# КРИПТОГРАФІЯ КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

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

#### 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.

#### 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 Н1 та Н2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення Н1 та Н2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 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

#### 1. Розрахунок частот та ентропії для літер

Маємо два випадки: з пробілом та без. Далі наведені таблиці з отриманими значеннями:

## 3 пробілом

 $H_1 = 4.354029047, R = 0.129194190592$ 

## Без пробілу

 $H_1 = 4.44876864953,\, R = 0.12919419$ 

Літера	Частота
	0,16821
П	0,02283
p	0,0348
e	0,07245
c	0,04403
T	0,05386
y	0,02467
Л	0,03823
Н	0,05414
И	0,05394
a	0,06627
к	0,02747
3	0,01281
Γ	0,01405
Ь	0,01932
Ы	0,01374
й	0,00833
o	0,09543
M	0,02615
В	0,03848
ж	0,00949
Ю	0,00467
б	0,01447
Д	0,02663
Ч	0,01506
Э	0,00293
ц	0,00231
Я	0,01777
X	0,00708
ш	0,00685
ф	0,00104
Щ	0,00249

Літера	Частота
П	0,02744
p	0,04184
e	0,0871
c	0,05294
T	0,06476
y	0,02965
Л	0,04597
Н	0,06509
И	0,06485
a	0,07967
К	0,03303
3	0,0154
Γ	0,01689
Ь	0,02323
Ы	0,01651
й	0,01001
o	0,11473
M	0,03144
В	0,04627
Ж	0,01141
Ю	0,00562
б	0,01739
Д	0,03202
Ч	0,0181
Э	0,00353
Ц	0,00277
R	0,02137
X	0,00851
Ш	0,00823
ф	0,00124
Щ	0,00299

#### 2. Розрахунок частот та ентропій для біграм

Загалом маємо 4 випадки, а саме: біграми з перетином з пробілом, біграми з перетином без пробілу, біграми без перетину з пробілом, біграми без перетину без пробілу. Далі будуть наведені невеликі частини таблиць з частотами біграми, бо тих біграм багато і не хочеться засоряти і так кривий протокол. Усі частоти можна буде переглянути у сѕу файлах, як мають бути у цій же папці)

#### 3 перетином біграм:

#### 3 пробілом

 $H_2 = 3.9432864, R = 0.1749084$ 

#### Без пробілу

 $H_2 = 4.126525$ , R = 0.1508674

Біграма	Частота
пр	0,00666
pe	0,00508
ec	0,00432
ст	0,0094
ту	0,00163
уп	0,00065
пл	0,00062
ле	0,00368
ен	0,00647
ни	0,00752
ие	0,00166
e	0,019
И	0,01165
И	0,01748
Н	0,01592
на	0,01002
ак	0,00552
ка	0,00685
аз	0,00338
за	0,00448

Біграма	Частота
пр	0,00801
pe	0,00611
ec	0,00751
ст	0,01164
ту	0,00216
уп	0,0015
пл	0,00075
ле	0,00482
ен	0,00962
ни	0,00936
ие	0,00245
еи	0,00163
ИН	0,00576
на	0,0121
ак	0,00768
ка	0,00829
аз	0,00457
за	0,0054
ан	0,00576
ег	0,00437

#### Без перетину біграм:

#### 3 пробілом

 $H_2 = 3.94394$ , R = 0.1703768

Біграма	Частота
пр	0,00667
ec	0,00431
ту	0,00169
пл	0,00061
ен	0,00644
ие	0,0017
И	0,01162
Н	0,0157
ак	0,00551
аз	0,00342
ан	0,00311
Γ	0,00302
иа	8,09E-05
ль	0,00456
ны	0,00247
й	0,00593
po	0,007
ма	0,00292
Н	0,00422
гл	0,00125

