00115088739475438	0.00102974135320128	0.00399781937125204	0.00321037010115694	0.00
ё	0	0	0	0	0	0	0
ж	6.05730207765461E-05	0.00109031437397783	0	0	0	0.00115088739475438	0.00
з	0.00133260645708401	0.0055121448906657	0	0.000969168332424738	0.000424011145435823	0.0007874492700951	0.00
и	0.013750075716276	0.000181719062329638	0.000424011145435823	0.00230177478950875	0.000666303228542007	0.00133260645708401	0.00
й	0.00769277363862136	0	0	0	0	6.05730207765461E-05	0.00
к	0.0047246956205706	0.00581500999454843	0	0.000121146041553092	0	0	0.00
л	0.00745048155551517	0.00569386395299534	0	0	0.000302865103882731	6.05730207765461E-05	0.00
м	0.00702647041007935	0.00224120176873221	0	0.000121146041553092	0	0	0.00
н	0.00266521291416803	0.00975225634502393	0	0	0	0.00193833666484948	0.0
о	0.0218668605003332	0	0.00363438124659277	0.0062995941607608	0.00381610030892241	0.00369495426736931	0.00
п	6.05730207765461E-05	0.00242292083106185	0	0	0	0	0.00
р	0.000484584166212369	0.00702647041007935	6.05730207765461E-05	0.000302865103882731	6.05730207765461E-05	6.05730207765461E-05	0.0
с	0.00339208916348658	0.0015748985401902	0	0.00121146041553092	0	0.000302865103882731	0.00
т	0.00424011145435823	0.00484584166212369	0.000121146041553092	0.0015748985401902	0	0.000121146041553092	0.00
у	0.00611787509843116	6.05730207765461E-05	0.0007874492700951	0.00109031437397783	0.000726876249318554	0.00175661760251984	6.057
ф	0	0.000181719062329638	0	0	0	0	0.000
х	0.00327094312193349	0.000424011145435823	0	0.000545157186988915	0	0	6.057
ц	0.000363438124659277	0.000424011145435823	0	6.05730207765461E-05	0	0	0.000
ч	6.05730207765461E-05	0.00187776364407293	0	0	0	0	0.00
ш	0.000181719062329638	0.000666303228542007	0	6.05730207765461E-05	0	0	0.00
щ	0	0.000181719062329638	0	0	0	0	0.00
ъ	0	0	0	0	0	0	0
ы	0.00436125749591132	0	0.000302865103882731	0.00121146041553092	6.05730207765461E-05	0	0.00
ь	0.00938881822036465	0	0	0	0	0	0.000
э	0	0	0	0	0	0	0
ю	0.00302865103882731	0	0.000121146041553092	0	0	0.000181719062329638	0.00
я	0.01841419831607	0	0	0.000242292083106185	0.000363438124659277	0.000666303228542007	6.057
Entropy	7.93763965416403						
R	1.19421730644476						

