КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

студентки групи ФБ-23

Сівашенко Анна,

Тарасенко Ангеліна

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 H1та H2 на тому ж тексті, в якому вилучено всі пробіли.

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

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

Хід роботи

Код програми містить 6 функцій підрахунку частоти букв та біграм з та без пробілів та відповідно з кроком 1 та 2. На цьому етапі труднощів не виникало.

- Підрахунок частоти букв з пробілами відбувався наступним чином:

Спочатку перевіряли чи належить літера до російського алфавіту, переводили її в нижній регістр, функцією change_bad, що заміняє літери «ъ» на «ь» та «ё» на «е».

```
char = char.lower()
if re.match(r'[а-яА-ЯёЁ ]', char):
    char = change_bad(char)
    freq_dict[char] += 1
```

Та ділили частоту знайденої літери на загальну кількість літер в тексті (за це відповідає функція calc_freq).

```
freq_dict = calc_freq(freq_dict, counter)
```

Для підрахунку літер без пробілу алгоритм аналогічний, але не враховуємо пробіл.

- Підрахунок біграм з кроком1 та 2 без пробілу.

```
if first_char == ' ':
      i += 1
      continue
```

Якщо перший символ виявляється пробілом, ітерація пропускає цей символ і переходить до наступного символа. Після знаходження першого символа, функція шукає другий символ, пропускаючи всі пробіли. Для цього змінна j ініціалізується як індекс наступного символа після i, і якщо це пробіл, вона збільшується, поки не знайде літеру або кінець тексту.

```
j = i + 1
while j < len(text) and text[j] == ' ':
    j += 1
if j >= len(text):
    break
```

Ну і формується біграма, після чого обраховується її частота в тексті.

- Підрахунок біграм з кроком1 та 2 з пробілу.

На кожному кроці обирається два символи та формується пара.

```
first_char = change_bad(text[i].lower())
second_char = change_bad(text[i + 1].lower())
pair = first_char + second_char
```

Далі вони перевіряються на відповідність вимогам та обраховується їх частота.

Наступним завдання треба було обрахувати ентропію відкритого тексту.

Для обрахунку значення H_1 та H_2 ми користувались формулами:

$$H_n = \frac{1}{n} H(x_{1,} x_{2\ldots} x_n),$$

де $H(x_1, x_{2,} \ldots x_n)$ – ентропія n-грами відкритого тексту $(x_1, x_{2,} \ldots x_n)$.

З цього слідує:

$$H_1 = -\sum_{i=1}^{n} p_i \times \log_2 p_i$$

$$H_2 = -\frac{1}{2}\sum_{i=1}^{n} p_i \times \log_2 p_i$$

P_i – частота появи букви в тексті, n – загальна кількість символів.

У коді це реалізовано наступним чином:

```
total_count = sum(freq_dict.values())
```

```
    for freq in freq_dict.values():
        if freq > 0:
            p = freq / total_count
            entropy -= p * math.log2(p)
```

Для обчислення надлишковості джерела відкритого тексту використаємо формулу:

$$R = 1 - \frac{H_\infty}{H_0}$$

Де H_∞ – ентропія джерела (H_1 , H_2), а $H_0 = \log_2 m$ (m – кількість букв в алфавіті).

В коді це реалізовано так:

```
H0 = math.log2(alphabet_size)
redundancy = 1 - (H / H0) if H0 > 0 else 0
```

Всі результати обчислень частоти виводяться в файли. Далі буде продемонстровано їх вміст.

Символ	Частота з пробілом	Частота без пробілу
г	0.01561860519503989	0.018875317559473147
у	0.023449168442816824	0.028338670152460665
с	0.043721198462854986	0.05283772108723371
т	0.05242743474309162	0.0633593376134944
о	0.09288645218213999	0.11225466423003672
й	0.008768689440366533	0.010597092102685324
	0.17253814957930488	-
а	0.06389622676609565	0.07721954399904933
р	0.035983650559952375	0.04348677892721906
м	0.026846886112011737	0.03244486268262681
з	0.013645351093063182	0.01649061051712011
н	0.05579591901818317	0.06743020115044049
п	0.020531144195045207	0.024812194283768896
л	0.038998676524521506	0.04713048281886825
я	0.017262319218033284	0.020861770496436588
е	0.07278849373608091	0.08796598139125575
к	0.02662896444687761	0.032181501096804656
ю	0.005372300561812683	0.006492505435847366
х	0.007168825508528178	0.00866363265555195
д	0.02632732897135659	0.031816970121306916
ж	0.00927230109332894	0.01120571430406701
и	0.05564709446638425	0.0672503444576839
в	0.03671448541253634	0.04437000375772019
ы	0.016361399163393412	0.01977299515992793
ь	0.016694925614299912	0.02017606685530208
б	0.014870496064122121	0.017971216505705632
щ	0.0029366273167464826	0.0035489579552863415

