КРИПТОГРАФІЯ
КОМП’ЮТЕРНИЙ ПРАКТИКУМ No1

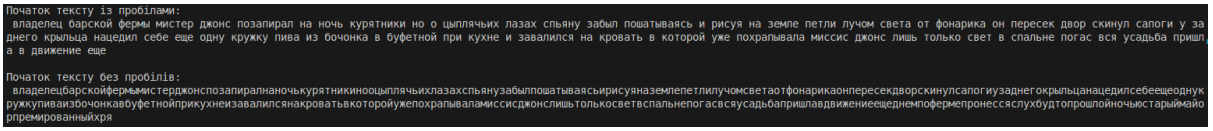
Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав студент ФБ-34 Синельник Максим

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.

Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H 2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

Отже на початку завдання треба відфільтрувати файл з текстом, як сказано у методичних вказівках, я замінив всі символи, окрім текстових із алфавіту на пробіли, а самі послідовності пробілів на один пробіл. В сам же алфавіт тексту із пробілами буде входити і сам пробіл, тоді в алфавіті буде 33 літери. Для другого тексту без пробілів, пробіли прибрав.



Далі відповідно порахував частоти для букв у двух текстах, кількості букв для них будуть однакові, а от частота різна, так як в тексті з пробілами, пробіл виступає в ролі літери, отже кількість літер буде більшою ніж у тексті без пробілів.

```
Частота букв для тексту з пробілами)
 : 111110 (0.15699)
о: 67506 (0.09538)
е: 51717 (0.07307)
и: 45201 (0.06387)
а: 43360 (0.06126)
н: 41168 (0.05817)
т: 33120 (0.04680)
с: 31987 (0.04520)
в: 29444 (0.04160)
л: 29363 (0.04149)
р: 27274 (0.03854)
к: 19272 (0.02723)
д: 18735 (0.02647)
м: 17719 (0.02504)
п: 17238 (0.02436)
у: 15351 (0.02169)
ы: 13362 (0.01888)
ь: 11738 (0.01658)
б: 10735 (0.01517)
я: 10449 (0.01476)
з: 10199 (0.01441)
г: 9998 (0.01413)
ч: 8035 (0.01135)
ж: 7743 (0.01094)
х: 5985 (0.00846)
й: 5572 (0.00787)
ш: 4201 (0.00594)
ю: 2896 (0.00409)
щ: 2129 (0.00301)
ц: 1866 (0.00264)
э: 1599 (0.00226)
ф: 1467 (0.00207)
ъ: 216 (0.00031)
```

```
Частота букв для тексту без пробілів
о: 67506 (0.11314)
е: 51717 (0.08668)
и: 45201 (0.07576)
а: 43360 (0.07267)
н: 41168 (0.06900)
т: 33120 (0.05551)
с: 31987 (0.05361)
в: 29444 (0.04935)
л: 29363 (0.04921)
р: 27274 (0.04571)
к: 19272 (0.03230)
д: 18735 (0.03140)
м: 17719 (0.02970)
п: 17238 (0.02889)
у: 15351 (0.02573)
ы: 13362 (0.02240)
ь: 11738 (0.01967)
б: 10735 (0.01799)
я: 10449 (0.01751)
з: 10199 (0.01709)
г: 9998 (0.01676)
ч: 8035 (0.01347)
ж: 7743 (0.01298)
х: 5985 (0.01003)
й: 5572 (0.00934)
ш: 4201 (0.00704)
ю: 2896 (0.00485)
щ: 2129 (0.00357)
ц: 1866 (0.00313)
э: 1599 (0.00268)
ф: 1467 (0.00246)
ъ: 216 (0.00036)
```

```
Довжина тексту з пробілами: 707755
Довжина тексту без пробілів: 596645
```