З пробілами

	а	б	в	г	д	е	
	0.0072992700729927	0.00354362903958567	0.00527001241784535	0.0130235939061695	0.00305903019656541	0.00787473119907926	0.00
а	0.0131447436169246	0.000151437138443832	0.00102977254141806	0.00242299421510131	0.000696610836841627	0.00202925765514735	0.000
б	0.000181724566132598	0.000696610836841627	0	0	3.02874276887664E-05	0	0.00
в	0.00517915013477905	0.00602719811006451	0	0	0	0.000181724566132598	0.00
г	0.000696610836841627	0.000969197686040525	0	0	0	0.000787473119907926	0.000
д	0.000787473119907926	0.00517915013477905	3.02874276887664E-05	0.000848047975285459	0.000121149710755066	0	0.0
е	0.017445583487294	0.000121149710755066	0.00118120967986189	0.00118120967986189	0.0039676530272284	0.00293788048581034	0.00
ё	0	0	0	0	0	0	0
ж	9.08622830662992E-05	0.00106005996910682	3.02874276887664E-05	0	0	0.000938910258351758	0.00
з	0.00145379652906079	0.00572432383317685	0.000151437138443832	0.000938910258351758	0.000393736559953963	0.000878335402974225	9.086
и	0.013871641881455	0.000121149710755066	0.00033316170457643	0.00266529363661144	0.000726898264530394	0.00133264681830572	0.00
й	0.00732955750068147	0	0	0	0	6.05748553775328E-05	0.00
к	0.00430081473180483	0.00608777296544205	0	0.000121149710755066	0	0	0.000
л	0.00829875518672199	0.00502771299633522	0	3.02874276887664E-05	0.000151437138443832	3.02874276887664E-05	0.00
м	0.00732955750068147	0.00257443135354514	0	6.05748553775328E-05	3.02874276887664E-05	0	0.00
н	0.00284701820274404	0.00917709058969622	0	0	3.02874276887664E-05	0.00227155707665748	0.00
о	0.0206863131114274	0	0.00330132961807554	0.00593633582699821	0.0037859284610958	0.0039676530272284	0.00
п	6.05748553775328E-05	0.00245328164279008	0	0	0	0	0.00
р	0.000666323409152861	0.00666323409152861	3.02874276887664E-05	0.000484598843020262	6.05748553775328E-05	9.08622830662992E-05	0.00
с	0.00311960505194294	0.00151437138443832	3.02874276887664E-05	0.00112063482448436	3.02874276887664E-05	0.000212011993821365	0.00
т	0.00411909016567223	0.00548202441166672	9.08622830662992E-05	0.00181724566132598	0	9.08622830662992E-05	0.00
у	0.00687524608534997	6.05748553775328E-05	0.000605748553775328	0.00112063482448436	0.000938910258351758	0.00190810794439228	0.000
ф	0	0.000242299421510131	0	0	0	0	0.000
х	0.00284701820274404	0.000393736559953963	0	0.000545173698397795	0	0	6.057
ц	0.00042402398764273	0.000302874276887664	0	0.000121149710755066	0	0	0.000
ч	6.05748553775328E-05	0.00236241935972378	0	0	0	0	0.00
ш	0.000181724566132598	0.000636035981464094	0	3.02874276887664E-05	0	0	0.00
щ	0	0.00033316170457643	0	0	0	0	0.00
ъ	0	0	0	0	0	0	3.028
ы	0.00402822788260593	0	0.000454311415331496	0.00115092225217312	0.000121149710755066	0.000121149710755066	0.00
ь	0.00969197686040525	0	3.02874276887664E-05	0	0	0	0.000
э	3.02874276887664E-05	0	0	0	0	0	0
ю	0.00299845534118787	0	0.000151437138443832	0	0	0.000121149710755066	0.00
я	0.0182027440409486	0	0	0.00042402398764273	0.00033316170457643	0.000726898264530394	9.086
Entropy	7.95070419981045						
R	1.27811449068543						

з пробілами

```
result_letters.txt
1  WITH SPACES:
2  | : 0.16427403234599308
3  o: 0.09373674965170513
4  a: 0.0669029014476952
5  e: 0.06602459264643527
6  н: 0.0554243140105397
7  и: 0.050911623962687017
8  т: 0.04906414682900236
9  с: 0.04639893391483433
10 л: 0.042340541522805746
11 р: 0.03579865527893876
12 в: 0.034042037676418925
13 к: 0.030195650857108243
14 м: 0.027591010963716762
15 у: 0.027166999818280937
16 д: 0.026561269610515477
17 п: 0.025561814767702466
18 я: 0.02519837664304319
19 ь: 0.01699073232782119
20 ы: 0.015688412381125446
21 г: 0.014476951965594525
22 з: 0.014204373372100067
23 б: 0.012720334363074686
24 ч: 0.01235689623841541
25 й: 0.009085953116481919
26 ж: 0.008722514991822642
27 х: 0.007359622024350355
28 ш: 0.007268762493185535
29 ю: 0.004936701193288509
30 щ: 0.0033920891634865833
31 э: 0.002301774789508753
32 ц: 0.0022714882791204797
33 ф: 0.0009085953116481918
34 ь: 0.00012114604155309225
35 Entropy: 4.382564155725319
36 R: 0.6618299636331342
37
```

Без пробілів

```
38
39 WITHOUT SPACES:
40 o: 0.11216206421685873
41 a: 0.08005363484815539
42 e: 0.07900268174240777
43 н: 0.06631876494890193
44 и: 0.06091904037109516
45 т: 0.05870841487279843
46 с: 0.05551931579328839
47 л: 0.05066318764948902
48 р: 0.04283539899978256
49 в: 0.04073349278828731
50 к: 0.03613104298035805
51 м: 0.03301442342538233
52 у: 0.032507066753642094
53 д: 0.031782271508298904
54 п: 0.03058635935348264
55 я: 0.030151482206276725
56 ь: 0.020330506631876494
57 ы: 0.018772196854388634
58 г: 0.017322606363702253
59 з: 0.016996448503297817
60 б: 0.015220700152207
61 ч: 0.014785823005001086
62 й: 0.010871928680147858
63 ж: 0.010437051532941943
64 х: 0.008806262230919765
65 ш: 0.008697542944118286
66 ю: 0.005907081249547003
67 щ: 0.004058853373921867
68 э: 0.002754221932304124
69 ц: 0.0027179821700369644
70 ф: 0.0010871928680147858
71 ъ: 0.0001449590490686381
72 Entropy: 4.472910081385867
73 R: 0.5270899186141333
```

