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 H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1 та H_2 на тому ж тексті, в якому вилучено всі пробіли.
- 2. За допомогою програми CoolPinkProgram оцінити значення $\mathbf{H}^{(10)}$, $\mathbf{H}^{(20)}$, $\mathbf{H}^{(30)}$.
- 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"
   со12 = "Ймовірність"
   со13 = "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):
```

```
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))
```

Алфавіт	Ймовірність	Результати	Алфавіт	Ймовірність
-	0,15733862	Ентропія	Ь	0,018589921
o	0,088435486	4,40997043	Я	0,017969994
a	0,070965734	Надлишковість	Ы	0,016908355
e	0,069921294	0,118005914	Γ	0,015717952
И	0,056144347		3	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		X	0,0074869
К	0,030686497		Ю	0,005628075
Д	0,026067266		ц	0,00352343
M	0,025822878		Э	0,003360983
V	0,024845568		Щ	0,00332706
П	0,02303691		ф	0,002215253

Н1 з пробілом

Ентропія - 4,40997043

Н1 без пробілу

Ентропія - 4,488239

Н2 з пробілом, перетинаються

Ентропія - 4,02682105336401

Н2 без пробілу, перетинаються

Ентропія - 4,18629661105217

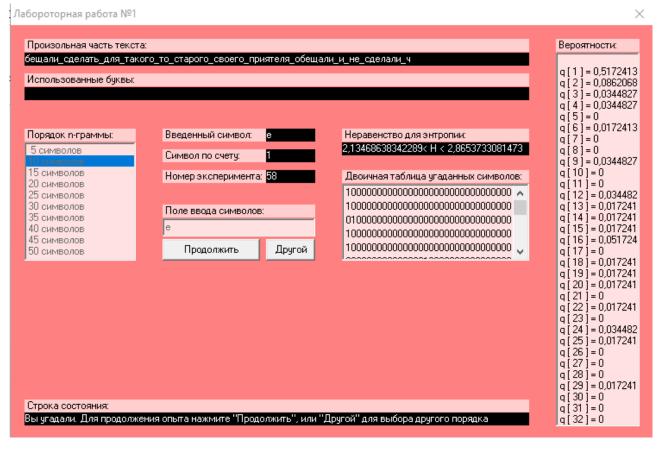
Н2 з пробілом, не перетинаються

Ентропія - 4,0267889213736

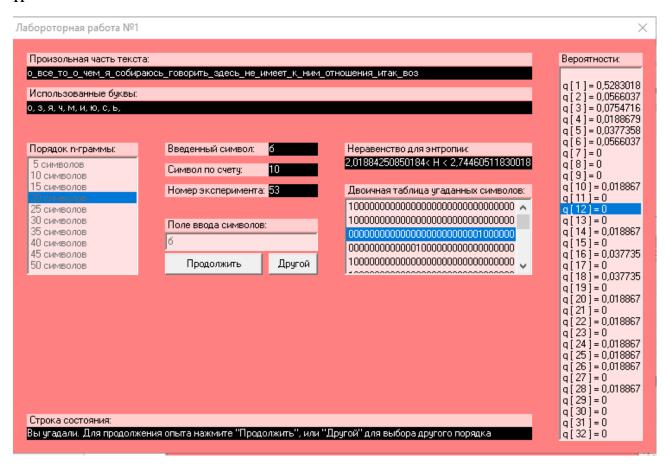
Н2 без пробілу, не перетинаються

Ентропія - 4,18626995986481

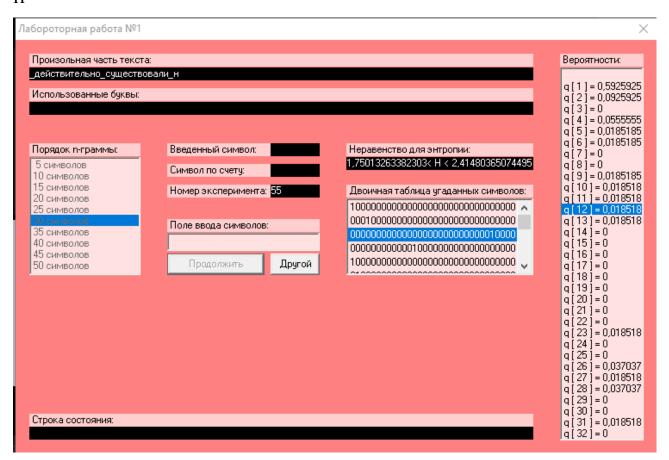
H (10)



 $H^{(20)}$



 $H^{(30)}$



2,13468638342289 < H < 2,8653733081473

2,01884250850184 < H < 2,74460511830018

1,75013263382303 < H < 2,41480365074495

Значення ентропії зменшується з кількістю відомих символів

3.

Н1 з пробілом

Надлишковість - 0,118005914

Н1 без пробілу

Надлишковість - 0,0940529955519138

Н2 з пробілом, перетинаються

Надлишковість - 0,587691687492101

Н2 без пробілу, перетинаються

Надлишковість - 0,574460770473293

Н2 з пробілом, не перетинаються

Надлишковість - 0,586705701421527

Н2 без пробілу, не перетинаються

Надлишковість - 0,573983686613964

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)