Далі треба обчислити біграми. Варіантів біграм буде 4 з перетином і без для тексту з пробілами та без. Опісля можна приступати до обчислення частоти біграм для кожної категорії. Таблиці з наведеними частотами біграм, робляться в кінці, на прикладі я виводив по 30 найчастіших біграм під кожну категорію.

```
30 найчастіших біграм що перетинаються(текст з пробілами)
и : 15848 (0.022392)
е : 13759 (0.019440)
о : 13327 (0.018830)
 п: 11510 (0.016263)
 в: 11490 (0.016234)
 с: 11333 (0.016013)
 н: 11234 (0.015873)
а : 10927 (0.015439)
то: 8547 (0.012076)
 о: 7385 (0.010434)
ст: 7234 (0.010221)
ни: 7225 (0.010208)
по: 7198 (0.010170)
на: 7194 (0.010165)
ь : 6995 (0.009883)
ли: 6860 (0.009693)
 и: 6847 (0.009674)
не: 6512 (0.009201)
но: 6469 (0.009140)
 к: 6207 (0.008770)
ер: 6031 (0.008521)
ко: 5951 (0.008408)
во: 5873 (0.008298)
я : 5765 (0.008145)
ал: 5665 (0.008004)
м : 5491 (0.007758)
ен: 5477 (0.007739)
ра: 5300 (0.007488)
ов: 5203 (0.007351)
от: 5192 (0.007336)
```

```
30 найчастіших біграм що неперетинаються(текст з пробілами)
и : 7927 (0.022400)
е : 6871 (0.019416)
о : 6654 (0.018803)
 п: 5775 (0.016319)
 в: 5719 (0.016161)
 с: 5664 (0.016006)
 н: 5621 (0.015884)
а : 5471 (0.015460)
то: 4266 (0.012055)
 о: 3678 (0.010393)
ст: 3634 (0.010269)
ни: 3631 (0.010261)
на: 3610 (0.010201)
по: 3594 (0.010156)
ь : 3482 (0.009840)
ли: 3439 (0.009718)
 и: 3418 (0.009659)
но: 3239 (0.009153)
не: 3237 (0.009147)
 к: 3108 (0.008783)
ер: 3025 (0.008548)
ко: 2991 (0.008452)
во: 2930 (0.008280)
я : 2904 (0.008206)
ал: 2817 (0.007960)
м : 2749 (0.007768)
ен: 2735 (0.007729)
ра: 2625 (0.007418)
от: 2602 (0.007353)
ов: 2597 (0.007339)
```

```
30 найчастіших біграм що перетинаються(текст без пробілів)
то: 8713 (0.014603)
ни: 7407 (0.012414)
ст: 7364 (0.012342)
на: 7219 (0.012099)
по: 7198 (0.012064)
ли: 7109 (0.011915)
ен: 6704 (0.011236)
но: 6682 (0.011199)
ов: 6673 (0.011184)
не: 6547 (0.010973)
ер: 6452 (0.010814)
он: 6267 (0.010504)
во: 6215 (0.010417)
ко: 6184 (0.010365)
ос: 6157 (0.010319)
от: 5835 (0.009780)
ал: 5801 (0.009723)
ра: 5310 (0.008900)
ро: 5110 (0.008565)
ол: 4854 (0.008136)
пр: 4606 (0.007720)
ка: 4512 (0.007562)
ло: 4445 (0.007450)
ив: 4341 (0.007276)
ре: 4337 (0.007269)
ом: 4260 (0.007140)
ва: 4198 (0.007036)
ин: 4196 (0.007033)
ве: 4159 (0.006971)
ор: 4104 (0.006878)
```

```
30 біграм що неперетинаються(текст без пробілів)
то: 4353 (0.014592)
ни: 3721 (0.012473)
ст: 3689 (0.012366)
на: 3612 (0.012108)
по: 3557 (0.011923)
ли: 3553 (0.011910)
ен: 3356 (0.011250)
но: 3339 (0.011193)
ов: 3309 (0.011092)
не: 3259 (0.010924)
ер: 3208 (0.010753)
он: 3140 (0.010526)
во: 3107 (0.010415)
ко: 3089 (0.010355)
ос: 3085 (0.010341)
от: 2926 (0.009808)
ал: 2891 (0.009691)
ра: 2624 (0.008796)
ро: 2590 (0.008682)
ол: 2439 (0.008176)
пр: 2291 (0.007680)
ка: 2252 (0.007549)
ло: 2224 (0.007455)
ив: 2188 (0.007334)
ре: 2172 (0.007281)
ом: 2138 (0.007167)
ве: 2101 (0.007043)
ва: 2100 (0.007039)
ин: 2086 (0.006992)
ор: 2051 (0.006875)
```