3 Оцінка значень $H(10)$, $H(20)$, $H(30)$ з використанням програми CoolPinkProgram

Результати експерименту для $H(10)$

Лабораторная работа №1

Произвольная часть текста:
й_или_бра

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: 52

Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:
 $3.20798298284626 < H < 3.70645156733443$

Двоичная таблица угаданных символов:

00000000000000010000000000000000
0000000000000000000000000000000100
1000000000000000000000000000000000
000000000000000000000000000000000100
0000010000000000000000000000000000

Вероятности:

q[1] = 0.3137254
q[2] = 0.0588235
q[3] = 0.0588235
q[4] = 0.0392156
q[5] = 0.0392156
q[6] = 0.0392156
q[7] = 0.0392156
q[8] = 0
q[9] = 0.0196078
q[10] = 0.019607
q[11] = 0
q[12] = 0.019607
q[13] = 0
q[14] = 0
q[15] = 0.019607
q[16] = 0.039215
q[17] = 0
q[18] = 0.078431
q[19] = 0
q[20] = 0.058823
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0.039215
q[26] = 0.019607
q[27] = 0
q[28] = 0.019607
q[29] = 0.019607
q[30] = 0.039215
q[31] = 0
q[32] = 0.019607

Строка состояния:

Результаты эксперименту для H(20)

Лабораторная работа №1

Произвольная часть текста:
обые_извинительные_

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: 52

Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:
 $3.2750965504615 < H < 3.57501579470232$

Двоичная таблица угаданных символов:

00000000000000010000000000000000
000000000000000000000000000000000000
1000000000000000000000000000000000
1000000000000000000000000000000000
0100000000000000000000000000000000

Вероятности:

q[1] = 0.2941176
q[2] = 0.0980392
q[3] = 0.0392156
q[4] = 0.0392156
q[5] = 0.0588235
q[6] = 0
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0
q[11] = 0.019607
q[12] = 0.039215
q[13] = 0.019607
q[14] = 0.019607
q[15] = 0.039215
q[16] = 0.078431
q[17] = 0
q[18] = 0.039215
q[19] = 0.039215
q[20] = 0.039215
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0.039215
q[25] = 0
q[26] = 0
q[27] = 0.019607
q[28] = 0.078431
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0

Строка состояния:

Результаты эксперименту для H(30)

Лабораторная работа №1

Произвольная часть текста:
люди_полагали_что_человеческа

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: 52

Неравенство для энтропии:
 $1.37457738123368 < H < 1.97184360774955$

Двоичная таблица угаданных символов:

```

01000000000000000000000000000000
00000000000000000100000000000000
00000000010000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

```

Поле ввода символов:

Продолжить Другой

Вероятности:

q[1] = 0.6470588
q[2] = 0.1176470
q[3] = 0
q[4] = 0.0392156
q[5] = 0.0196078
q[6] = 0.0196078
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0.019607
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0.058823
q[16] = 0.019607
q[17] = 0
q[18] = 0.019607
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] = 0.019607
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0.019607
q[32] = 0

Строка состояния:

Результати експериментів у наведених таблицях:

	Найвище значення H	Найнижче значення H
H10	3,706	3,207
H20	3,575	3,275
H30	1,971	1,374

Результати з першого та другого завдань

	H	R
літери з пробілами	4.38256415572531 9	0.6618299636331342
літери без пробілів	4.47291008138586 7	0.5270899186141333
біграми(перет.) з пробілами	7.95070419981045	1.27811449068543
біграми(не перет.) з пробілами	8.27623382066583	1.15630807972243
біграми(перет.) без пробілів	7.93763965416403	1.19421730644476

біграми(не перет.) без пробілів	8.26283907640888	1.05908901847848
---------------------------------	------------------	------------------

Висновок

У результаті виконання цього практикуму ми ознайомилися з поняттями частоти зустрічі букв і біграм у тексті, а також обрахували ентропію та надлишковість певного текстового джерела (мови). Використовуючи програму CoolPinkProgram, ми провели експерименти, щоб визначити, як кількість букв впливає на ентропію та надлишковість.