НАЦIОНАЛЬНИЙ ТЕХНIЧНИЙ УНIВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ ПОЛIТЕХНIЧНИЙ IНСТИТУТ IМЕНI СIКОРСЬКОГО» ФIЗИКО-ТЕХНIЧНИЙ IНСТИТУТ

**Комп’ютерний практикум 1 Експериментальна оцiнка ентропiї на символ джерела вiдкритого тексту**

Виконали:

Толмачов Євгеній

Котович Анна

Перевiрив: Деркач О.Г.

Київ 2020

**Мета роботи**

Засвоєння понть ентропiї на символ джерела та його надлишковостi, вивчення та порiвняння рiзних моделей джерела вiдкритого тексту для наближеного визначення ентропiї, набуття практичних навичок щодо оцiнки ентропiї на символ дерела

**Постановка задачi**

1. Написати програми для пiдрахунку частот букв i частот бiграм в текстi, а також пiдрахунку H1 та H2 за безпосереднiм означенням. Пiдрахувати частоти букв та бiграм, а також значення H1 та H2 на довiльно обраному текстi росiйською мовою достатньої довжини (щонайменше 1Мб), де iмовiрностi замiнити вiдповiдними частотами. Також одержати значення H1 та H2 на тому ж текстi, в якому вилучено всi пробiли.

2. За допомогою програми CoolPinkProgram оцiнити значення H(10), H(20), H(30) .

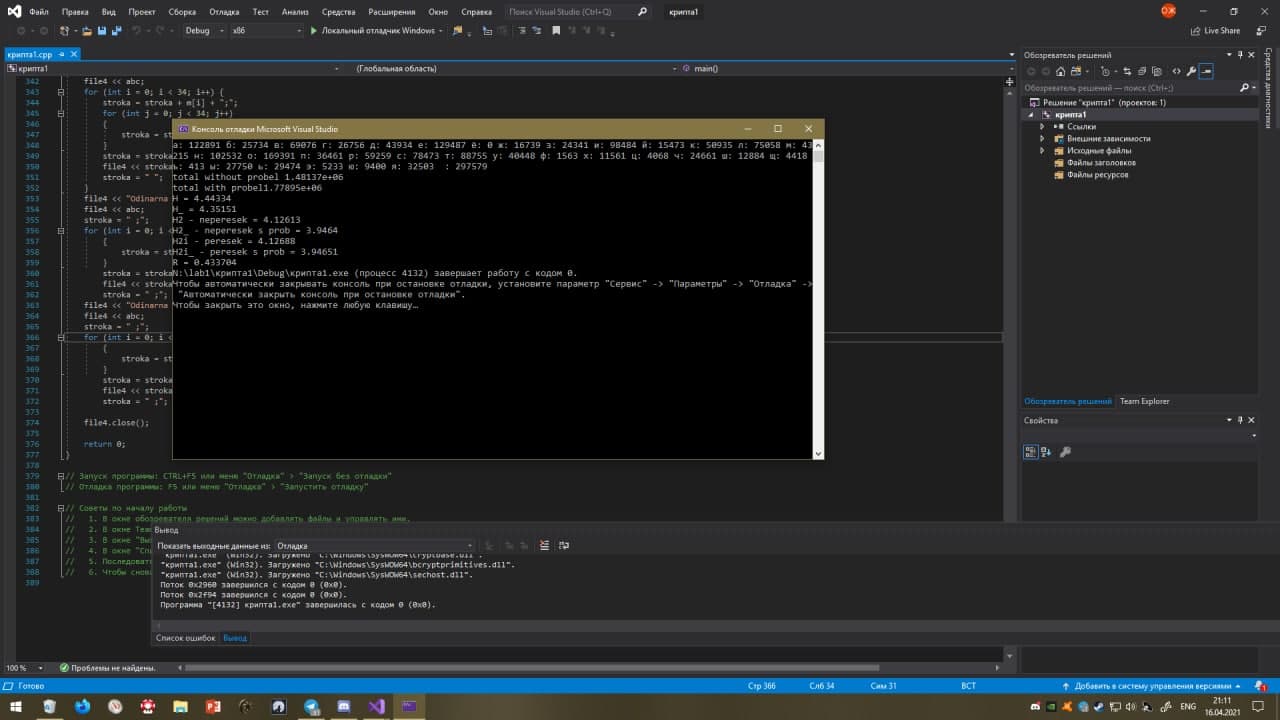
3. Використовуючи отриманi значення ентропiї, оцiнити надлишковiсть росiйської мови в рiзних моделях джерела.