Далі йде обчислення ентропії. Для букв буду використовувати звичайну формулу ентропії, так як це найпростіша модель джерела повідомлень - це H1. H2 це ентропія,

яка враховує залежність між сусідніми символами, тобто ми дивимося вже не на окремі літери, а на пари, відповідно будем ділити все на 2.

```
H1 (з пробілами): 4.37887
H1 (без пробілів): 4.450491
H2 (з пробілами, перетинаються): 3.964017
H2 (з пробілами, неперетинаються): 3.963849
H2 (без пробілів, перетинаються): 4.121706
H2 (без пробілів, неперетинаються): 4.121597
```

Код для виконання цього завдання:

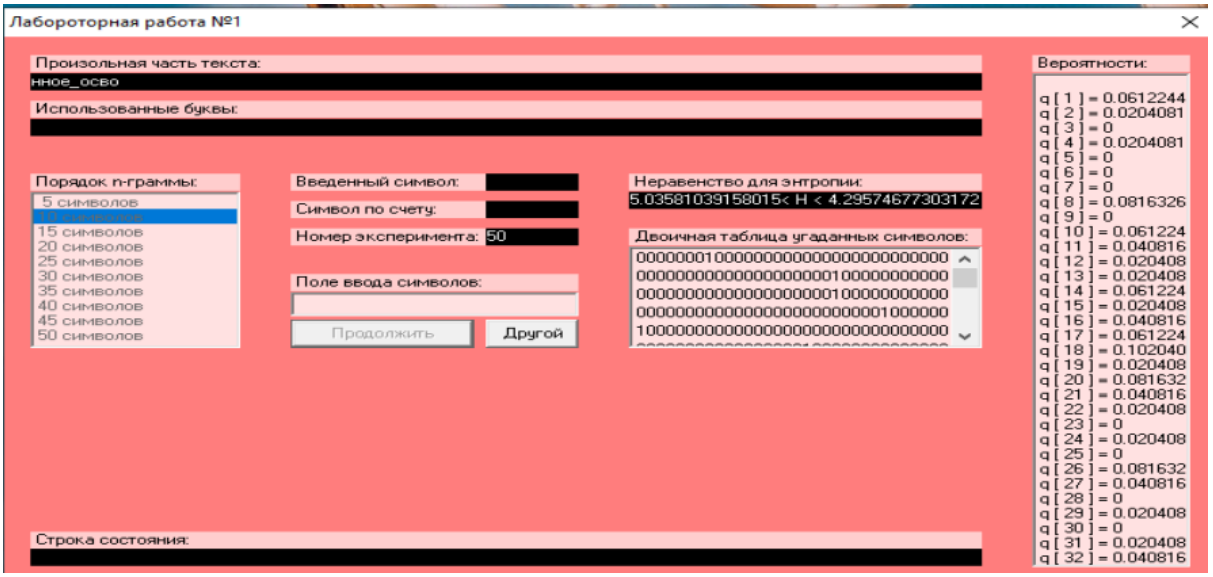
```
import re
import math
from collections import Counter
with open("crypto_lab1.txt", "r", encoding="utf-8") as f:
    text = f.read()
text = text.lower()

text_with_spaces = re.sub(r"[^а-яёі\s]", " ", text)
text_with_spaces = re.sub(r"\s+", " ", text_with_spaces).strip()

text_no_spaces = re.sub(r"\s+", "", text_with_spaces)

print("Початок тексту із пробілами:\n", text_with_spaces[:400])
print("\nПочаток тексту без пробілів:\n", text_no_spaces[:400])
print("\nДовжина тексту з пробілами:", len(text_with_spaces))
print("Довжина тексту без пробілів:", len(text_no_spaces))

def chastota_bukv(text):
    counts = Counter(text)
    total = sum(counts.values())
    return {ch: counts[ch] / total for ch in counts}, counts, total

letter_freq_with, letter_counts_with, total_with =
chastota_bukv(text_with_spaces)
letter_freq_no, letter_counts_no, total_no = chastota_bukv(text_no_spaces)

print("\nЧастота букв для тексту з пробілами)")
for ch, cnt in letter_counts_with.most_common(33):
    print(f"{ch}: {cnt} ({letter_freq_with[ch]:.5f})")

print("\nЧастота букв для тексту без пробілів")
for ch, cnt in letter_counts_no.most_common(32):
    print(f"{ch}: {cnt} ({letter_freq_no[ch]:.5f})")


def bigrams_count_func(text, step=1):
    bigrams = Counter()
    for i in range(0, len(text) - 1, step):
