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

КРИПТОГРАФІЯ
КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1
Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:
студентки 3 курсу
групи ФБ-04
Подвисоцька Ольга,
Стоян Анастасія

Перевірив:
Чорний О.М

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.

Хід роботи

1. Обрали текст, з яким будемо працювати (Deathly Hallows.txt).
2. Редагування тексту двома способами. Перший спосіб – вилучення всіх знаків пунктуації, другий – вилучення всіх пробілів. Для обох способів усі прописні літери були замінені на рядкові, а літери «ё» та »ъ» – замінені на «е» та «ь» відповідно. Відредаговані тексти містяться у файлах with_spaces.txt та without_spaces.txt.
3. Написання функцій для аналізу монограм та біграм: підрахунок кількості, частоти, ентропії та надлишквості.

Опис труднощів

Методом спроб і помилок було вирішено використати регулярні вирази замість методу replace. Спочатку не вдавалося коректно прибрати пунктуацію, бо вона замінювалася на додаткові пробіли. Рішення було знайдене у комбінуванні різних наборів для регулярних виразів. Ще одна складність була в нерозумінні того, що таке біграми з перетином і як їх рахувати. Але виявилося, що ми просто неуважно прочитали методичку.

Результати

MONOGRAMS WITH SPACES

Entropy: 4.371791807388634

Redundancy: 0.12564163852227317

MONOGRAMS WITHOUT SPACES

Entropy: 4.45294571199343

Redundancy: 0.10117697543444781

У таблицях наведено по 10 моно-/біграм, що зустрічалися в тексті найчастіше, з їхніми частотами відповідно. Повний перелік значень зберігається у файлі results.txt.

Monograms with spaces	Monograms without spaces
' ': 0.16144661738914923	'о': 0.11015714011754486
'о': 0.0923726424643047	'а': 0.08389454803885073
'а': 0.07035005704058679	'е': 0.07886674111927082
'е': 0.06613397254105882	'и': 0.06910893769670716
'и': 0.05795153347421633	'н': 0.06429659210125555
'н': 0.053916124796857944	'т': 0.05763195596015658
'т': 0.04832747161686888	'л': 0.05389240470630704
'л': 0.0451916582635067	'р': 0.05383148893927601
'р': 0.04514057714100849	'с': 0.05206154748609654
'с': 0.04365638674842169	'в': 0.041646079393549695

BIGRAMS WITH SPACES

Entropy: 3.9832061160733745

Redundancy: 0.1959934838023557

BIGRAMS WITHOUT SPACES

Entropy: 4.152079466989734

Redundancy: 0.16190655217183025

Bigrams with spaces	Bigrams without spaces
'и ': 0.01966623216180986	'то': 0.014500208692905569
'о ': 0.019482718499501485	'на': 0.01316908637630151
'а ': 0.008676223250991825	'он': 0.012539623450314168
' п': 0.016930554268016974	'по': 0.011754486897469739
'е ': 0.016188932045080037	'ал': 0.011686802711879702
' с': 0.015203255570207219	'ст': 0.010926483693751621
' н': 0.015106769005488383	'не': 0.010545196114927748
' в': 0.014669741624114831	'но': 0.01043464527846402
' о': 0.011898117755230612	'ер': 0.01001725946732546
'то': 0.011625685101906839	'ро': 0.00978036481776033