**Хід роботи**

Під час виконання лабораторної роботи особливих проблем не виникло. Єдиними труднощами були ознайомлення з модулями для роботи з файлами (що не було чимось важким саме по собі), уникання повторених поспіль пробілів та підрахунок біграмм що не перетинаються

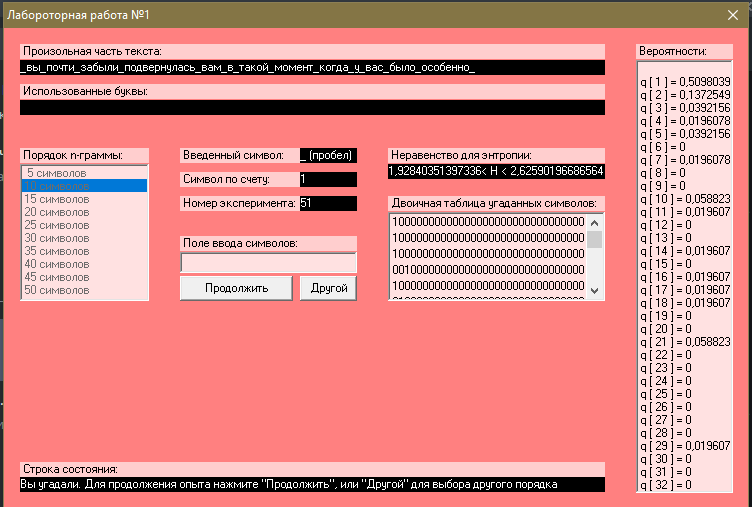
**Результати**

**1.** Код та таблиці біграмм наведені у додатку до листа та за посиланням

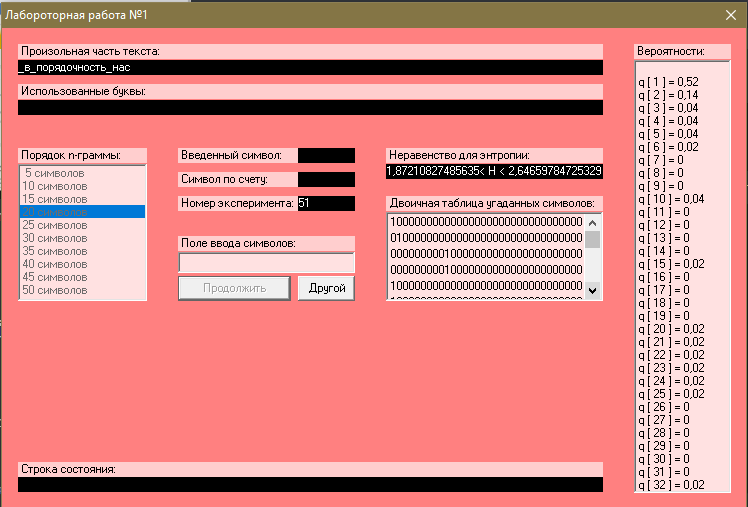


2. **Результати роботи CoolPinkProgram**

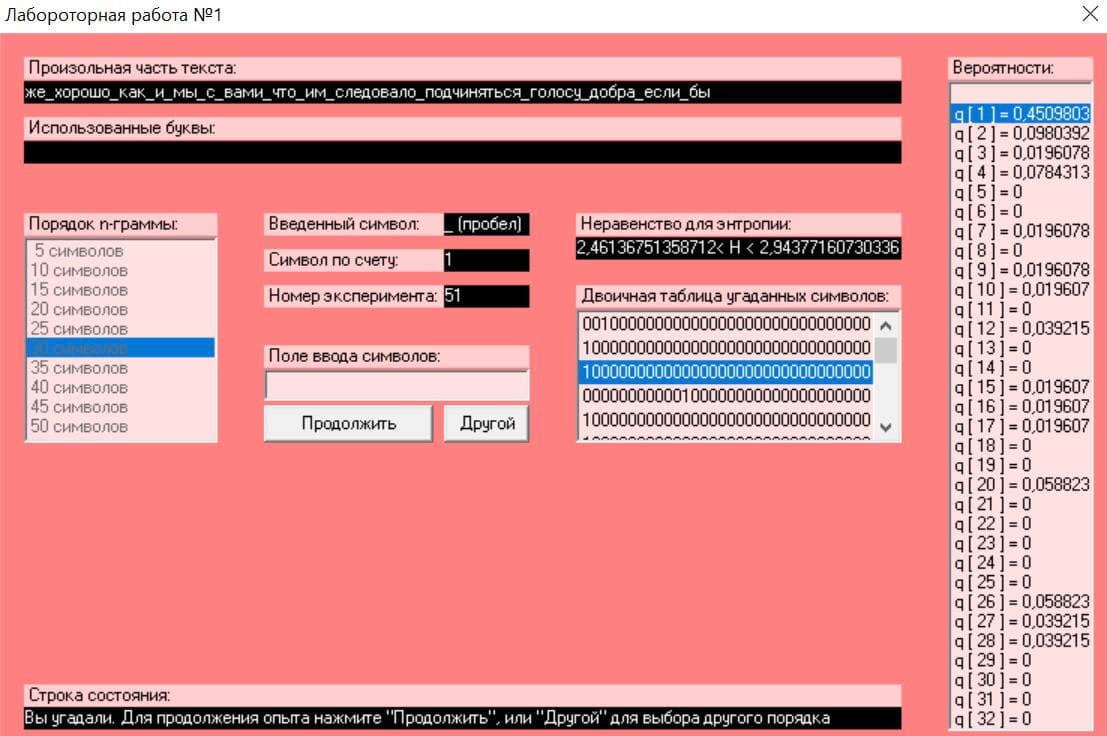
Оцiнка для H(10) : 1.96 < H(10) < 2.81 ⇒ H(10) ≈ 2.387



Оцiнка для H(20) : 1.14 < H(20) < 1.88 ⇒ H(20) ≈ 1.53



Оцiнка для H(30) : 0.787 < H(30) < 1.34 ⇒ H(30) ≈ 1.1

****

**Значення ентропiй**

Ентропiя монограм для тексту без пробiлiв: 4.46423

Ентропiя монограм для тексту з пробiлами: 4.37022

Ентропiя бiграм для тексту без пробiлiв, бiграми перетинаються: 4.13289

Ентропiя бiграм для тексту без пробiлiв, бiграми не перетинаються: 4.13271