        pair = text[i:i+2]
        if len(pair) == 2:
            bigrams[pair] += 1
```

```
    return bigrams


bigrams_with_overlap = bigrams_count_func(text_with_spaces, step=1)
bigrams_with_nonoverlap = bigrams_count_func(text_with_spaces, step=2)
bigrams_no_overlap = bigrams_count_func(text_no_spaces, step=1)
bigrams_no_nonoverlap = bigrams_count_func(text_no_spaces, step=2)

def bigram_chastota(counter):
    total = sum(counter.values())
    return {bg: counter[bg] / total for bg in counter}, total

bigrams_freq_with_overlap, total_with_overlap =
bigram_chastota(bigrams_with_overlap)
bigrams_freq_no_overlap, total_no_overlap =
bigram_chastota(bigrams_no_overlap)

print("\n30 найчастіших біграм що перетинаються(текст з пробілами)")
for bg, cnt in bigrams_with_overlap.most_common(30):
    print(f"{bg}: {cnt} ({bigrams_freq_with_overlap[bg]:.6f})")

print("\n30 найчастіших біграм що неперетинаються(текст з пробілами)")
for bg, cnt in bigrams_with_nonoverlap.most_common(30):
    freq = cnt / sum(bigrams_with_nonoverlap.values())
    print(f"{bg}: {cnt} ({freq:.6f})")

print("\n30 найчастіших біграм що перетинаються(текст без пробілів)")
for bg, cnt in bigrams_no_overlap.most_common(30):
    print(f"{bg}: {cnt} ({bigrams_freq_no_overlap[bg]:.6f})")

print("\n30 біграм що неперетинаються(текст без пробілів)")
for bg, cnt in bigrams_no_nonoverlap.most_common(30):
    freq = cnt / sum(bigrams_no_nonoverlap.values())
    print(f"{bg}: {cnt} ({freq:.6f})")

def entropy_H1(text):
    counts = Counter(text)
    total = sum(counts.values())
    H = 0.0
    for c in counts:
        p = counts[c] / total
        H -= p * math.log2(p)
    return H

def entropy_H2(counter):
    total = sum(counter.values())
    H = 0.0
    for count in counter.values():
        p = count / total
        H -= p * math.log2(p)
    return H / 2

H1_with = entropy_H1(text_with_spaces)
```

