Міністерство освіти і науки України
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"
Фізико-технічний інститут

«Криптографія»

Комп’ютерний практикум

№ 1

Виконав:
студент групи **ФБ-83**
Гах Валерій

Перевірив:
_____

Київ 2020

Назва: 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;
Постановка задачі:
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 1 H та 2 H за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1 H та 2 H на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1 H та 2 H 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;
Варіант: 6(номер у списку групи), номер бригади відсутній(робота виконана самостійно);

Характеристики обладнання:

- Ноутбук - Lenovo G510;
- ОС - Windows 10 Home x64;
- Процесор - Intel Core i5-4200M, CPU - 2.5GHz;
- Тип системи: 64-розрядна ОС, процесор x64;
- ОЗУ - 6.00 ГБ;

Хід роботи:
Першочергово було написано заголовкові файли – тобто уся задача практикуму була розбита на логічні етапи та розділена між мінімальною кількістю функцій – етап фільтрування тексту від не належних до алфавіту літер, підрахунок кількості пробілів у тексті та загальної кількості літер, генерація усіх можливих n-грам та підрахунок кількості поява кожної в тексті, вивід пар значень <n-грама, частота> у файл(бажано відсортований вивід), підрахунок ентропії тексту за формулою, в яку підставляються значення частот. За завданням практикуму, алгоритм аналізуватиме текст російською мовою, з узятим алфавітом довжиною 32 – «ъ» = «ь», «ё» = «е», та пробіл. Під кінець написання функціонального коду затребуваними стануть функції вибору імен файлів вводу-виводу для програми та прозорий вивід стану виконання програми в консолі(без зайвих повідомлень, залишених після відлагодження коду). Далі буде виконано експеримент з відгадуванням наступної літери, що йде після заданих n штук. Робота проводиться в інтерфейсі програми CoolPinkProgram.exe, який випадковим чином вибирає послідовних n літер з деякого, заданого викладачем, тексту, а експериментатор відгадуватиме, яка літера йде наступна, виходячи зі змісту вибраного фрагменту осмисленого тексту та з правил граматики російської мови(текст саме на ній написаний). Вгадування для різних n має відбуватися не менше 50 разів, щоб отримуване значення ентропії було порівняне з

фактичною ентропією на символ джерела. Під кінець з отриманих значень ентропії для трьох моделей джерела робиться висновок про надлишковість проаналізованих текстів.

Результати:
За посиланням на ПР на гітхабі можна ознайомитися з програмними кодами, розробленими для вирішення задачі даного практикуму. Програмні файли “complete_funcs.cpp”, “funcs.cpp”, “cryptography_formulas_funcs.cpp”, “Lab1_crypto_EntropyCalc.cpp” та заголовкові файли “complete_funcs_h”, “header.h”, “cryptography_formulas_header.h” реалізують алгоритм, що виконує такі етапи розв’язку задачі практикуму:

1. Фільтрує текст довільного змісту, що міститься у заданому самою програмою текстовому файлі, або заданому користувачем файлі(буде виведено запрошення, вибрати файл за замовченням, або інший, і затребувано вказати ім’я файлу, що користувач намагається проаналізувати) та записує вже текст, який містить лише літери з алфавіту, у файл;
2. У циклі перебирає усі можливі n-грами, довжини від 1 до n(яке задасть користувач в консолі на цьому етапі виконання програми) та рахує кількість появ кожної з них у відфільтрованому тексті двома способами – з перетином n-грам та без. Паралельно підраховує кількість пробілів та загальну кількість зчитаних n-грам – їх кількість різна для підрахунку з перетинами та без. Виводить дані про текст(номер мови, довжину шуканих n-грам, довжину тексту та к-ість пробілів) та відповідні пари <n-грама, частота> у лексикографічному порядку в унікальний файл, ім’я якого вказує на зміст даних у ньому, наприклад “1-gramm_rate_data_ANDblanks.txt” – файл з частотами монограм, порахованих для тексту з включеними пробілами;
3. Потім зайвий раз ті самі пари <n-грама, частота> сортуються та виводяться в ще один файл, з приставкою “SORTED” в назві(монограми сортуються за частотою і так само виводяться у файл в стовпчик за спаданням значень частоти, а біграми виводяться у вигляді двовимірної матриці. n-грами більшої довжини не сортуються.);
4. Далі частоти зчитуються зі створених файлів та підставляються у функцію підрахунку ентропії, що реалізує формулу $H_n = -\frac{\sum_{i=1}^{m} p_i * log_2 p_i}{n}$. В окремий файл зайвий раз виводиться інформація про проаналізований текст та відповідні значення ентропій H_1, H_2, … H_n – ентропії монограм, біграм, … n-грам для одного і того ж самого тексту;
5. Етапи 1-5 повторюються ще раз, але відфільтрований текст не міститиме пробілів. Проте все одно вважається, що вони входять в алфавіт;

