Міністерство освіти і науки України
Національний технічний університет України
“Київський політехнічний інститут ім. Ігоря Сікорського”
Фізико-технічний інститут

Лабораторна робота № 1

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали:
Студенти 3 курсу
групи ФБ-92
Сидоренко Андрій
Варгіч Дмитро

Перевірила:
Селюх П.В.

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 1 Мб), 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 $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.

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

Виконання

1. код програми міститься у файлі **base.py**;

текст, що аналізувався - у файлі **raw_text.txt;**

результати аналізу: **Char_spaces.csv, Char_no_spaces.csv, Bigrams_spaces.csv, Bigrams_no_spaces.csv, H_and_R.txt**

Результати

H_1 з пробілами=4.389834609607416

H_1 без пробілів=4.449759325578299

H_2 з пробілами =3.9947022441300635

H_2 без пробілів =4.131904131896036

Частота букв з пробілами:

Літера	Частота
	0.15117749875452802
о	0.09643246164368524
е	0.07258695083129976
а	0.06301834854638644
и	0.06019471582129377
н	0.054081348318401744
т	0.0516056033572858
с	0.0470864399766949
р	0.04386763377213351
л	0.04218899087216813
в	0.03655523562640907
к	0.02914995482525395
д	0.025853464945241454
п	0.02533501085038293
м	0.024176510820829358
у	0.023560107743880298
ы	0.017196801459102077
ь	0.016832025939592498
б	0.0163676126624391
г	0.015773163667682746
я	0.015614418765673948
з	0.01378716361701948
ч	0.01111045436506261
й	0.009479097180589214
ж	0.008835673694787594
х	0.0078105869339435445
ш	0.006682484864349103
ю	0.004259091945384999
щ	0.0029857551782080404
ц	0.0023592194479392717
ф	0.002127857197139214
э	0.0019066275996588674

Частота букв без пробілів:

Літера	Частота
о	0.11360733428035955
е	0.08551487587192068
а	0.074242080592727
и	0.07091555152339911
н	0.06371337734219877
т	0.060796695753900516
с	0.055472657602272865
р	0.05168057362730937
л	0.04970296005379746
в	0.043065818322171955
к	0.03434163771870138
д	0.030458034403313797
п	0.029847242283526917
м	0.028482410380680664
у	0.027756224309924277
ы	0.020259596598663818
ь	0.01982985360884634
б	0.019282727117180567
г	0.01858240520784838
я	0.018395387425242625
з	0.016242693374397665
ч	0.013089255231524035
й	0.01116734908261809
ж	0.010409330197801148
х	0.009201672814378881
ш	0.007872652827350755
ю	0.005017647316294819
щ	0.0035175259536912138
ц	0.0027794025776621185
ф	0.0025068341072686253
э	0.0022462029421478394

Частота біграм з пробілами:

Біграма	Частота
('с', 'ш')	9.96379259099138e-05
('а', 'э')	9.794914750466103e-05
('я', 'г')	9.794914750466103e-05
('е', 'я')	9.626036909940825e-05
('ь', 'т')	9.457159069415548e-05
('т', 'м')	9.457159069415548e-05
('щ', 'у')	9.28828122889027e-05
('й', 'к')	9.28828122889027e-05
('я', 'я')	9.28828122889027e-05
('ч', 'ш')	9.28828122889027e-05
('я', 'е')	9.28828122889027e-05
('э', 'р')	9.119403388364992e-05
('б', 'ь')	9.119403388364992e-05
('м', 'п')	8.781647707314436e-05