ц	0.0025964569126346727	0.003137856943271272
ш	0.006873833986212469	0.008307130996695134
ч	0.013205521390871739	0.01595906975549125
э	0.003279455302140416	0.003963270693957779
ф	0.0008916185201524389	0.0010775342932113734

Біграма з кроком 1 з пробілом:

Character	Frequency
гу	0.00048235100270541776
ус	0.0009713459586161442
ст	0.01084558921234606
то	0.0129185026124024
ой	0.003286099255345725
й	0.005129131874498381
а	0.001940034335950165
ар	0.0021566272104432314
ро	0.00633035861401821
ом	0.005002896763597515
ма	0.0024423171982715093
ат	0.00446207897268538
т	0.004217581494730016
р	0.0037671214674100807
оз	0.0012291313429821252
з	0.0012291313429821252
н	0.015016663034638915
на	0.009807803721676828
ап	0.0007441227589945838
по	0.00804051216906469
ол	0.005515809951047353
лн	0.0002989778942388953
ня	0.00127829659670141
ял	0.0006803408082236195
л	0.0048567297930807215
м	0.006295810057350604
ас	0.004214923913447893
те	0.00562742836489654
ер	0.005591551017587873

Біграма з кроком 1 без пробілу:

Character	Frequency
гу	0.0006070163380535009
ус	0.0016267395514502551
ст	0.01333348321391063
то	0.01588841176905116
ой	0.003971300010598698
йа	5.138762650188368e-05

ар	0.0029290947106073695
ро	0.007676026708718874
ом	0.006942147167738848
ма	0.0029772706104528855
ат	0.006071769243863193
тр	0.0035826810851782025
оз	0.002066746103372634
зн	0.0025131761086077484
на	0.011860906541966027
ап	0.002055505060075347
по	0.009717078998840566
ол	0.007149303537074566
лн	0.0008784072405165741
ня	0.0015544757016819812
ял	0.0010261466667094896
лм	0.00017664496610022515
ас	0.006442723672673666
те	0.006972658570974341
ер	0.007203902890232818
рс	0.00023927363589939588
ск	0.004269990589640896
ку	0.0017792965676277223
ую	0.0011497981429796473

Біграма з кроком 2 з пробілом:

Character	Frequency
гу	0.0002498126405196103
ст	0.0055211251136116
ой	0.0016742762077378136
а	0.0009899490275910089
ро	0.0032223173045747605
ма	0.001257035946444422
т	0.0021885181858287135
з	0.0006072573229652229
на	0.00486337374628603
по	0.003991687085749518
лн	0.0001634412488505961
ял	0.0003361840321886245
м	0.003126644378418314
ас	0.002078228562620588
те	0.0027904603462296893
рс	7.042590397627311e-05
ку	0.0006630665298898167
ю	0.0010590461409262202
ху	9.965929807963177e-05
до	0.002266916833651357
жн	0.0004252130051397622

ик	0.0009766611211803912
к	0.004246814888833375
ог	0.0023546170159614334
да	0.0022748895774977276
в	0.007692369021106511
с	0.007671108370849522
ад	0.0009593868428465884
у	0.0020210905650549322

Біграма з кроком 2 без пробілу:

Character	Frequency
гу	0.00033241370893406005
ст	0.006670756265275775
ой	0.002049081606762612
ар	0.0014613356286473172
ом	0.003460635472236229
ат	0.0030366875535956885
ро	0.0038926127075176887
зн	0.0012622085759525178
ап	0.0011000163798059474
ол	0.0036308569850237184
ня	0.0007290619509954747
лм	7.708143975282552e-05
ас	0.0032020914763986265
те	0.0034799058321744354
рс	0.0001381042462238124
ку	0.0008944658737984127
юх	9.63517996910319e-06
уд	0.0014019185685504514
ож	0.0013826483255663076
ни	0.0047886844446442854
ка	0.004001811413834192
ко	0.004814378257895227
гд	0.0009281890036902739
ав	0.0021582803130791146
са	0.0009474593636284803
ду	0.0010566580699449832
по	0.004880218654350766
дн	0.0010534463432886154
им	0.0019061597705542477

Обрахунки H_1 , H_2 та R:

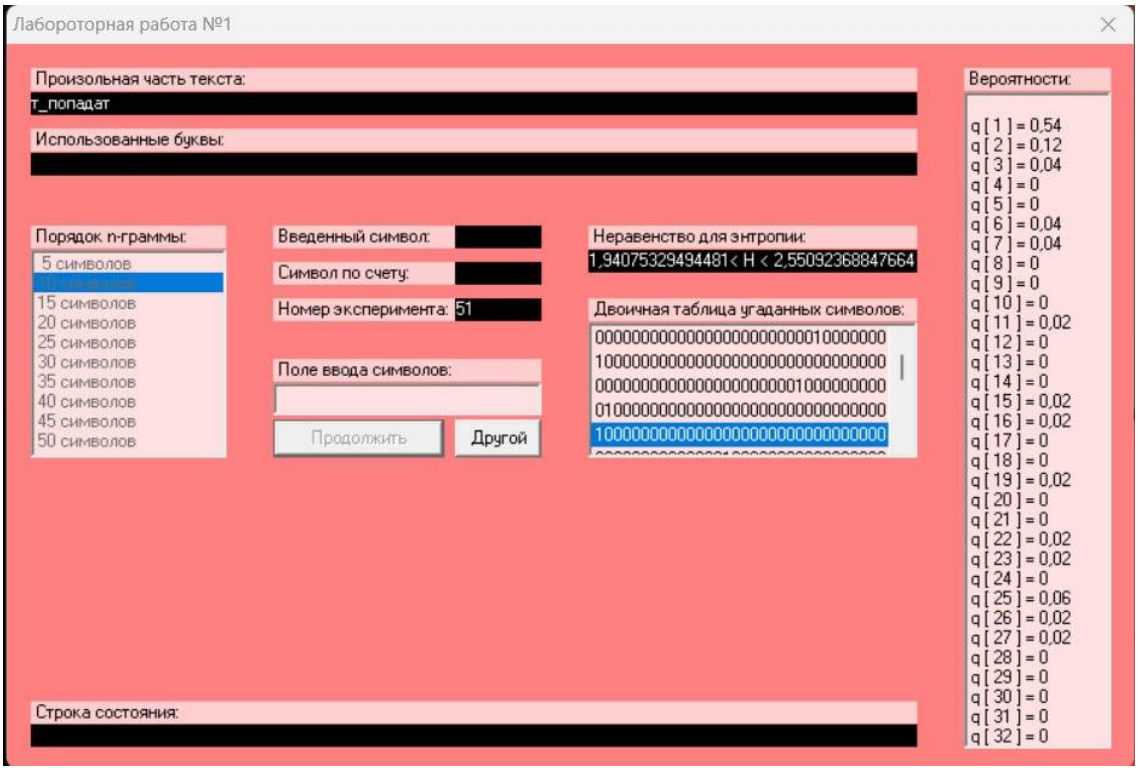
Модель відкритого тексту	Ентропія	Надлишковість
H1	4.353302086326825	0.1443078363090472
H1_w	4.459207272689106	0.11600736041294324
H2	3.9972747728131646	0.21428914617275918
H2_w	4.1160860241323	0.1840276697777563
H2_2	3.997425960368093	0.2142594285001892
H2_w2	4.117570056763143	0.1837334753520773

Оцінки для H^{10}, H^{20}, H^{30}

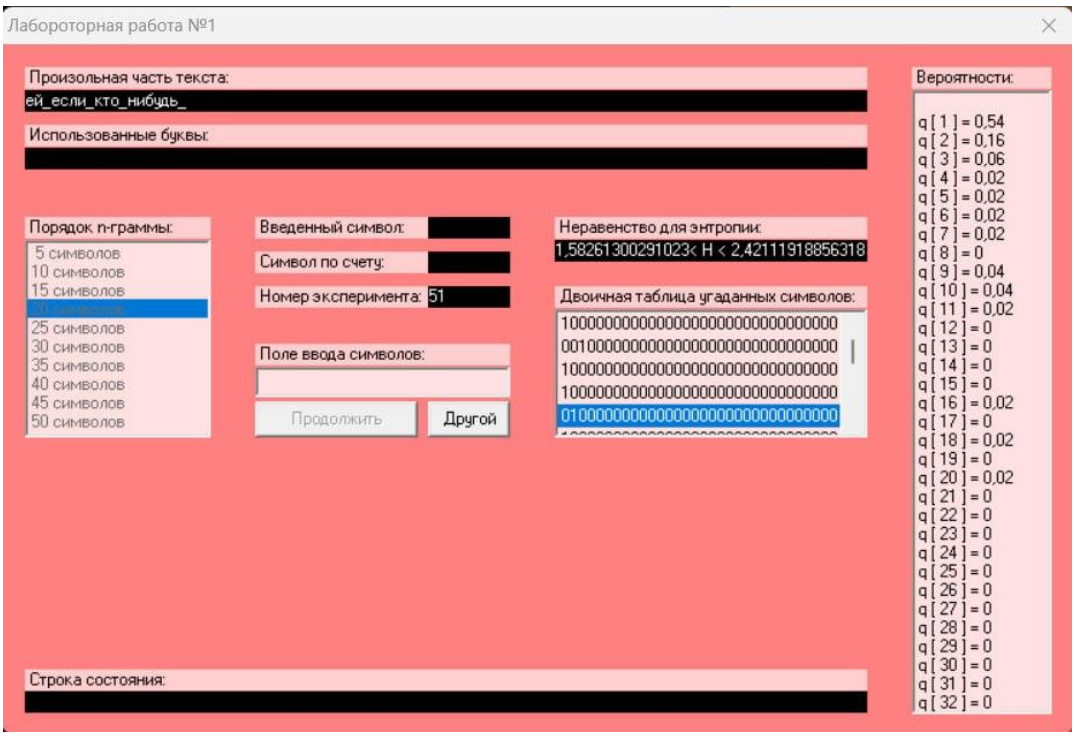
H	R
1,9407533< H^{10} <2,5509237	0,61184934> R^{10} >0,48981526
1,582613< H^{20} <2,4211192	0,6834774> R^{20} >0,51577616
1,6443866< H^{30} <2,3244549	0,67112268> R^{30} >0,53510902