Ентропiя бiграм для тексту з пробiлами, бiграми перетинаються: 3.95195

Ентропiя бiграм для тексту з пробiлами, бiграми не перетинаються: 3.95166.

**Оцiнки надлишковостi**

1. H1 (монограми, нерiвноiмовiрнi)

1.1. З пробiлами R = 1 − H1/ H0 = 1 − 4.37/ log 34 = 0.13

1.1. Без пробiлiв R = 1 − H1/ H0 = 1 − 4.46 /log 33 = 0.12

2. H2 (бiграми, нерiвноiмовiрнi)

2.1. З перетином

а) З пробiлами R = 1 − H2 /H0 = 1 − 3.95195/ 5.087 = 0.22

б) Без пробiлiв R = 1 − H2 /H0 = 1 − 4.13289/ 5.044 = 0.18

2.1. Без перетинiв

а) З пробiлами R = 1 − H2 /H0 = 1 − 3.9516 /5.087 = 0.22

б) Без пробiлiв R = 1 − H2 H0 = 1 − 4.13271/ 5.044 = 0.18

3. Надлишковiсть росiйської мови: R = 1 − H∞/ H0 = 1 − 1.659 5 = 0.66

**Висновок**

В ході цієї лабораторної було пiдраховно частоти букв та бiграм, а також оцiнено вiдповiднi значення ентропiй та надлишковость джерела. За допомогою програми CoolPinkProgram практичним шляхом оцiнили значення ентропiй H(10), H(20), та H(30)