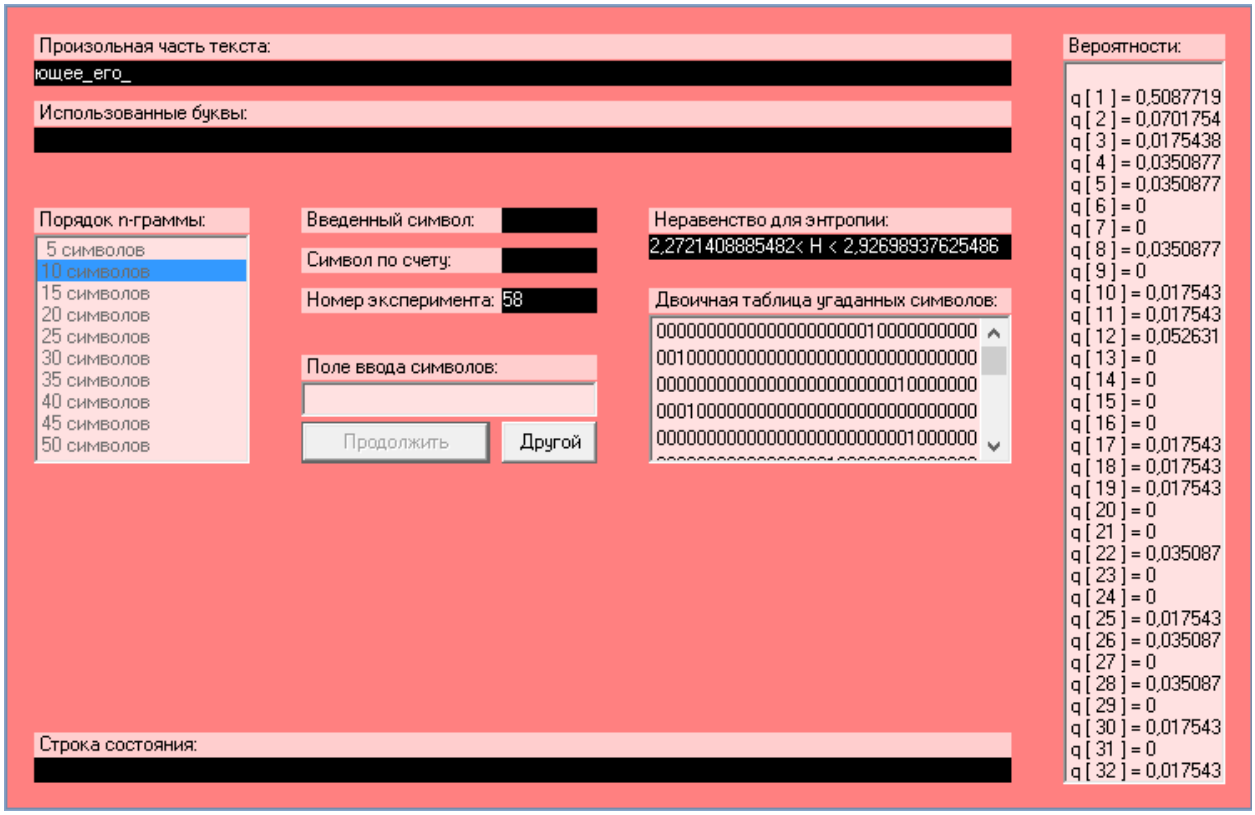
Частота біграм без пробілів:

Біграма	Частота
('э', 'п')	9.94777418552599e-06
('ж', 'л')	9.94777418552599e-06
('п', 'м')	9.94777418552599e-06
('ц', 'ч')	9.94777418552599e-06
('н', 'э')	9.94777418552599e-06
('ю', 'я')	9.94777418552599e-06
('ы', 'ю')	9.94777418552599e-06
('к', 'ц')	9.94777418552599e-06
('х', 'ю')	9.94777418552599e-06
('х', 'ш')	9.947774185525989e-05
('я', 'ш')	9.748818701815469e-05
('т', 'ю')	9.748818701815469e-05
('н', 'ю')	9.549863218104949e-05
('ш', 'т')	9.549863218104949e-05