Обрахували значення надлишковості для кожної моделі, де H_∞ – ентропія джерела, а $H_0 = \log_2 32 = 5$.

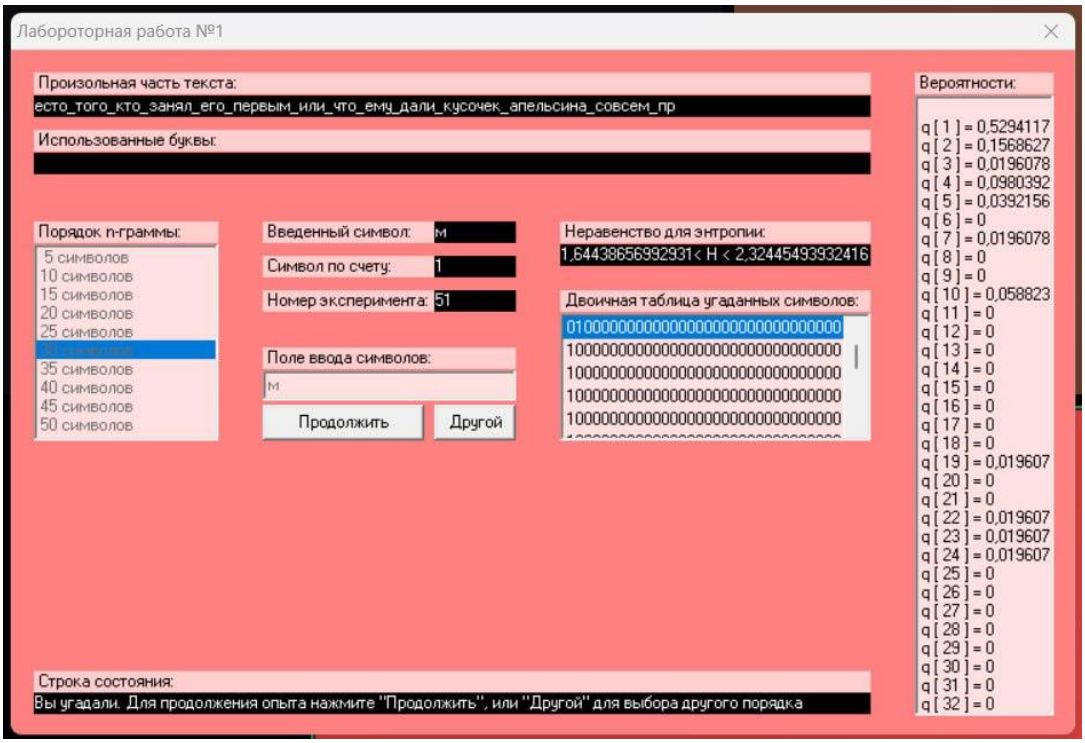
Результат роботи програми для H^{10} :



Результат роботи програми для H^{20} :



Результат роботи програми для H^{30} :



Висновки

У результаті виконання комп’ютерного практикуму ми ознайомились з понятям ентропії та надлишковості джерела відкритого тексту. Ми порівняли різні моделі відкритого тексту (монограми, біграми з кроком 1 та 2) з пробілом та без нього. Визначили ентропію та надлишковість для кожної моделі.

У ході виконання практикуму було відмічено, що H_1 з пробілами менша за H_1 без пробілів, але в загальному ентропія для всіх моделей тексту дорівнює близько 4 і надлишковість теж не сильно відрізняється.

При аналізі програми CoolPinkProgram зробили висновок, що чим менше літер дано, тим важче вгадати наступний символ або ж $H_{10} > H_{30}$ і чим більше дано тексту, ти більша його надлишковість ($R_{10} < R_{30}$).