BIGRAMS WITH SPACES & WITH INTERSECTIONS

Entropy: 3.9837174281579757

Redundancy: 0.19589027592512054

BIGRAMS WITHOUT SPACES & WITH INTERSECTIONS

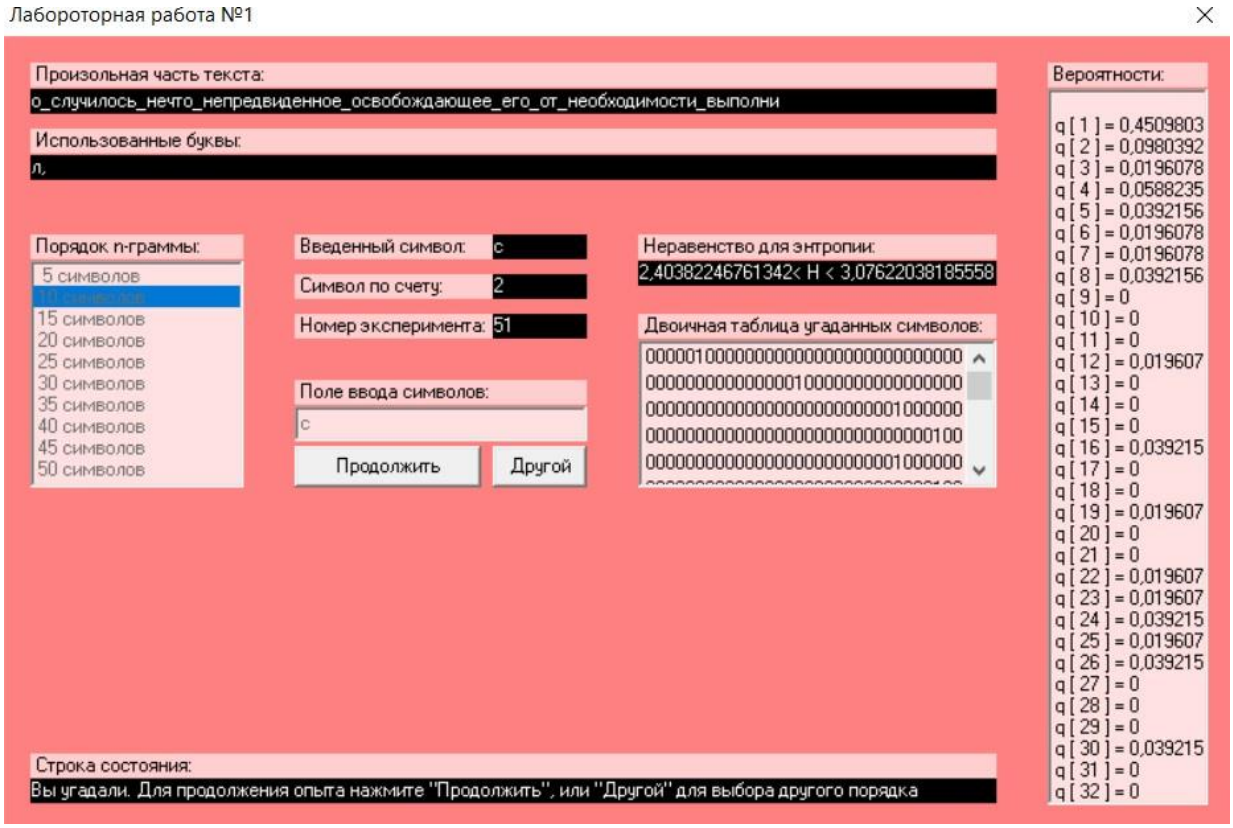
Entropy: 4.151908428712033

Redundancy: 0.16194107609195474

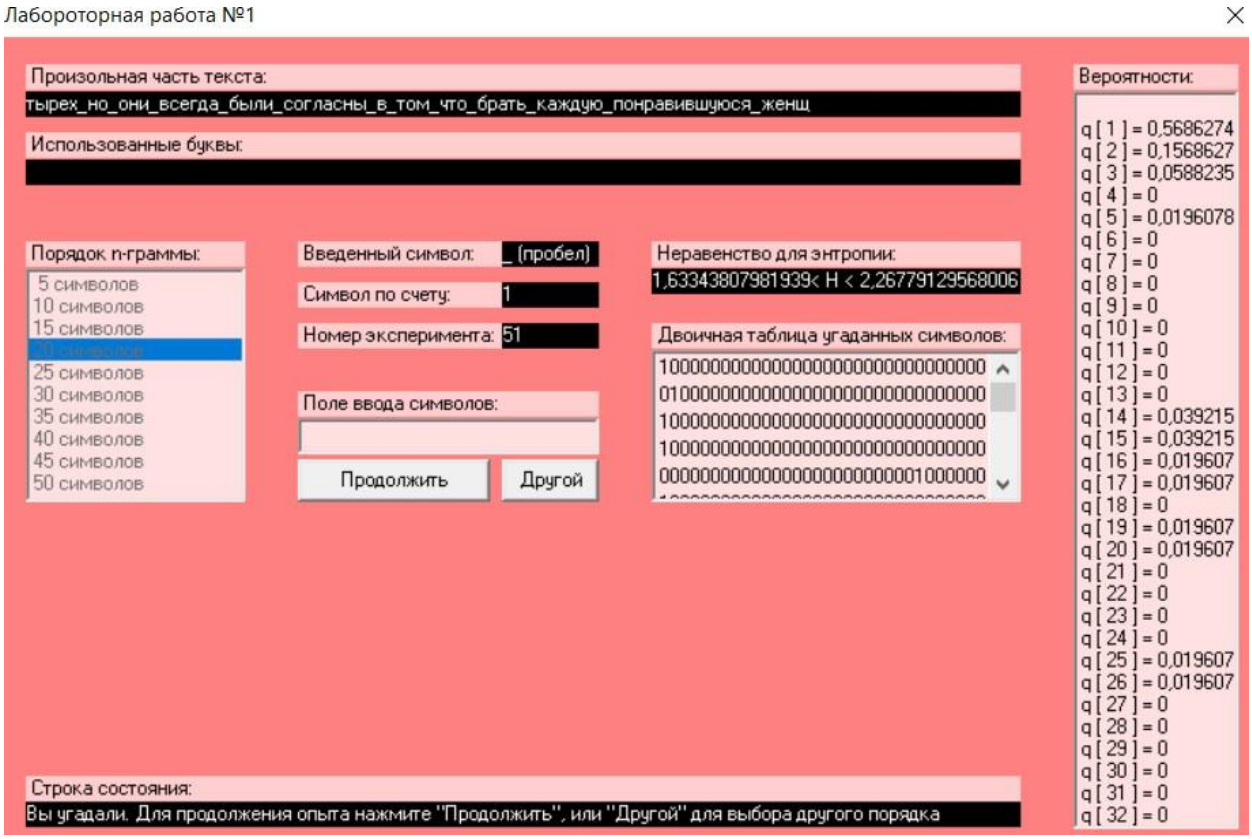
Bigrams with spaces & with intersections	Bigrams without spaces & with intersections
'и ': 0.019535691515425554	'то': 0.014406578902839352
'о ': 0.019490286073204924	'на': 0.013084481144313964
'а ': 0.017501906082626553	'он': 0.012650174286777895
' п': 0.016852040690843804	'ал': 0.011756743036989408
'е ': 0.016223932073458437	'по': 0.011680034293320698
' с': 0.015387715179228524	'ст': 0.011056211716132526
' н': 0.01520041773006843	'не': 0.010803524089929722
' в': 0.014577984792960644	'но': 0.010511354022132728
' о': 0.011794063616808338	'ер': 0.010093968210994168
'то': 0.011729739240329114	'ро': 0.009726217469288301

