**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 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 імовірності замінити відповідними частотами. Також одержати значення H1 та H2 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**

0. Опис функцій програмного коду

**Функція пошуку кількості повторень букв в заданому тексті по заданому алфавіті та кількості букв у тексті:**

def ocr(dictionary\_ocr, alph, text):

sum = 0

for letter in alph:

dictionary\_ocr[letter] = text.count(letter)

sum += dictionary\_ocr[letter]

return sum

**Функція пошуку ймовірності повторень букв в заданому алфавіті:**

def apr(dictionary\_apr, dictionary\_ocr, alph, sum):

for letter in alph:

dictionary\_apr[letter] = dictionary\_ocr[letter] / sum

**Функція пошуку ентропії по заданому ансамблю:**

def ngramma(dictionary\_apr):

ngram = 0

for letter in dictionary\_apr:

if dictionary\_apr[letter] != 0:

ngram -= (dictionary\_apr[letter] \* math.log(dictionary\_apr[letter], 2))

return ngram

**Функція виводу результатів у ексель-файл:**

def toexl(dictionary\_apr, entr, redun, name):

col1 = "Алфавіт"

col2 = "Ймовірність"

col3 = "Результати"

arr1 = []

arr2 = []

for letter in dictionary\_apr:

arr1.append(letter)

arr2.append(dictionary\_apr[letter])

arr3 = ["Ентропія", entr, "Надлишковість", redun]

for i in range(4, len(arr1),1):

arr3.append(" ")

data = pandas.DataFrame({col1: arr1, col2: arr2, col3: arr3})

with pandas.ExcelWriter("results.xlsx", mode="a", engine="openpyxl") as writer:

data.to\_excel(writer, sheet\_name=name, index=False)

**Головна зв’язна функція:**

def func(alph, text, n, name):

dictionary\_ocr = {}

dictionary\_apr = {}

sum = 0

sum = ocr(dictionary\_ocr, alph, text)

apr(dictionary\_apr, dictionary\_ocr, alph, sum)

entrop = 1 / n \* ngramma(dictionary\_apr)

red = 1 - (entrop / math.log(len(alph), 2))

toexl(dictionary\_apr, entrop, red, name)

return entrop

**Функція створення алфавіту з тексту по заданому кроці:**

def alph\_wht\_step\_n(text, n):

alph = []

for i in range(0, len(text) - n, n):

temp = text[i] + text[i + 1]

if temp not in alph:

alph.append(temp)

return alph

**Функція пошуку надлишковості:**

def redun(entr, alphabet):

return 1 - (entr / math.log(len(alphabet), 2))

1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Алфавіт** | **Ймовірність** | **Результати** | **Алфавіт** | **Ймовірність** |  |
|  | 0,15733862 | Ентропія | ь | 0,018589921 |
| о | 0,088435486 | 4,40997043 | я | 0,017969994 |
| а | 0,070965734 | Надлишковість | ы | 0,016908355 |
| е | 0,069921294 | 0,118005914 | г | 0,015717952 |
| и | 0,056144347 |  | з | 0,014654162 |
| н | 0,054576732 | б | 0,014410491 |
| т | 0,050166054 | ч | 0,012181383 |
| л | 0,043969175 | й | 0,009497422 |
| с | 0,040638532 | ж | 0,008463494 |
| р | 0,04048564 | ш | 0,00764242 |
| в | 0,035321973 | х | 0,0074869 |
| к | 0,030686497 | ю | 0,005628075 |
| д | 0,026067266 | ц | 0,00352343 |
| м | 0,025822878 | э | 0,003360983 |
| у | 0,024845568 | щ | 0,00332706 |
| п | 0,02303691 | ф | 0,002215253 |

