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: Студент групи ФБ-05 Даниленко Данило, Студентка ФБ-05 Мірошніченко Ілона

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 1 Н та 2 Н за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1 Н та 2 Н на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1 Н та 2 Н на тому ж тексті, в якому вилучено всі пробіли.
- 2. За допомогою програми CoolPinkProgram оцінити значення (10) Н, (20) Н, (30) Н. З. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи

Перед виконанням роботи були розглянуті теоретичні відомості в методичних вказівках. В якості експериментального тексту була взята книга «Good Omens» в перекладі на російську мову. Оригінал тексту можна знайти у файлі badtxt.txt. Відредагований текст з пробілами міститься у файлі spaces.txt, а без пробілів у nospaces.txt. В ході виконання роботи було прийнято рішення пробіли замінити на «_» для кращого сприйняття. Усі таблиці також наведені у файлах з відповідними назвами. Єдина відмінність - таблиці біграм були виконанні у двох варіантах(таблиці у вигляді матриці і звичайні)

Monograms with space

h1 with space - 4.401617503940018

r1 with space - 0.13481087896098765

0.156776	к - 0.031018	ч - 0.011965
o - 0.086727	y - 0.029778	x - 0.009649
a - 0.074319	м - 0.02446	й - 0.009144
e - 0.066044	д - 0.024208	ж - 0.009027
и - 0.06327	п - 0.022487	ш - 0.008112
н - 0.055251	я - 0.018134	ю - 0.00499
т - 0.052805	ы - 0.01596	щ - 0.003901
c - 0.042807	ь - 0.015801	ц - 0.003482
p - 0.04054	б - 0.015512	э - 0.002737
л - 0.036091	з - 0.014949	ф - 0.000767
в - 0.035572	г - 0.013423	ъ-0.000152

аи - 0.010635

Monograms without space

h1	without	cnace -	1	176051	36/12	70007
$\Pi \perp$	without	Space -	4.	4/0951	304/	30097

r1 without sp	ace - 0.12742394828978676
---------------	---------------------------

	r1 without space - 0.12742394828978676			
o - 0.102851	y - 0.035314	x - 0.011443		
a - 0.088136	м - 0.029008	й - 0.010844		
e - 0.078324	д - 0.028709	ж - 0.010705		
и - 0.075033	п - 0.026668	ш - 0.00962		
н - 0.065523	я - 0.021505	ю - 0.005917		
т - 0.062623	ы - 0.018928	щ - 0.004627		
c - 0.050766	ь - 0.018739	ц - 0.004129		
p - 0.048078	6 - 0.018396	э - 0.003246		
л - 0.042801	з - 0.017728	ф - 0.00091		
в - 0.042185	г - 0.015919	ъ-0.00018		
к - 0.036785	ч - 0.014189	ë - 0.00017		
Bigrams with space				
	h2 with space - 4.277004366767287			
r2 with space - 0.15930504060131223				
_o - 0.024882	т 0.014839	00 - 0.010092		
_a - 0.021953	a 0.010715	_p - 0.009965		
_e - 0.019365	o 0.010699	л 0.008965		
н 0.016837	и 0.010652	e 0.008937		
0.015917	_т - 0.010442	c 0.008747		
Bigrams without space				
	h2 without space - 4.3809269041590895			
	r2 without space - 0.13152564995927474			
00 - 0.013896	ои - 0.010589	ao - 0.008691		
иа - 0.012311	ио - 0.009931	ee - 0.008605		
aa - 0.010726	oa - 0.009891	ae - 0.008512		
0.040605	0.0004	0.000400		

eo - 0.0094

ea - 0.008499

Bigrams with spaces without intersection

h2 with space without intersection - 4.002665513513063

r2 with space without intersection - 0.21322953338184325

и 0.021899	_в - 0.015072	я 0.011541
o 0.019434	a 0.014843	_и - 0.011321
e 0.017351	_н - 0.014016	_o - 0.01036
_п - 0.015506	то - 0.011633	ь 0.010237
c - 0.015074	но - 0.011563	на - 0.009907

Bigrams without space without intersection

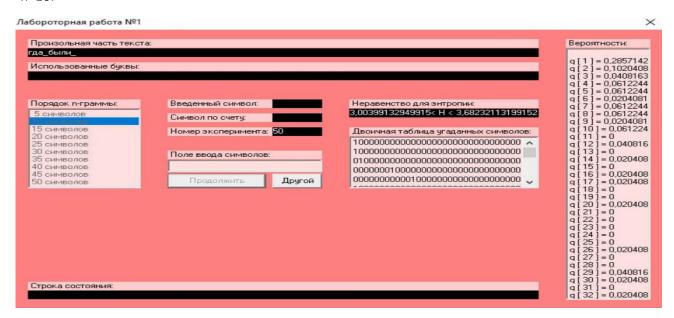
h2 without space without intersection - 4.162934395360695

r2 without space without intersection - 0.1747404550756756

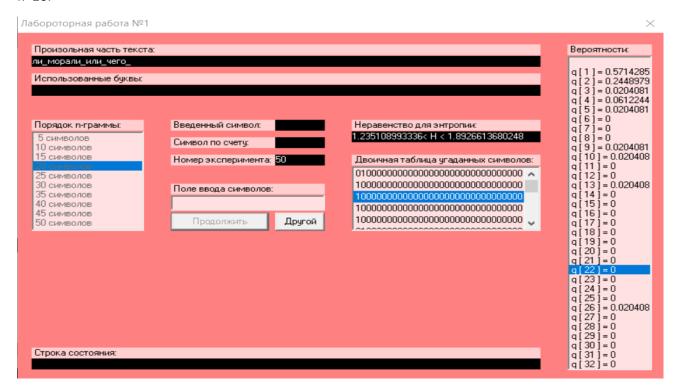
то - 0.014266	по - 0.010293	oc - 0.00934
но - 0.013924	не - 0.010097	от - 0.009156
на - 0.011785	ен - 0.009536	ов - 0.009003
ст - 0.011472	ка - 0.00944	он - 0.008692
pa - 0.010736	ни - 0.00939	ак - 0.008522

Результати експериментів у CoolPinkProgram

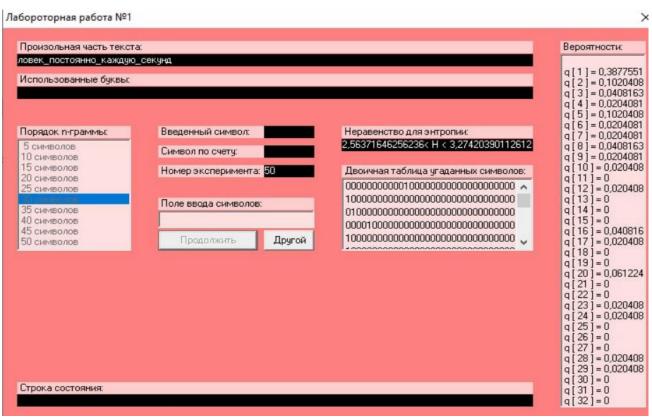
n=10:



n=20:



n=30:



Висновки

Під час виконання даної лабораторної роботи, ми здобули навички з аналізу тексту. Використовуючи regex нам вдалося в досить простій формі отримати працююче рішення. Крім того, нами було засвоєно визначення ентропії та надлишковості, які ми інтегрували в наш розв'язок. В результаті ми перевірили статистику тексту за допомогою CoolPinkProgram. На нашу думку, дані навички стануть у нагоді у подальшому розвитку.