Криптографія

Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали: студенти групи ФБ-13

Гапонов Максим та Квітницький Віталій

Мета: 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 H1() 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 H(10), H(20), H(30)

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

Хід роботи:

1. Підрахунок частот букв і біграм в тексті, підрахунок H1, H2, R1, R2.

Частоти шукаються таким чином:

$$\frac{\text{К-сть разів,скільки зустрічається буква чи біграма в тексті}}{\text{к-сть існуючих різновидів букв чи біграм в тексті}}$$

Таблиця частот символів (з пробілом)

0.16286958701085816	пробіл
0.09547859189465627	о
0.07019528418828024	а
0.0667908501601967	е
0.055606207368371144	и
0.05456035656467815	н
0.04747241433032358	т
0.04391760538098367	с
0.042630612811672346	л
0.03885633776103935	в
0.037757652525553576	р
0.029940866078262695	к
0.025359172531514384	д
0.0247048384146961	м
0.023885227357397835	у

0.021389816501954196	п
0.019252054107877073	я
0.01735407875039795	г
0.016632008182563284	ь
0.01595870785946041	ы
0.014739451741165473	з
0.014518630910852056	б
0.011478618989236678	ч
0.009590126734899852	й
0.00847924893823113	ж
0.007967161368547256	ш
0.007219350949326361	х
0.005378274210701004	ю
0.003309602996660593	ц
0.0025482452872364206	э
0.0023626474114515246	щ
0.0017963706809545421	ф

Таблиця частот символів (без пробіла)

0.1140546208967977	о
0.08385226853440185	а
0.07978547801376203	е
0.0664247846041312	и
0.06517545619906398	н
0.05670850514296071	т
0.052462083206566416	с
0.050924697215874885	л
0.04641611051054161	в
0.045103668364803895	р
0.03576607134765638	к
0.03029297722079452	д
0.02951133781710609	м
0.028532265686192007	у
0.02555135516529327	п
0.022997676119702493	я
0.02073043635869319	г
0.019867881902863097	ь
0.019063586284575003	ы
0.017607115346025128	з
0.01734333227604332	б
0.013711864735925635	ч
0.011455953082216166	й
0.010128946227092948	ж
0.009517228432902002	ш
0.00862392625725808	х
0.006424655140047725	ю

0.003953509447641488	ц
0.0030440242615692344	э
0.0028223170187010867	щ
0.0021458671827968125	ф

Таблиця частот біграм (з пробілами), біграми не перетинаються, (10 найчастіших)

0.021372233978275534	о
0.017844514830238675	и
0.016806790994833058	а
0.016446432587237894	е
0.01565526976755528	с
0.015156728812686784	п
0.015070026037927045	в
0.014777404173112928	н
0.013067733833319335	то
0.011981239687111362	о

Таблиця частот біграм (з пробілами), біграми перетинаються (10 найчастіших)

0.021373588709131155	о
0.017950183836977107	и
0.016793243686276847	а
0.016427466355259202	е
0.015578050108784887	с
0.015208208585200379	п
0.014999580033434757	в
0.014896620488407569	н
0.013289909693641164	то
0.012040847844758682	о

Таблиця частот біграм (без пробілів), біграми не перетинаються (10 найчастіших)

0.01620890324501725	то
0.012914042321808872	ст
0.01240589579435925	на
0.012376766375588252	ов
0.011043286316293702	ал
0.01083938038489672	го
0.010561032605529411	он
0.01034741686787543	но
0.010237372396962773	не
0.010127327926050115	по

Таблиця частот біграм (без пробілів), біграми перетинаються (10 найчастіших)

0.016184654921018688	то
0.012912444916989113	ст
0.012384877923263277	ов
0.012282924792573929	на
0.01099637338149405	ал
0.010776284083497996	го
0.010543248356208056	не
0.010475279602415156	он
0.010226060838507859	но
0.01012410770781851	ко

Тепер вирахуємо ентропії та надлишковості для символів та біграм:

Формула для розрахунку ентропії:

$$H(x_1, x_2, ..., x_n) = -\sum_{z_1, z_2, ..., z_n} P(x_1 = z_1, ..., x_n = z_n) \cdot \log_2 P(x_1 = z_1, ..., x_n = z_n).$$

Формула для розрахунку надлишковості:

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

Результати:

H1 з пробілами: 4.380476562555626

R1 з пробілами: 0.12390468748887484

H2 без перетину з пробілами: 3.97224239356651

H2 з перетином з пробілами: 3.972316749167432

R2 без перетину з пробілами: 0.1668321602319094

R2 з перетином з пробілами: 0.1708102308887568

H1 без пробілів: 4.466861997959686

R1 без пробілів: 0.09836798582354367

H2 без перетину без пробілів: 4.147897312612289

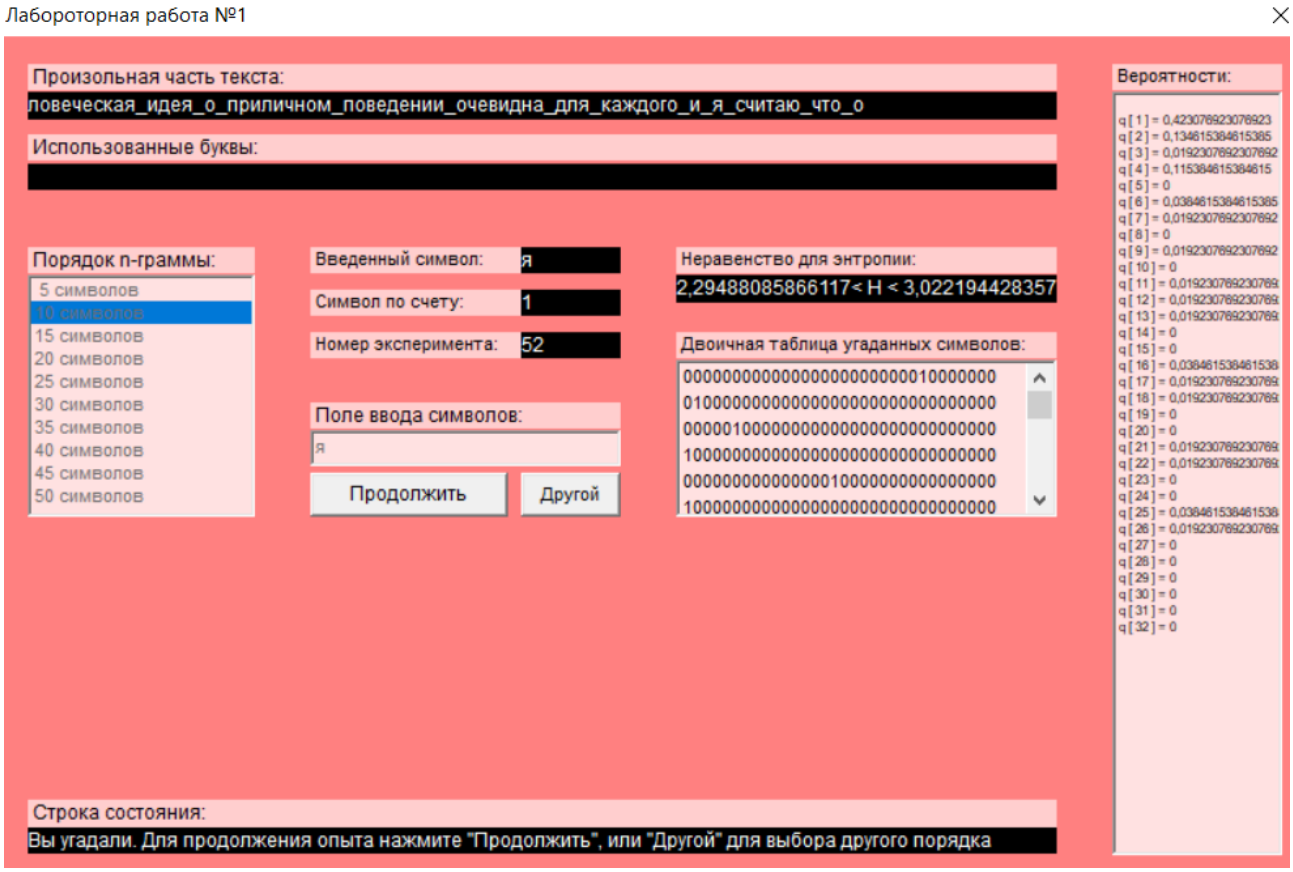
H2 з перетином без пробілів: 4.148552359704669

R2 без перетину без пробілів: 0.1443425182157595

R2 з перетином без пробілів: 0.14841673050230597

2. Оцінка значення H(10), H(20), H(30) за допомогою програми CoolPinkProgram.