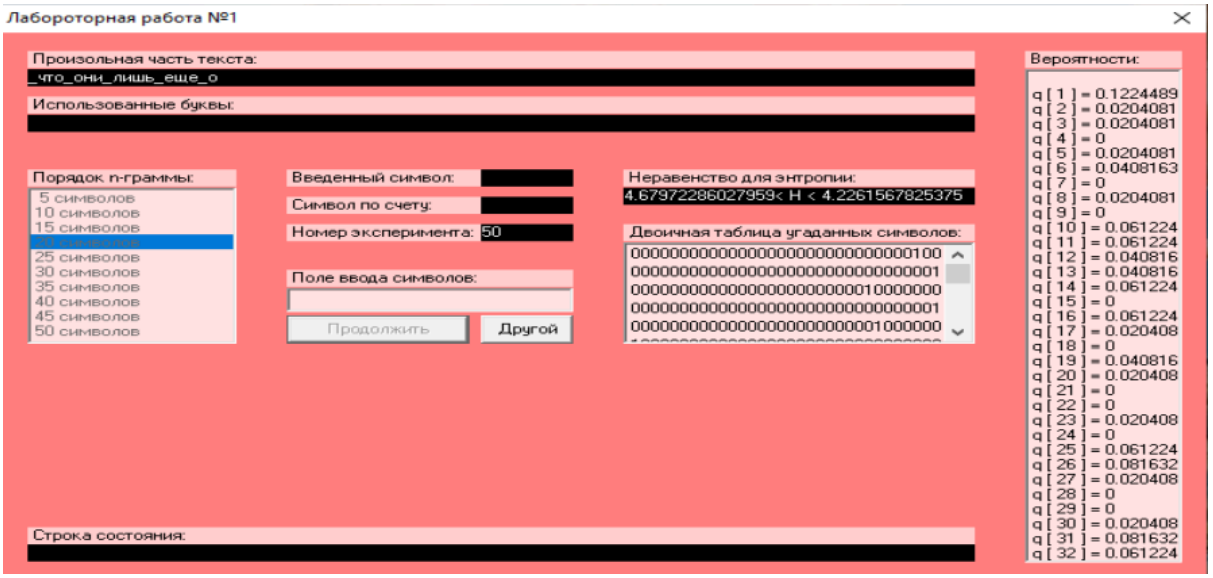
```
H1_no = entropy_H1(text_no_spaces)

H2_with_overlap = entropy_H2(bigrams_with_overlap)
H2_with_nonoverlap = entropy_H2(bigrams_with_nonoverlap)
H2_no_overlap = entropy_H2(bigrams_no_overlap)
H2_no_nonoverlap = entropy_H2(bigrams_no_nonoverlap)

print("\nH1 (з пробілами):", round(H1_with, 6))
print("H1 (без пробілів):", round(H1_no, 6))
print("H2 (з пробілами, перетинаються):", round(H2_with_overlap, 6))
print("H2 (з пробілами, неперетинаються):", round(H2_with_nonoverlap, 6))
print("H2 (без пробілів, перетинаються):", round(H2_no_overlap, 6))
print("H2 (без пробілів, неперетинаються):", round(H2_no_nonoverlap, 6))

import csv

def save_bigrams(counter, total, filename):
    with open(filename, "w", newline="", encoding="utf-8") as f:
        writer = csv.writer(f)
        writer.writerow(["Біграма", "Кількість", "Частота"])
        for bg, cnt in counter.most_common():
            freq = cnt / total
            writer.writerow([bg, cnt, f"{freq:.8f}"])

save_bigrams(bigrams_with_overlap, total_with_overlap,
"bigrams_with_overlap.csv")
save_bigrams(bigrams_with_nonoverlap, sum(bigrams_with_nonoverlap.values()),
"bigrams_with_nonoverlap.csv")
save_bigrams(bigrams_no_overlap, total_no_overlap, "bigrams_no_overlap.csv")
save_bigrams(bigrams_no_nonoverlap, sum(bigrams_no_nonoverlap.values()),
"bigrams_no_nonoverlap.csv")
```

За допомогою програми CoolPinkProgram оцінити значення H (10) , H (20) , H (30)