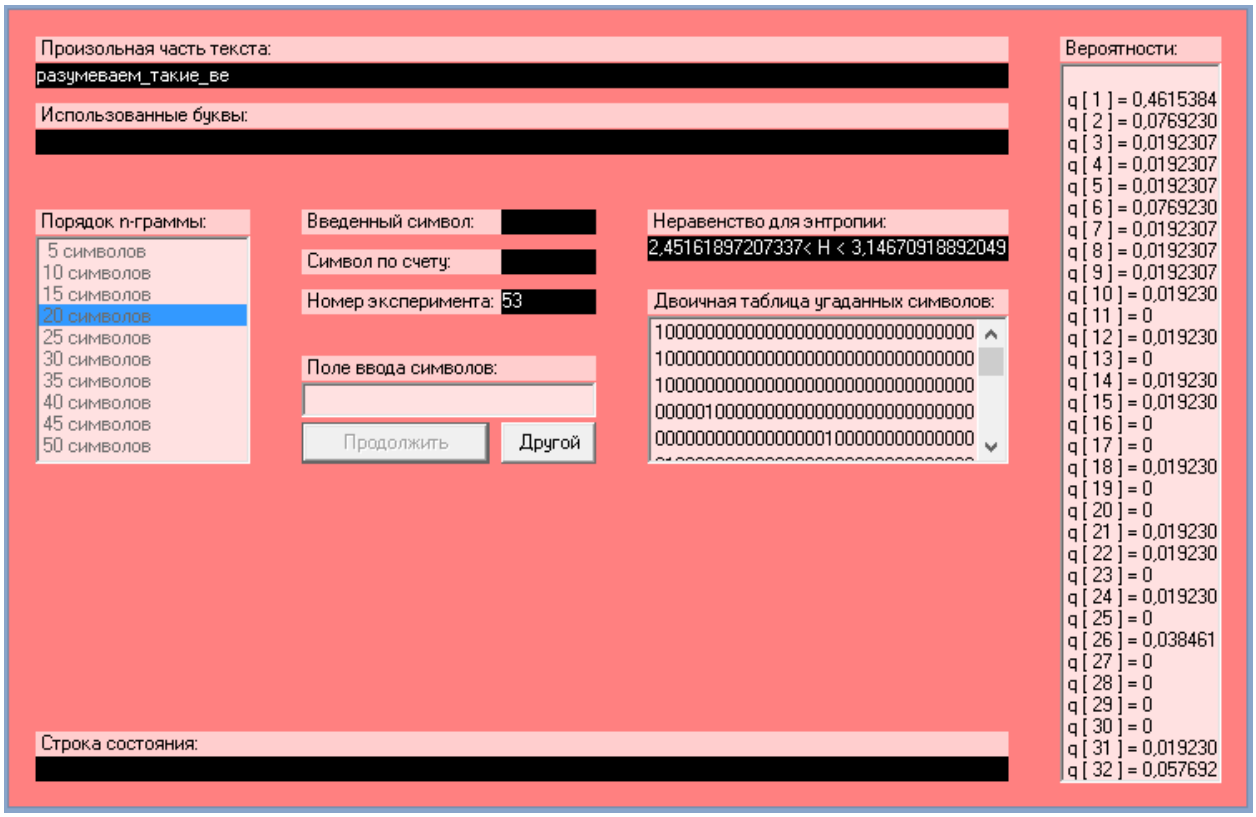
*Наведені приклади 14 найуживаніших біграм

2.