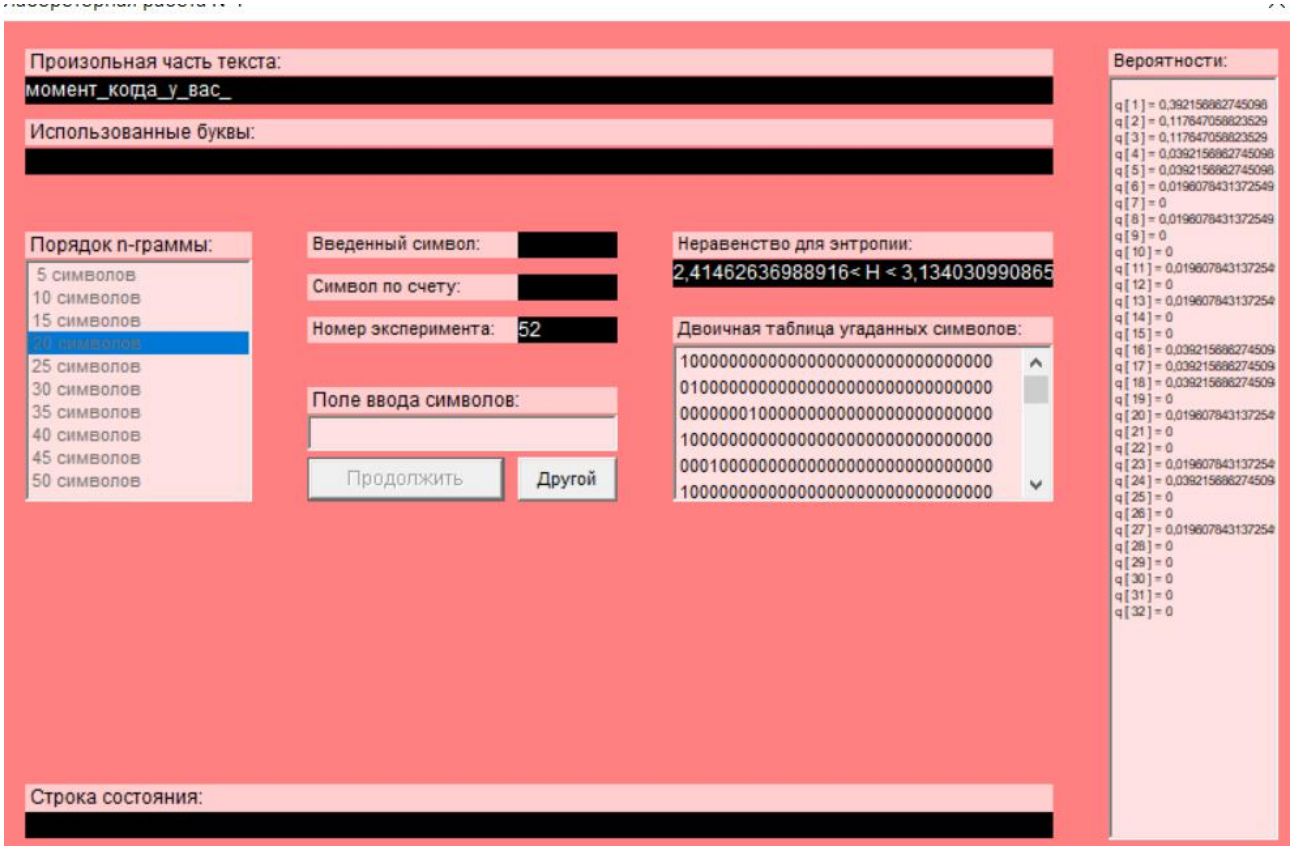
H(10):