Cool pink program

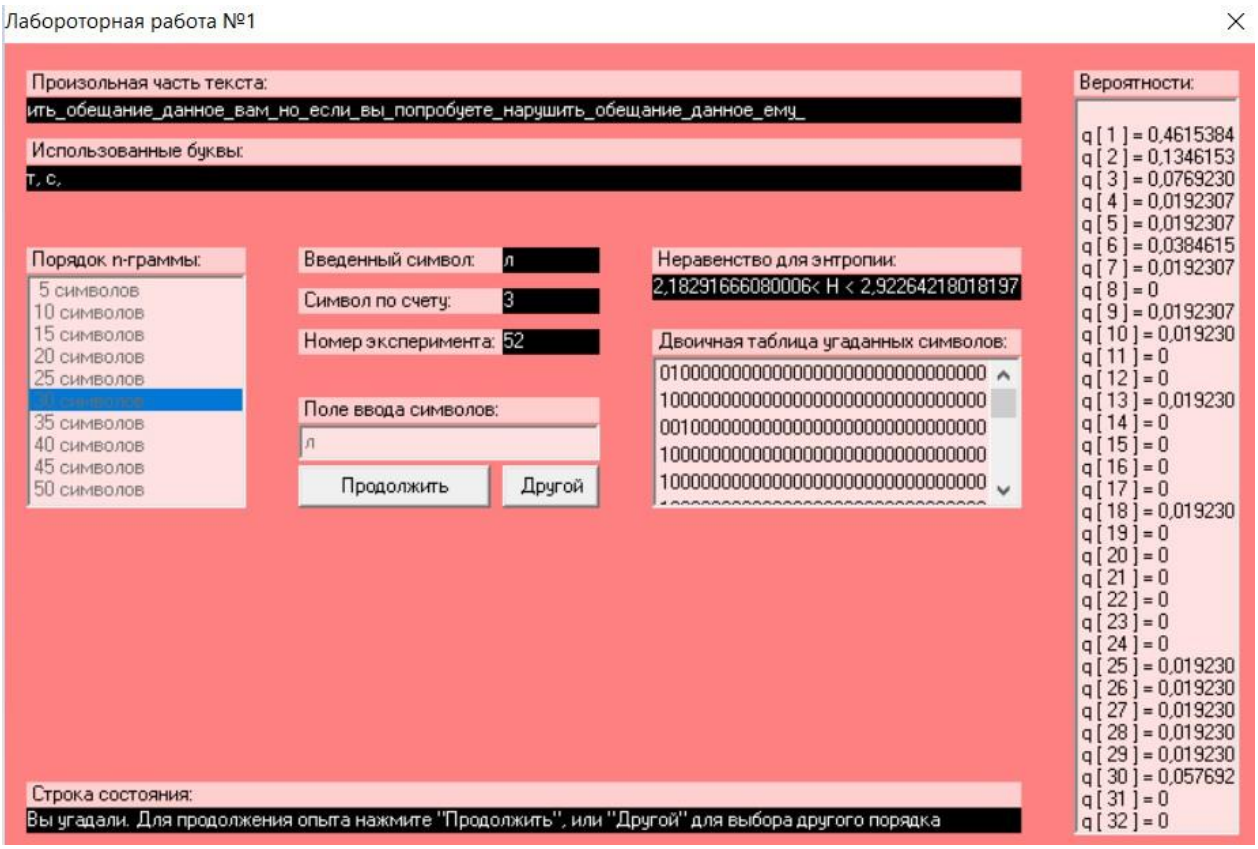
R = 0.451995715053078



R = 0.609872790612604



R = 0.489444121181743



Висновки

У ході лабораторної роботи засвоїли поняття ентропії та надлишковості, порівняли різні моделі джерела відкритого тексту, набули практичних навичок щодо оцінки ентропії на символ джерела.