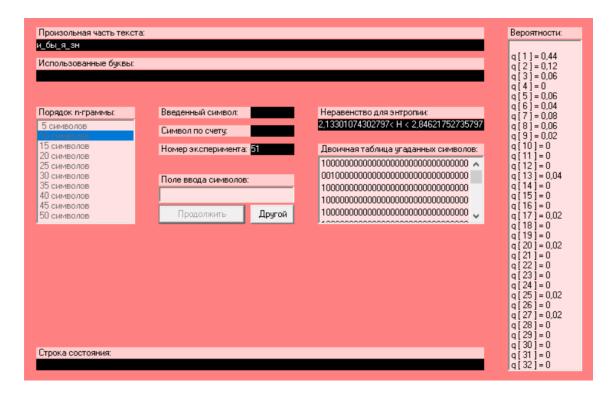
#### Без пробілу

 $H_2 = 4.126635$ , R = 0.1465657

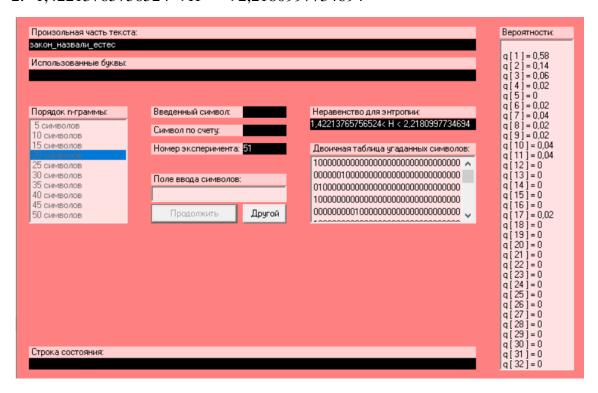
Біграма	Частота
пр	0,00805
ec	0,00762
ту	0,00215
пл	0,00079
ен	0,00969
ие	0,00244
ИН	0,00592
ак	0,00771
аз	0,00453
ан	0,00593
ге	0,00026
ни	0,00925
ал	0,00867
ьн	0,00421
ый	0,00155
po	0,00877
ма	0,00354
НΓ	9,25E-05
ла	0,00685
ВН	0,00243

# 3. Оцінка значення $\mathbf{H}^{(10)}, \mathbf{H}^{(20)}, \mathbf{H}^{(30)}$ за допомогою програми CoolPinkProgram

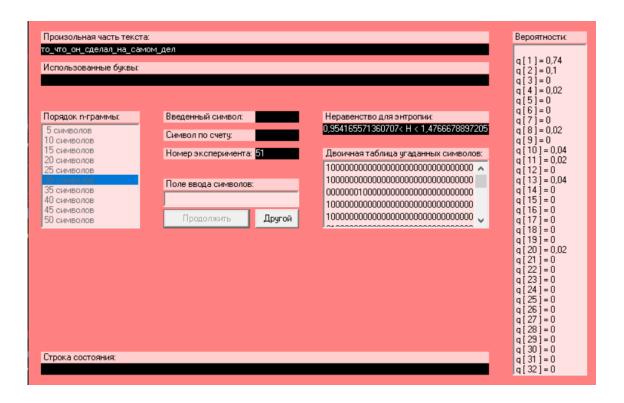
1.  $2,13301074302797 < H^{(10)} < 2,84621752735797$ 



2.  $1,42213765756524 < H^{(20)} < 2,2180997734694$ 



3.  $0.954165571360707 < H^{(30)} < 1.4766678897205$ 



Оцінимо надлишковість російської мови, використовуючи отримані дані:

- 1. Для  $H^{(10)}$  надлишковість 0.430756494529 < R < 0.573397851394
- 2. Для  $H^{(20)}$  надлишковість 0.556380045306 < R < 0.715572468487
- 3. Для  $H^{(30)}$  надлишковість 0.704666422056 < R < 0.809166885728

#### Труднощі

Під час виконання даної лабораторної роботи ми зіткнулися з наступними труднощами:

- 1. На етапі розробки коду для виконання підрахунків ентропії для літер та біграм виникла проблема підрахунку для випадків, коли якась літера (в нас це була літера "ё"), та деякі біграми не були наявними у тексті, тому під час підрахунку за формулою log<sub>2</sub>0 отримували помилку виконання коду. Після виявлення причини помилки, ми прийшли до висновку: необхідно виконати об'єднання літер "ё" з "е", та "ъ" з "ь", а також трохи змінити код так, що перед виконанням обрахунків перевірялося умова не рівності частоти нулю.
- 2. Після виконання програми та аналізу отриманих результатів, довелося визначитися з оформленням даних для біграм, оскільки досить складно надати дані для всіх біграм (враховуючи їх загальну кількість), тому ми вирішили надати як приклад 20 довільно вибраних біграм й прикріпити сsv-файл з повним списком отриманих результатів.

#### Висновок

В результаті виконання даної лабораторної роботи ми ознайомилися з поняттям ентропії, отримали навички її підрахунку для подальшого розгляду надлишковості вибраної мови. Також провели експерименти з наданою програмою CoolPinkProgram, де спробували вгадати зашифровану літеру по попереднім п-грамам. Загалом поставлене завдання допомогло нам розібратися у значенні криптографії для захисту інформації, розвити навички роботи у команді та автоматизації роботи з даними за допомогою програмування.