Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут імені Ігоря Сікорського”

Фізико-технічний інститут

Криптологія

Комп’ютерний практикум

Робота №1

Виконали:

студенти групи ФБ-14

Мартиненко Даніїл

Цуканов Данило

Київ-2023

**Мета: оцінити надлишковість російської мови в різних моделях джерела**

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 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 ймовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

Код на python:

import math

no\_target\_symb='\n '

file\_name = "res1.txt"

def counf\_symb():

with open(file\_name, "r", encoding="cp1251") as f:

count = 0

while True:

c = f.read(1).lower()

if not c:

break

if c not in no\_target\_symb:

count += 1

return count

def count\_c1():

c1 = dict()

with open(file\_name, "r", encoding="cp1251") as f:

while True:

c = f.read(1).lower()

if not c:

break

if c not in no\_target\_symb:

if c not in c1.keys():

c1[c] = 0

c1[c] += 1

return c1

def count\_c2():

c2 = dict()

with open(file\_name, "r", encoding="cp1251") as f:

a = f.read(1).lower()

while True:

b = f.read(1).lower()

if not b:

break

if b not in no\_target\_symb:

if a+b not in c2.keys():

c2[a+b] = 0

c2[a+b] += 1

a = b

return c2

def count\_h(p):

res = 0

for i in p.keys():

res += p[i] \* math.log2(p[i])

return -res

count\_of\_symbols = counf\_symb()

c1 = count\_c1()

c2 = count\_c2()

p1 = dict()

p2 = dict()

for i in c1.keys():

p1[i] = c1[i] / count\_of\_symbols

for i in c2.keys():

p2[i] = c2[i] / count\_of\_symbols

h1 = count\_h(p1) / 1

h2 = count\_h(p2) / 2

print("\*\*\*\*\*\*\*\*\*\*\*\*p1\*\*\*\*\*\*\*\*\*\*\*\*")

print(p1)

print("\*\*\*\*\*\*\*\*\*\*\*\*h1\*\*\*\*\*\*\*\*\*\*\*\*")

print(h1)

print("\*\*\*\*\*\*\*\*\*\*\*\*p2\*\*\*\*\*\*\*\*\*\*\*\*")

for i in p2:

print(f"{i}: {p2[i]}", end=',')

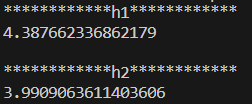
print()

print("\*\*\*\*\*\*\*\*\*\*\*\*h2\*\*\*\*\*\*\*\*\*\*\*\*")

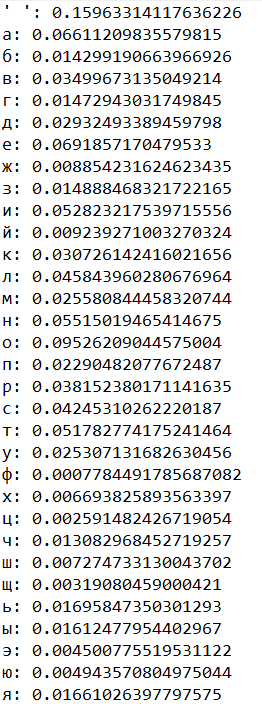
print(h2)

Із пробілами:

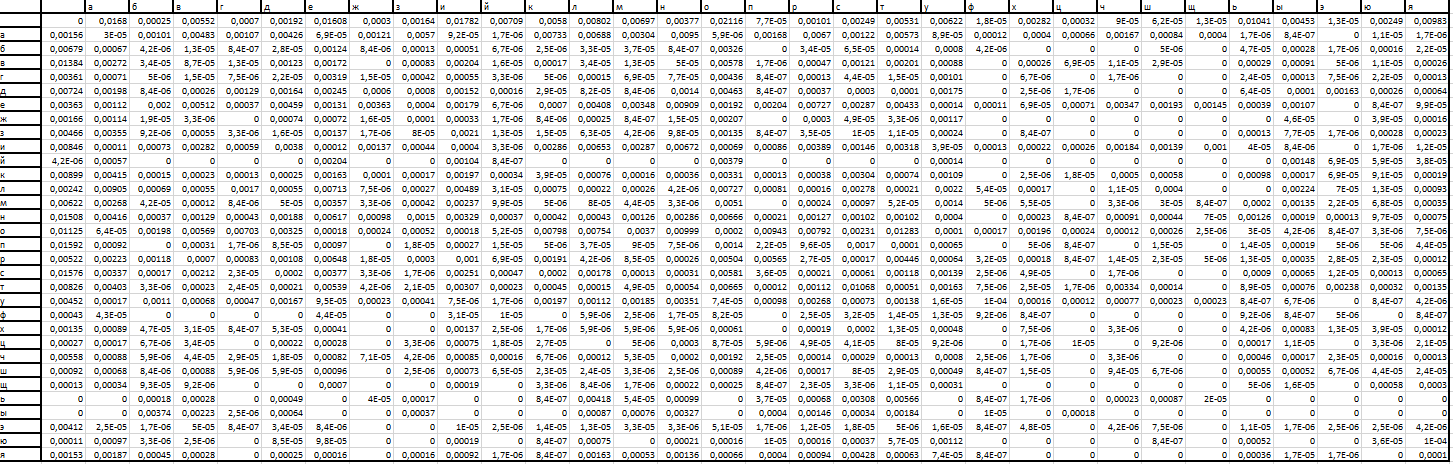
H1 i H2



Частоти букв:

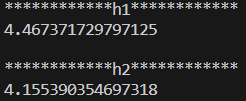


Частоти біграм:

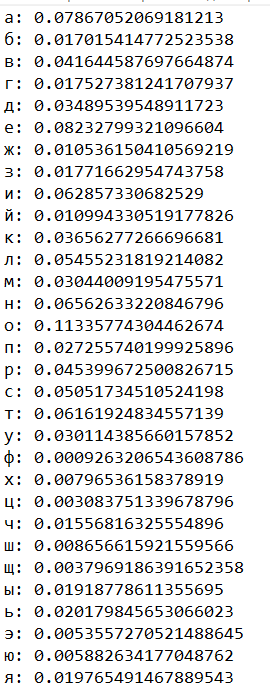


Без пробілів:

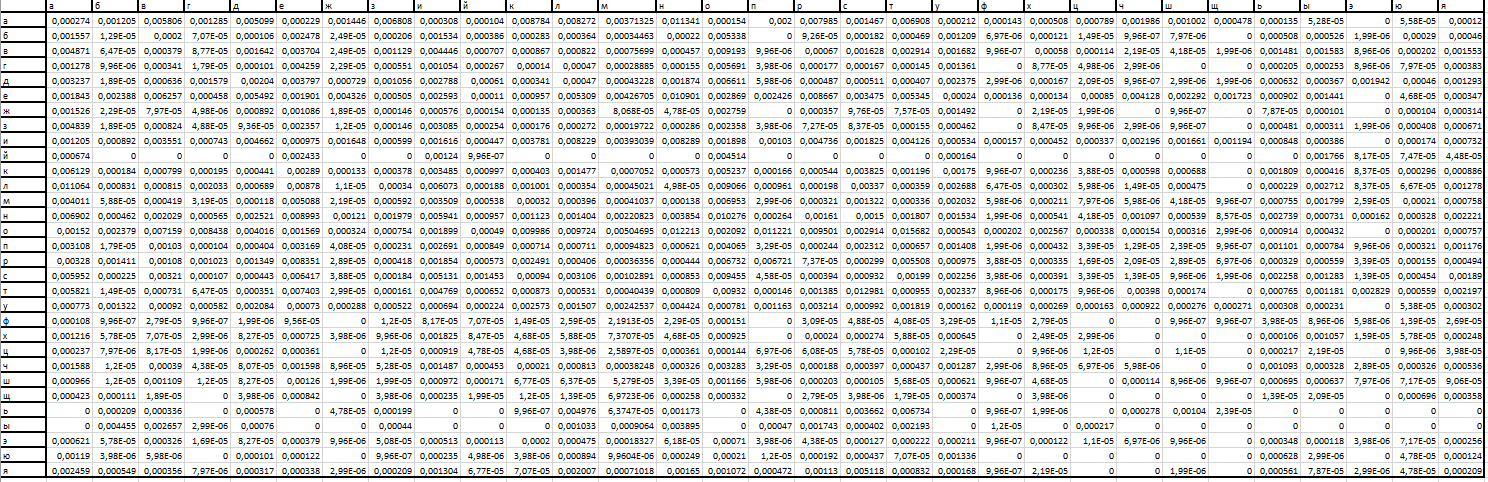
H1 i H2:



Частоти букв:

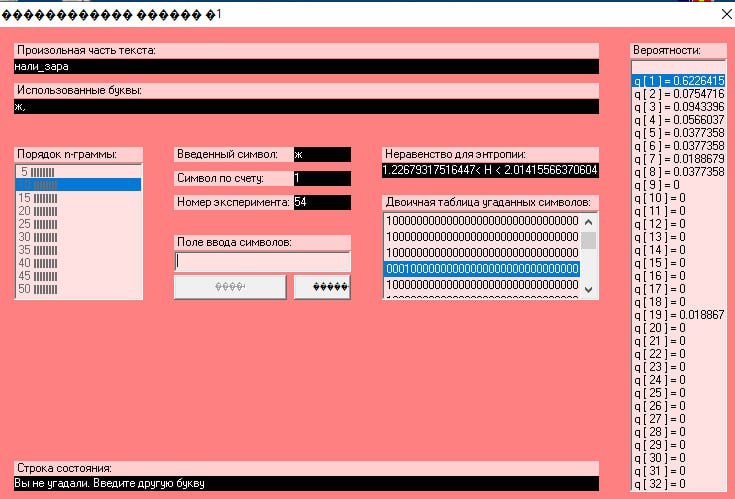


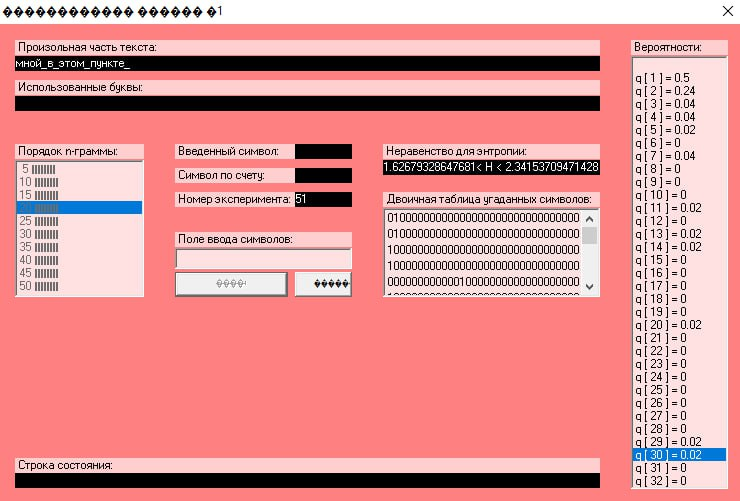
Частоти біграм:



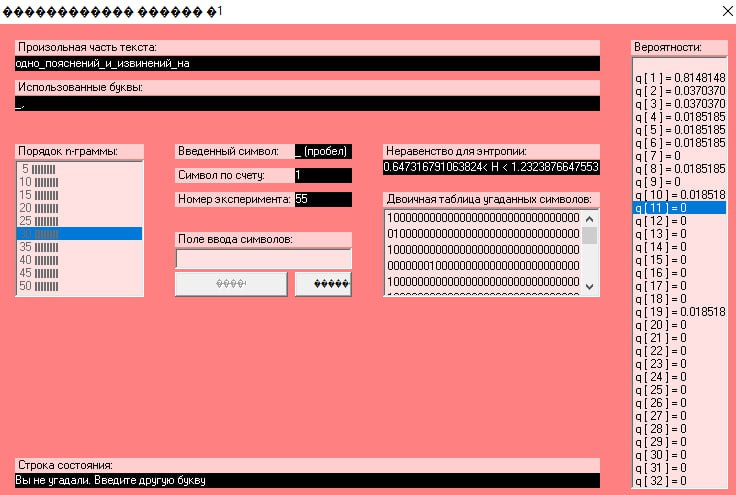
2. За допомогою програми CoolPinkProgram оцінити значення (10) H , (20) H , (30) H.

H(10)



H(20)

H(30)



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



H0=log232=5 (32 це довжина алфавіта)

H(10)

RL0,754641364967

RR0,597168867258

H(20)

RL0,674641342704

RR0,531692581057

H(30)

RL0,870536641787

RR0,753522467048

Висновки: вирахувавши частоту появи символів та груп символів у довільному тексті, можна визначити їх ентропію, яка є мірою кількості інформації. Коли джерело даних має менш імовірне значення (наприклад, коли стається низькоймовірна подія), то ця подія несе більше «інформації» («неочікуваності»), ніж коли джерело даних має більш імовірне значення. Визначена таким чином кількість інформації, що передається кожною подією, стає випадковою змінною, чиє математичне сподівання є інформаційною ентропією. За отриманими результатами можна вирахувати приблизне значення надлишковості мови, тобто наскільки легко вгадати наступний символ, знаючи n попередніх.

російська мова - лайно, вивчайте українську