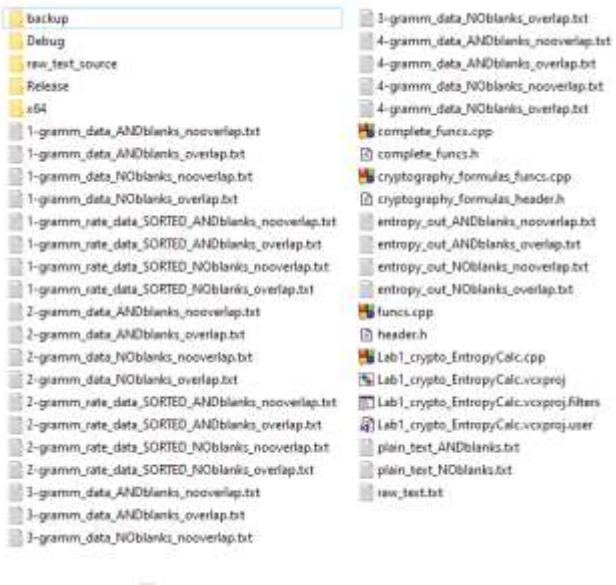
В результаті виконання програми маємо n по 4 файлы з виводом полічених появ n-грам та їх частот у тексті з пробілами та без, з врахуванням перетинів та без урахування, а також відсортовані для 1- та 2- грам частоти та 4 файли ентропій.
Написаний код крім текстів російською мовою придатний і для аналогічного аналізу текстів, написаних англійською.
Складнощі та деталі написання коду:

1. Першою проблемою стало некоректне зчитування та виведення(і у файл, і в консоль) російських літер. Адекватний вивід у консоль був досягнутий зміною кодування символів у локальній консолі, а для роботи програми з файлами необроблених текстів вони повинні мати кодування Windows(1251)(встановлювалось в текстовому редакторі Notepad++ перед копіюванням в них самого тексту російською мовою з джерел. Файли, створювані

самою програму, в які опісля записується російський текст одразу мають коректне кодування.) На гітхабі, тим не менш, файли знову стають непридатними для читання людиною, тому їхнє кодування потрібно змінити самостійно. Тип данних, посередництвом якого відбувається зчитування, зберігання та виведення символів є wchar_t, wstring;

2. Спочатку програма відлагоджувалась на коротких текстах, а тому не було помічено головний недолік реалізації – наївний підрахунок кількості появ n-грами в тексті. Першим було реалізовано підрахунок кількості появ кожної n-грами в тексті шляхом проходу всього тексту один раз. Тобто для кожної n-грами текст повністю переглядався на кількість появ у ньому тільки оцієї даної n-грами. Для невеликих тестів очевидні величезні витрати по часу не були помічені(і не були очевидні на той час), а от на тексті 1 Мб розміром лише з 3 спроби і часом виконання в 45 хв було зроблено висновок оптимізувати саме механізм підрахунку частот. Вже відомий набагато швидший механізм – хеш-таблиці(в таблиці усіх можливих n-грам їм у відповідність ставиться ціле додатнє число – сама кількість появ у тексті, яка ставиться у 0.На єдиному проході по тексту при появі деякої n-грам її відповідне значення в таблиці збільшується на 1. Тобто замість n^m проходів лише один(n – довжина n -грами, m – довжина алфавіту). У мові C++ така таблиця реалізована об’єктом STL під назвою map. Стара неоптимізована версія підрахунку частот була надалі видалена, але в перших комітах на гітхабі залишилася;
3. В останньому коміті на гітхабі було перероблено механізм передачі підрахованих частот n –грам між функціями підрахунку появ n –грам у тексті та функції обчислення ентропії(раніше частоти записувалися спочатку у файл, а наступна функція зчитувала їх уже з нього, а в новому коміті частоти передаються як аргумент функції за вказівником на їх хеш-таблицю). Після цього програма стала коректно обчислювати ентропії і для n>3. Зроблено висновок про жахливу реалізацію зчитування дробових чисел з файлів функцією fwscanf, недоліки якої були помічені ще на перших етапах написання коду, але, навідміну від перших етапів, в даній ситуації fwscanf не викликала помилок при зчитуванні і явно не спотворювала зчитані дробові частоти, через що я цього й не помічав;