**H1 з пробілом**

Ентропія - 4,40997043

**H1 без пробілу**

Ентропія - 4,488239

**H2 з пробілом, перетинаються**

Ентропія - 4,02682105336401

**H2 без пробілу, перетинаються**

Ентропія - 4,18629661105217

**H2 з пробілом, не перетинаються**

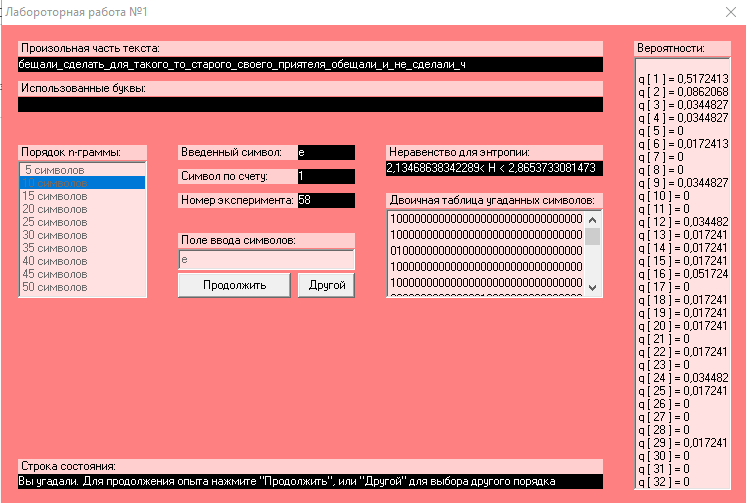
Ентропія - 4,0267889213736

**H2 без пробілу, не перетинаються**

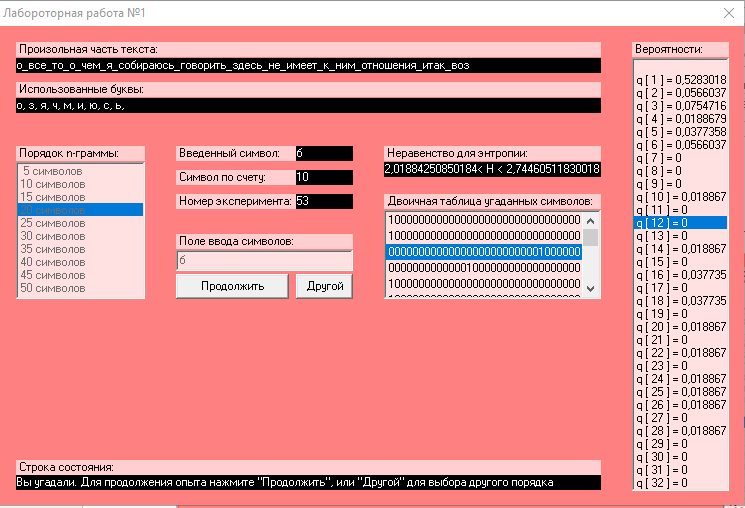
Ентропія - 4,18626995986481

2.

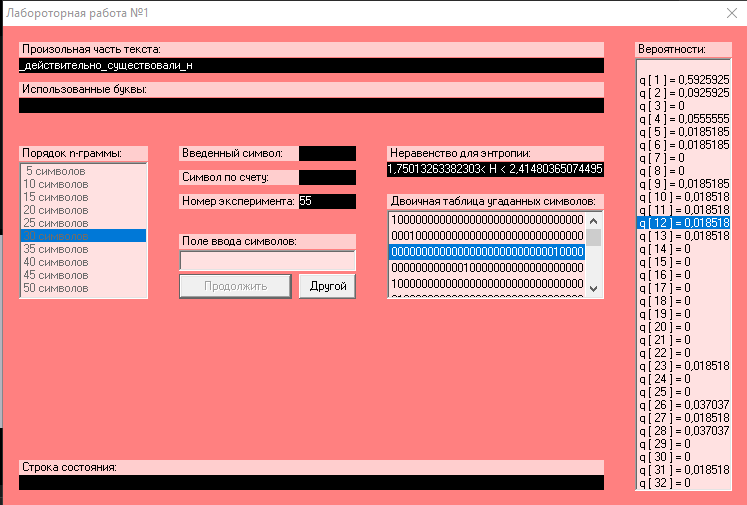
H (10)



H (20)



H (30)



2,13468638342289 < H < 2,8653733081473

2,01884250850184 < H < 2,74460511830018

1,75013263382303 < H < 2,41480365074495

***Значення ентропії зменшується з кількістю відомих символів***

3.

**H1 з пробілом**

Надлишковість - 0,118005914

**H1 без пробілу**

Надлишковість - 0,0940529955519138

**H2 з пробілом, перетинаються**

Надлишковість - 0,587691687492101

**H2 без пробілу, перетинаються**

Надлишковість - 0,574460770473293

**H2 з пробілом, не перетинаються**

Надлишковість - 0,586705701421527

**H2 без пробілу, не перетинаються**

Надлишковість - 0,573983686613964

**Висновок**

Під час виконання комп’ютерного практикуму №1 засвоїли такі поняття, як ентропія на символ джерела та надлишковість джерела тексту, навчилися визначати їх наближені значення, для чого порівнювали різні моделі джерел відкритого тексту. В результаті, отримали практичні навички щодо оцінки ентропії на символ джерела.

(код програми та результати всіх експериментів прикріплюються)