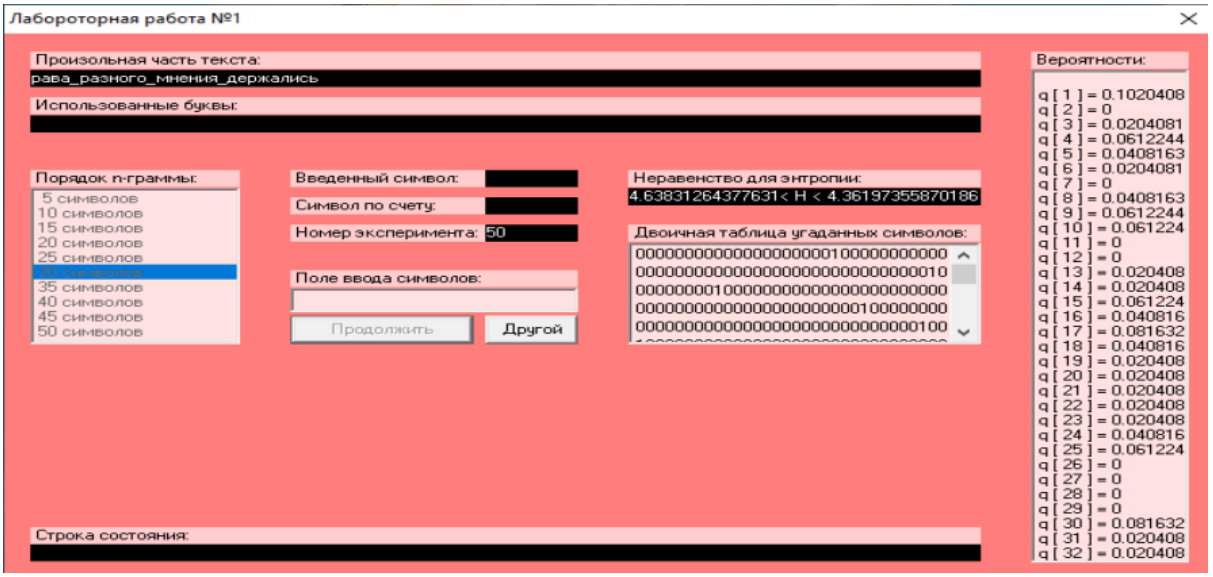
$5.03<H(10)<4.29$



4.68<H(20)<4,23



4.63<H(30)<4.36



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

Для цього використаю формулу:

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

```
H1 (з пробілами): 4.37887
H1 (без пробілів): 4.450491
H2 (з пробілами, перетинаються): 3.964017
H2 (з пробілами, неперетинаються): 3.963849
H2 (без пробілів, перетинаються): 4.121706
H2 (без пробілів, неперетинаються): 4.121597
```

Відповідно будемо ділити ентропію нашої моделі на максимально можливу ентропію (log2(N));

Для алфавіту без пробілів це буде 33 літери, тоді H0 буде log2(33)=5.0444

Для алфавіту з пробілами це буде 34 літери, тоді H0 буде log2(34)=5.0875

R1(з пробілом) : 1-(4.3789/5.0875)=0.139

R1(без пробілу) : 1-(4.4504/5.0444)=0.118

R2 (з пробілами, перетинаються): 1-(3.9640/5.0875)=0.214

R2 (з пробілами, неперетинаються): 1-(3.9638/5.0875)=0.221

R2 (без пробілів, перетинаються): 1-(4.1217/5.0444)=0.183

R2 (без пробілів, неперетинаються): 1-(4.1215/5.0444)=0.183

Висновки: Отримані результати показали, що символи та їх сполучення мають нерівномірний розподіл і виявляють закономірності. Зменшення ентропії при переході від окремих символів до біграм свідчить про наявність залежностей між елементами тексту. Врахування пробілів призводить до зменшення ентропії, оскільки пробіл є одним із найчастіших і найбільш передбачуваних символів. Підрахована надлишковість підтверджує структурованість і передбачуваність текстових повідомлень, що є характерною властивістю природних інформаційних джерел.