Скриншот виконання програми та директорії, в якій вона створила файли виводу:



Відсортований вивід частот для монограм:

Без пробілів:

о 0,115351647

е 0,085549779

а 0,077844642

и 0,069124520

н 0,066658504

т 0,065472811

с 0,052918211

л 0,047681872

р 0,044295505

в 0,042439964

к 0,033300489

м 0,031332687

д 0,028753484

п 0,028695961

у 0,026734654

я 0,020283861

ь 0,020218916

ы 0,019753175

з 0,017542297

г 0,017197166

б 0,016585765

ч 0,015616244

й 0,010635041

ж 0,009877052

ш 0,008398185

х 0,008372207

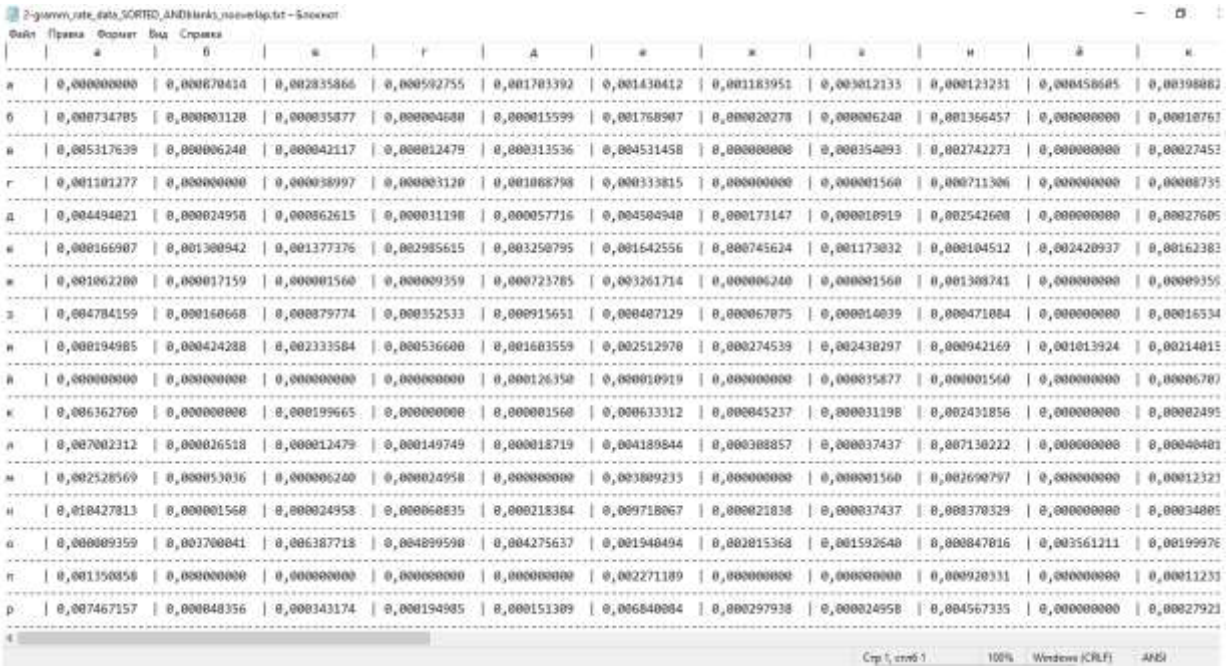
ю 0,005741048

щ 0,004286303
э 0,003910556
ц 0,003724074
ф 0,001703388

З пробілами:

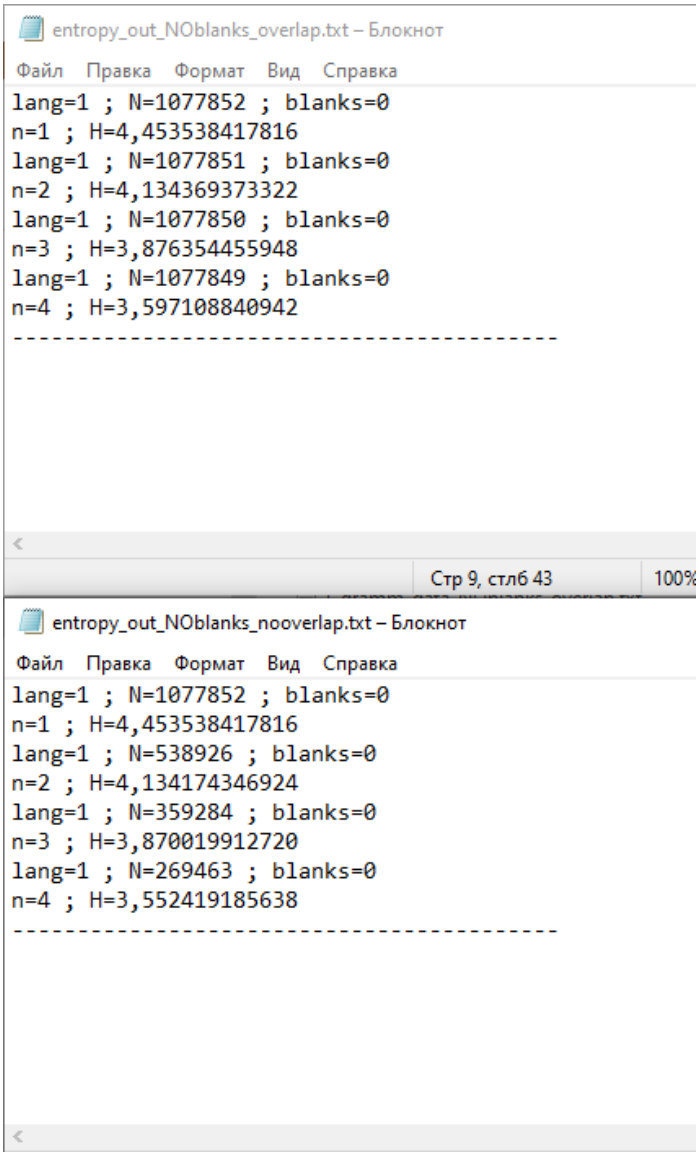
0,159338087(пробіл)
о 0,096971646
е 0,071918376
а 0,065441743
и 0,058110297
н 0,056037214
т 0,055040449
с 0,044486284
л 0,040084295
р 0,037237510
в 0,035677627
к 0,027994428
м 0,026340174
д 0,024171937
п 0,024123579
у 0,022474784
я 0,017051853
ь 0,016997257
ы 0,016605727
з 0,014747128
г 0,014456989
б 0,013943008
ч 0,013127970
й 0,008940466
ж 0,008303254
ш 0,007060027
х 0,007038189
ю 0,004826276
щ 0,003603328
э 0,003287452
ц 0,003130684
ф 0,001431972

Матриця частот біграм надто велика, щоб додати її у цей звіт, тому ознайомитися з нею можна тільки на гітхабі у відповідному txt файлі. Тут же міститься лише оглядовий скриншот файлу з матрицею:

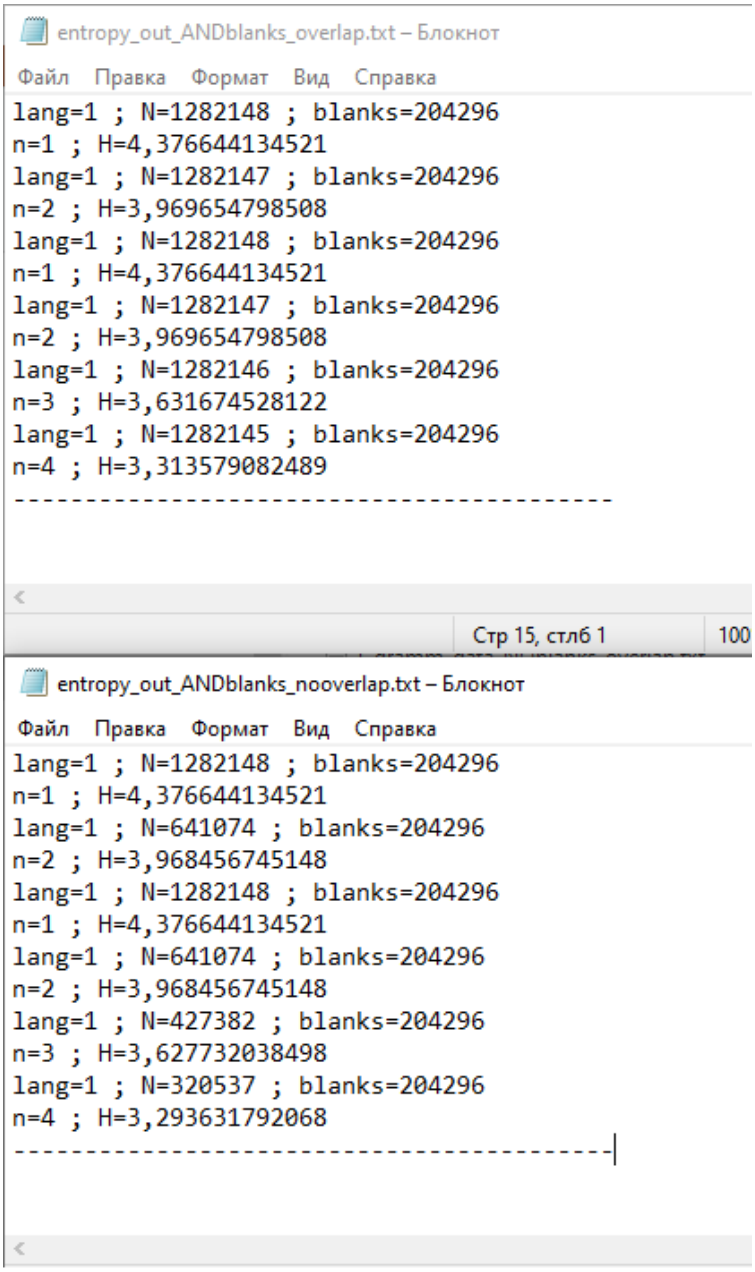


	а	б	в	г	д	е	ж	з	и	й	к
а	0,000000000	0,000670414	0,002835866	0,000592755	0,001703392	0,001430412	0,001183951	0,003012133	0,000123231	0,000458605	0,0039888
б	0,000734785	0,000003120	0,000035877	0,000004680	0,000015599	0,001768907	0,000020278	0,000006240	0,001366457	0,000000000	0,0001076
в	0,005317639	0,000006240	0,000042117	0,000012479	0,000313536	0,004531458	0,000000000	0,000354093	0,002742273	0,000000000	0,0002745
г	0,001101277	0,000000000	0,000038997	0,000003120	0,001068798	0,000333815	0,000000000	0,000001560	0,000711306	0,000000000	0,0000873
д	0,004494821	0,000024958	0,000862615	0,000031198	0,000057716	0,004504948	0,000173147	0,000018919	0,002542608	0,000000000	0,0002760
е	0,000166907	0,001300942	0,001377376	0,002985615	0,003250795	0,001642556	0,000745624	0,001173032	0,000104512	0,002420937	0,0016238
ж	0,001062280	0,000017159	0,000001560	0,000009359	0,000723785	0,003261714	0,000006240	0,000001560	0,001308741	0,000000000	0,0000935
з	0,004784159	0,000160668	0,000879774	0,000352533	0,000915651	0,000407129	0,000067075	0,000014039	0,000471084	0,000000000	0,0001653
и	0,000194985	0,000424288	0,002333584	0,000536600	0,001603559	0,002512970	0,000274539	0,002438297	0,000942169	0,001013924	0,0021401
й	0,000000000	0,000000000	0,000000000	0,000000000	0,000126350	0,000010919	0,000000000	0,000035877	0,000001560	0,000000000	0,0000670
к	0,006362760	0,000000000	0,000199665	0,000000000	0,000001560	0,000633312	0,000045237	0,000031198	0,002431856	0,000000000	0,0000249
л	0,007002312	0,000026518	0,000012479	0,000149749	0,000018719	0,004189844	0,000388857	0,000037437	0,007130222	0,000000000	0,0004048
м	0,002528569	0,000053036	0,000006240	0,000024958	0,000000000	0,003809233	0,000000000	0,000001560	0,002690797	0,000000000	0,0001232
н	0,010427813	0,000001560	0,000024958	0,000060835	0,000218384	0,009718067	0,000021838	0,000037437	0,008370329	0,000000000	0,0003480
о	0,000089359	0,001700841	0,006387718	0,004899598	0,004275637	0,001948494	0,002015368	0,001592640	0,000847016	0,003561211	0,0019997
п	0,001350858	0,000000000	0,000000000	0,000000000	0,000000000	0,002271189	0,000000000	0,000000000	0,000920331	0,000000000	0,0001123
р	0,007467157	0,000848356	0,000343174	0,000194985	0,000151389	0,006840084	0,000297938	0,000024958	0,004567335	0,000000000	0,0002792

