

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

Київ – 2022

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. 3.

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
о - 0.086727	у - 0.029778	х - 0.009649
а - 0.074319	м - 0.02446	й - 0.009144
е - 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
с - 0.042807	ь - 0.015801	ц - 0.003482
р - 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

ë - 0.000143

Monograms without space

h1 without space - 4.476951364230097

r1 without space - 0.12742394828978676

о - 0.102851	у - 0.035314	х - 0.011443
а - 0.088136	м - 0.029008	й - 0.010844
е - 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
с - 0.050766	ь - 0.018739	ц - 0.004129
р - 0.048078	б - 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	ё - 0.00017

Bigrams with space

h2 with space - 4.277004366767287

r2 with space - 0.15930504060131223

о - 0.024882	т - 0.014839	оо - 0.010092
а - 0.021953	а - 0.010715	_р - 0.009965
е - 0.019365	о - 0.010699	л_ - 0.008965
н_ - 0.016837	и_ - 0.010652	е_ - 0.008937
__ - 0.015917	_т - 0.010442	с_ - 0.008747

Bigrams without space

h2 without space - 4.3809269041590895

r2 without space - 0.13152564995927474

оо - 0.013896	ои - 0.010589	ао - 0.008691
иа - 0.012311	ио - 0.009931	ее - 0.008605
аа - 0.010726	оа - 0.009891	ае - 0.008512
аи - 0.010635	ео - 0.0094	еа - 0.008499

oe - 0.008485

OH - 0.007465

OT - 0.007064

Bigrams with spaces without intersection

h2 with space without intersection - 4.002665513513063

r2 with space without intersection - 0.21322953338184325

$$u_- = 0.021899$$

B - 0.015072

$$r_{\text{я}} = 0.011541$$
$$o_- = 0.019434$$

a_ = 0.014843

$$_H = 0.011321$$
$$e_- = 0.017351$$

$\alpha_H = 0.014016$

$$\rho = 0.01036$$

$\pi = 0.015506$

TO - 0.011633

$$b_- = 0.010237$$

_c - 0.015074

HO - 0.011563

на - 0.009907

Bigrams without space without intersection

h2 without space without intersection - 4.162934395360695

r2 without space without intersection - 0.1747404550756756

TO - 0.014266

по - 0.010293

oc - 0.00934

HO - 0.013924

He - 0.010097

от - 0.009156

на - 0.011785

ен - 0.009536

OB - 0.009003

CT - 0.011472

ка - 0.00944

OH - 0.008692

pa - 0.010736

ни - 0.00939

ак - 0.008522

Результати експериментів у CoolPinkProgram

n=10:

Лабораторная работа №1

Произвольная часть текста:	
гда_были__	
Использованные буквы:	
Порядок n-граммы:	
5 символов	
15 символов	
20 символов	
25 символов	
30 символов	
35 символов	
40 символов	
45 символов	
50 символов	
Введенный символ:	
Символ по счету:	
Номер эксперимента: 50	
Поле ввода символов:	
Продолжить	Другой
Неравенство для энтропии:	
3,00399132949915 < H < 3,68232113199152	
Двоичная таблица угаданных символов:	
10000000000000000000000000000000 ^	
10000000000000000000000000000000	
01000000000000000000000000000000	
00000001000000000000000000000000	
00000000001000000000000000000000 v	
Вероятности:	
q[1] = 0,2857142	
q[2] = 0,1020408	
q[3] = 0,0408163	
q[4] = 0,0612244	
q[5] = 0,0612244	
q[6] = 0,0204081	
q[7] = 0,0612244	
q[8] = 0,0612244	
q[9] = 0,0204081	
q[10] = 0,061224	
q[11] = 0	
q[12] = 0,040816	
q[13] = 0	
q[14] = 0,020408	
q[15] = 0	
q[16] = 0,020408	
q[17] = 0,020408	
q[18] = 0	
q[19] = 0	
q[20] = 0,020408	
q[21] = 0	
q[22] = 0	
q[23] = 0	
q[24] = 0	
q[25] = 0	
q[26] = 0,020408	
q[27] = 0	
q[28] = 0	
q[29] = 0,040816	
q[30] = 0,020408	
q[31] = 0	
q[32] = 0,020408	
Строка состояния:	

Лабораторная работа №1

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.