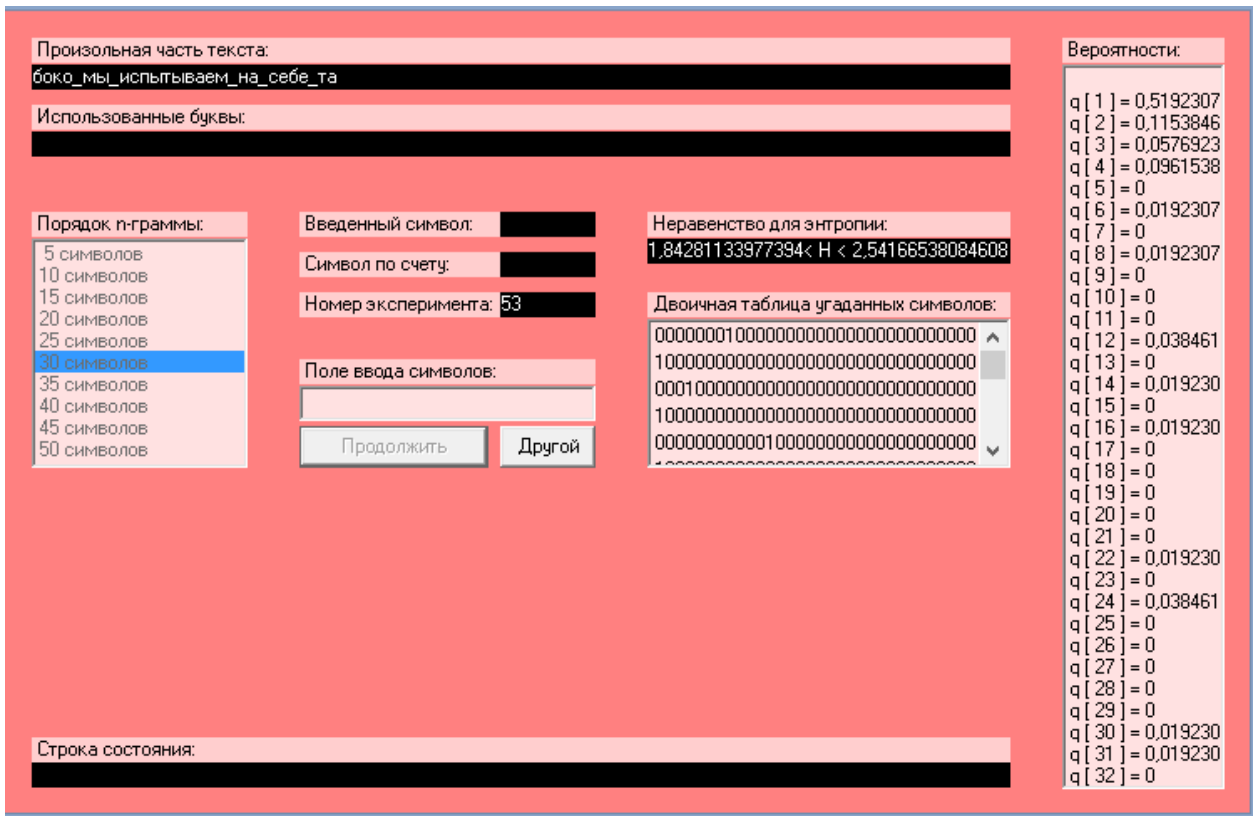
$$H^{(10)}$$



$H^{(20)}$


Произольная часть текста:
разумеваем_такие_ве
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 53
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
2,45161897207337< H < 3,14670918892049
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0,4615384
q [2] = 0,0769230
q [3] = 0,0192307
q [4] = 0,0192307
q [5] = 0,0192307
q [6] = 0,0769230
q [7] = 0,0192307
q [8] = 0,0192307
q [9] = 0,0192307
q [10] = 0,019230
q [11] = 0
q [12] = 0,019230
q [13] = 0
q [14] = 0,019230
q [15] = 0,019230
q [16] = 0
q [17] = 0
q [18] = 0,019230
q [19] = 0
q [20] = 0
q [21] = 0,019230
q [22] = 0,019230
q [23] = 0
q [24] = 0,019230
q [25] = 0
q [26] = 0,038461
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0
q [31] = 0,019230
q [32] = 0,057692

$H^{(30)}$


Произольная часть текста:
боко_мы_испытываем_на_себе_та
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 53
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,84281133977394< H < 2,54166538084608
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0,5192307
q [2] = 0,1153846
q [3] = 0,0576923
q [4] = 0,0961538
q [5] = 0
q [6] = 0,0192307
q [7] = 0
q [8] = 0,0192307
q [9] = 0
q [10] = 0
q [11] = 0
q [12] = 0,038461
q [13] = 0
q [14] = 0,019230
q [15] = 0
q [16] = 0,019230
q [17] = 0
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0
q [22] = 0,019230
q [23] = 0
q [24] = 0,038461
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0,019230
q [31] = 0,019230
q [32] = 0

H	R
H_1 з пробілами	0.12203307807851682
H_1 без пробілів	0.10182014462183964
H_2 з пробілами	0.20105955117398733
H_2 без пробілів	0.16597892513198087
$H^{(10)}$	$0.414 < R < 0.546$
$H^{(20)}$	$0.37 < R < 0.51$
$H^{(30)}$	$0.492 < R < 0.632$

Висновки

ПІд час виконання лабораторної роботи ми засвоїли поняття ентропії на символ джерела та його надлишковості, вивчили та порівняли різні моделі джерела відкритого тексту для наближеного визначення ентропії, набули практичних навичок щодо оцінки ентропії на символ джерела.

У результаті виконання лабораторної роботи, найуживанішими виявилися:

- символи пробілу, "о", "е", "а" - для випадку з пробілом;
- символи "о", "е", "а" - для випадку без пробілів;
- біграми (сш),(аэ),(яг) - для випадку з пробілом;
- біграми (эп),(жл),(пм) - для випадку без пробілів

У російській мові можливе ущільнення тексту деякою схемою кодування символів без втрати його змісту, причому російська мова більш надлишкова, ніж англійська.