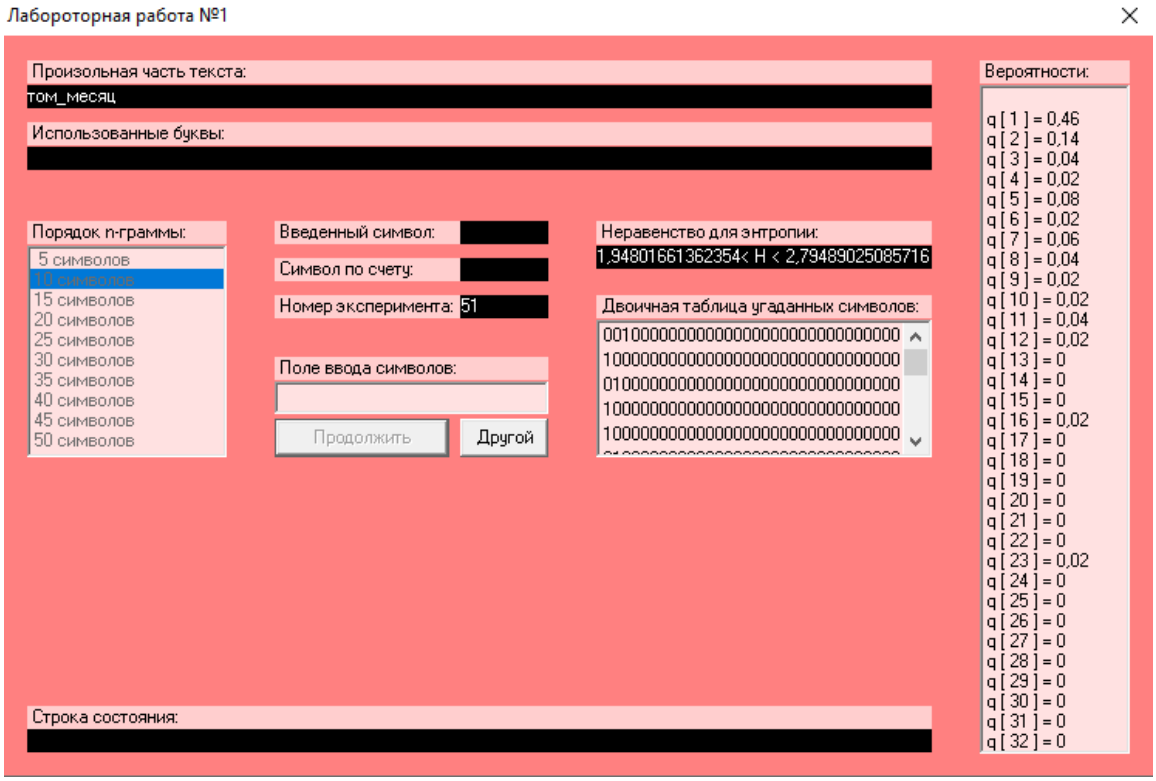
Результати виконання програми підрахунку ентропії у файлах(з перетином n-грам та без):
Без пробілів:

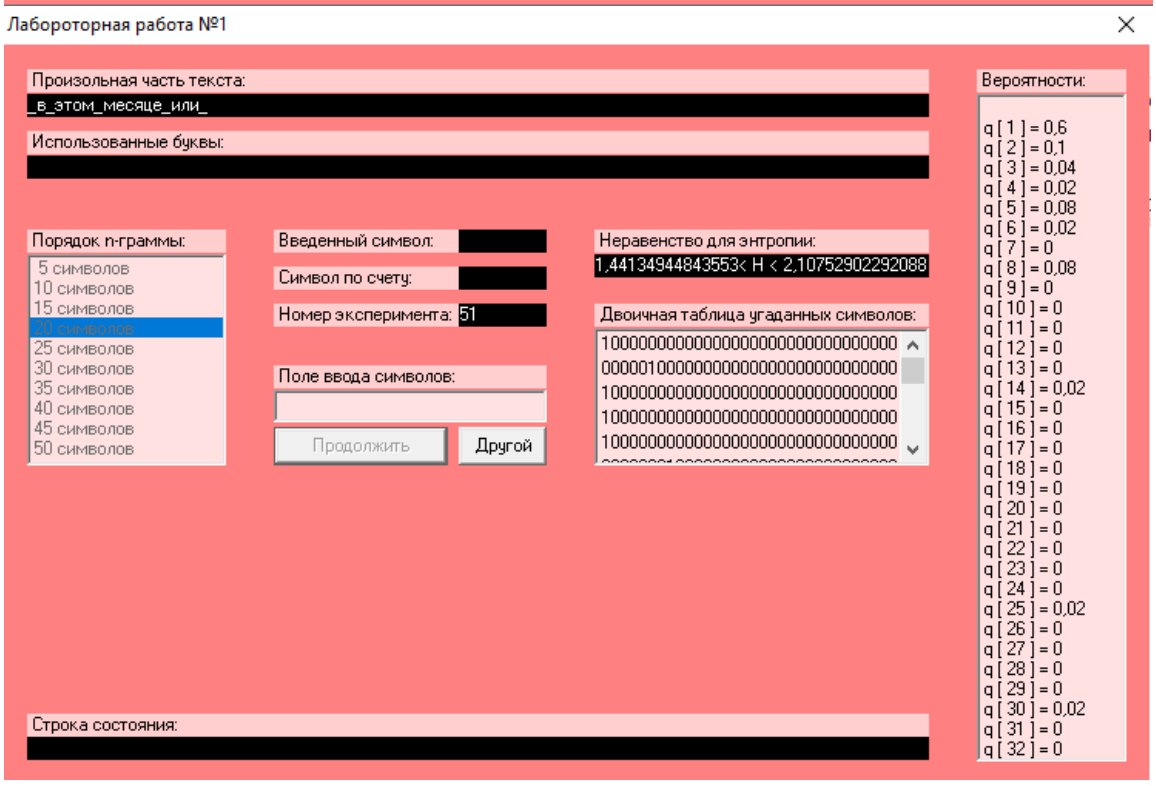

entropy_out_NOblanks_overlap.txt – Блокнот
Файл Правка Формат Вид Справка
lang=1 ; N=1077852 ; blanks=0
n=1 ; H=4,453538417816
lang=1 ; N=1077851 ; blanks=0
n=2 ; H=4,134369373322
lang=1 ; N=1077850 ; blanks=0
n=3 ; H=3,876354455948
lang=1 ; N=1077849 ; blanks=0
n=4 ; H=3,597108840942
Стр 9, стлб 43
100%
entropy_out_NOblanks_nooverlap.txt – Блокнот
Файл Правка Формат Вид Справка
lang=1 ; N=1077852 ; blanks=0
n=1 ; H=4,453538417816
lang=1 ; N=538926 ; blanks=0
n=2 ; H=4,134174346924
lang=1 ; N=359284 ; blanks=0
n=3 ; H=3,870019912720
lang=1 ; N=269463 ; blanks=0
n=4 ; H=3,552419185638

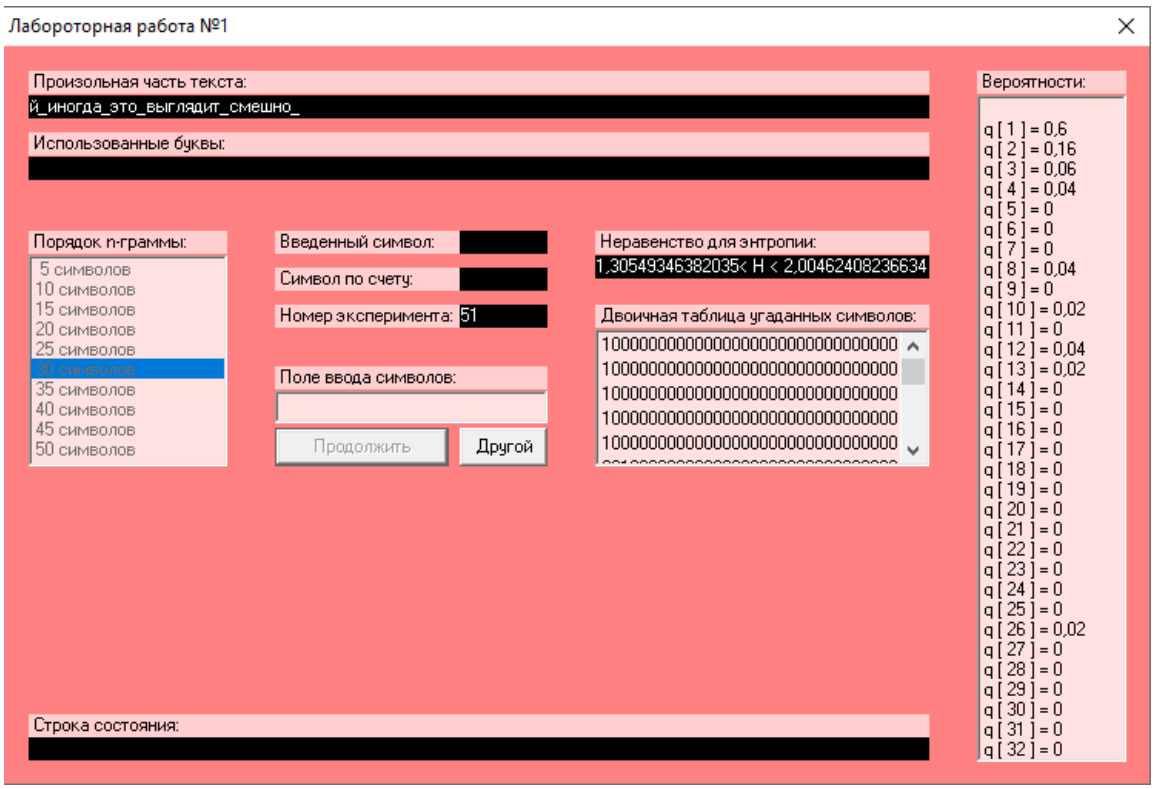
З пробілами:



Результати експериментів у CoolPinkProgram (було проведено 50 відгадувань для узятих 10, 20 та 30 послідовних символів з тексту):


Лабороторная работа №1
Произольная часть текста:
том_месяц
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: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,948016613623541 < H < 2,794890250857161
Двоичная таблица угаданных символов:
Вероятности:
q [1] = 0,46
q [2] = 0,14
q [3] = 0,04
q [4] = 0,02
q [5] = 0,08
q [6] = 0,02
q [7] = 0,06
q [8] = 0,04
q [9] = 0,02
q [10] = 0,02
q [11] = 0,04
q [12] = 0,02
q [13] = 0
q [14] = 0
q [15] = 0
q [16] = 0,02
q [17] = 0
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0
q [22] = 0
q [23] = 0,02
q [24] = 0
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0
Строка состояния:


Лабороторная работа №1
Произольная часть текста:
_в_этом_месяце_или_
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: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,441349448435531 < H < 2,107529022920881
Двоичная таблица угаданных символов:
Вероятности:
q [1] = 0,6
q [2] = 0,1
q [3] = 0,04
q [4] = 0,02
q [5] = 0,08
q [6] = 0,02
q [7] = 0
q [8] = 0,08
q [9] = 0
q [10] = 0
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0,02
q [15] = 0
q [16] = 0
q [17] = 0
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0
q [22] = 0
q [23] = 0
q [24] = 0
q [25] = 0,02
q [26] = 0
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0,02
q [31] = 0
q [32] = 0
Строка состояния:



Таблиця значень ентропії для різних моделей джерел відкритого тексту(ентропії тексту взяті з файлів виводу програми з урахуванням перетинів n –грам; з тих же файлів виводу ентропій видно, що вихідні значення, з урахуванням перетинів та без, рівні з точністю до 2 знаку після коми):

	текст з пробілами	текст без пробілів
H1	4,376644135	4,453538418
H2	3,969654799	4,134369373
H3	3,631674528	3,876354456
H4	3,313579082	3,597108841
1,948016614	< H(10) <	2,794890251
1,441349448	< H(20) <	2,107529023
1,305493464	< H(30) <	2,004624082

Підраховані надлишковості текстів:

	текст з пробілами	текст без пробілів
R1	0,124671173	0,101057338
R2	0,20606904	0,165481318
R3	0,273665094	0,217561394
R4	0,337284184	0,273926866
0,44102195	< R(10) <	0,610396677
0,578494195	< R(20) <	0,71173011
0,599075184	< R(30) <	0,738901307

Висновки: обчислені програмою значення ентропії тексту, взятого у вигляді набору n – грам, виявляються однаковими, незалежно того, узяті вони послідовно, чи з перетинами. Теоретично це обґрунтовується тим, що кількість появ n –грам, узятих з перетинами в середньому більша за середню кількість появ n –грам, узятих без перетинів, але відповідно кількість усіх розглянутих n –грам з перетинами більша, ніж без. З цього логічно випливає, що в середньому частота появ n –грам у обох моделях приблизно рівна, а отже і ентропія рівна(експериментально – до 3 знаків після коми). Також було експериментально доведено, що ентропія тексту з пробілами менша за ентропію тексту без них, з чого випливає, що надлишковість тексту з пробілами більша. Інтуїтивно це зрозуміло, адже текст з пробілами залишиться читабельним навіть після вилучення цих пробілів, а отже і кількісна міра можливого ущільнення тексту з пробілами більша.
На прикладі значень ентропій для різних моделей представлення тексту видно, що ентропія тексту 4-грам дуже сильно відрізняється від значення умовної ентропії. На основі значень H_1 … H_4 можна очікувати, що навіть H_7 не попаде в окіл приближення, отриманий в програмі CoolProgram.exe, а його обчислення буде тривати годинами. Хоч це було очікувано, але результат експерименту показує, що представлення тексту в моделі ланцюга Маркова дає змогу отримати краще приближення ентропії H_inf, проте його недолік в тому, що такий підхід неможливо автоматизувати.