Ентропія: 2.29488085866117 < H < 3.022194428357

Надлишковість: 0.39556111432859997 < R < 0.541023828267766

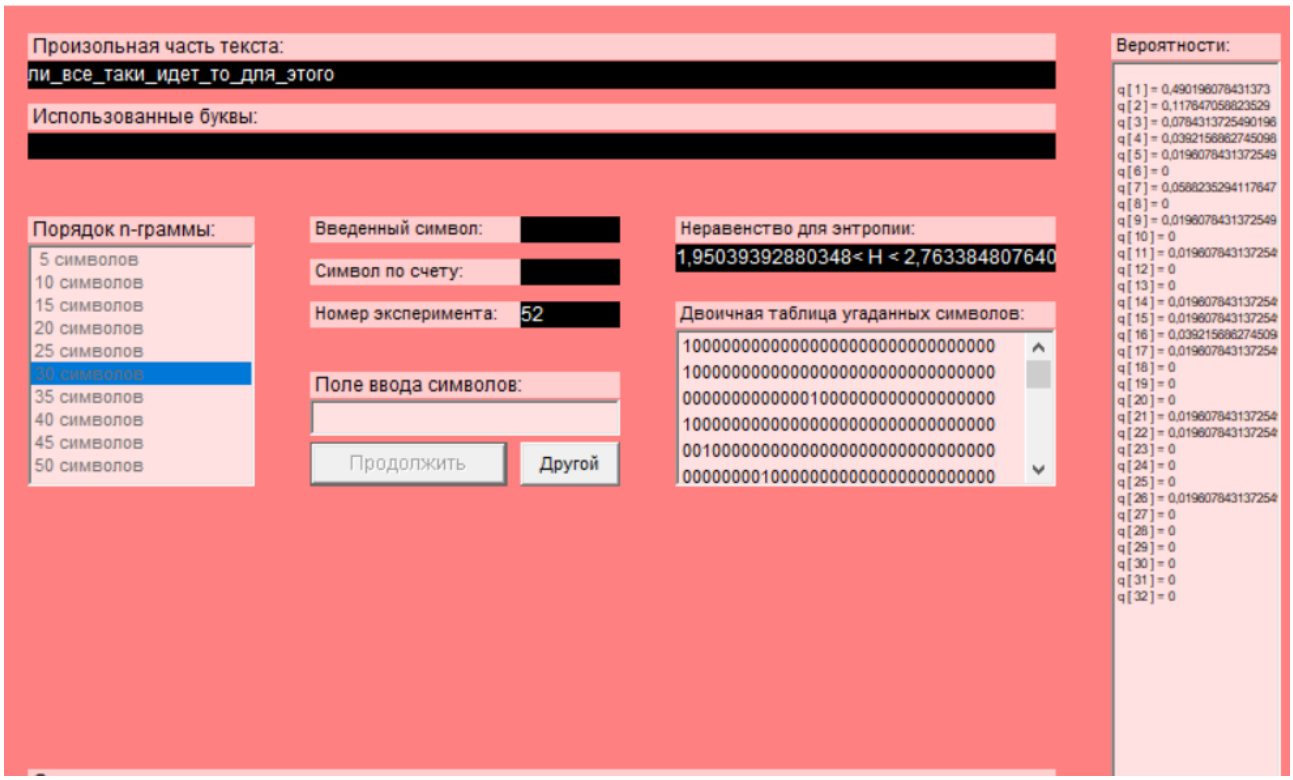
H(20):



Ентропія: 2.41462636988916 < H < 3.134030990865

Надлишковість: 0.373193801827 < R < 0.517074726022168

H(30):



Ентропія: 1.95039392880348 < H < 2.763384807640

Надлишковість: 0.44732303847199995 < R < 0.609921214239304

Труднощі які виникали:

Були труднощі з фільтрацією тексту, коли ми проходились по основному тексту просто через for letter in data, значно пришвидшили код шляхом роботи з індексами, змінений рядок коду: for letter in range(len(data)).

Були труднощі через те, що спочатку було не дуже зрозуміло, що підставляти у формули при пошуку H та R, і куди саме, в який рядок коду, в яке місце в коді заносити цю формулу, однак, пізніше все було зроблено.

Були труднощі, коли намагалися знайти частоту біграм таким же способом, як і частоту символів, в результаті частоти шукало дуууууже довго, довелося оптимізувати код.

Висновки: У ході виконання лабораторної роботи ми засвоїли поняття ентропії на символ його джерела та надлишковості, навчилися шукати частоту, ентропію та надлишковість букв і діаграм у довільному тексті російською мовою. Також ми ознайомилися з не дуже зручною програмою CoolPinkProgram, за допомогою якої погралися і встановили H та R для певного тексту вже з більшими n-грамами: 10, 20, 30.