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

Лабораторна робота № 1 «Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали:

Студенти 3 курсу

групи ФБ-92

Сидоренко Андрій

Варгіч Дмитро

Перевірила:

Селюх П.В.

#### 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  $H_1$  та  $H_2$  за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення  $H_1$  та  $H_2$  на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1 Мб), 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.

#### Виконання

1. код програми міститься у файлі base.py;

текст, що аналізувався - у файлі raw\_text.txt;

результати аналізу: Char\_spaces.csv, Char\_no\_spaces.csv, Bigrams\_spaces\_step1.csv, Bigrams\_spaces\_step2.csv, Bigrams\_no\_spaces\_step1.csv, Bigrams\_no\_spaces\_step2.csv, H\_and\_R.txt

(також результати перших 6-ти файлів представлені разом у файлі **Результати.xlsx**)

## Результати

Н<sub>1</sub> з пробілами=4.389834609607416

Н<sub>1</sub> без пробілів=4.449759325578299

H<sub>2</sub> з пробілами (перехресні біграми) =3.9947022441300635

H<sub>2</sub> з пробілами (НЕперехресні біграми) = 3.9952480901955263

H<sub>2</sub> без пробілів (перехресні біграми) =4.131904131896036

 $H_2$  без пробілів (НЕперехресні біграми) = 4.130475636556303

## Частота букв з пробілами:

Літера	Частота	
	0.15117749875452802	
0	0.09643246164368524	
e	0.07258695083129976	
a	0.06301834854638644	
И	0.06019471582129377	
Н	0.054081348318401744	
Т	0.0516056033572858	
С	0.0470864399766949	
р	0.04386763377213351	
Л	0.04218899087216813	
В	0.03655523562640907	
К	0.02914995482525395	
Д	0.025853464945241454	
П	0.02533501085038293	
M	0.024176510820829358	
У	0.023560107743880298	
Ы	0.017196801459102077	
Ь	0.016832025939592498	
б	0.0163676126624391	
Γ	0.015773163667682746	
Я	0.015614418765673948	
3	0.01378716361701948	
Ч	0.01111045436506261	
й	0.009479097180589214	
ж	0.008835673694787594	
X	0.0078105869339435445	
Ш	0.006682484864349103	
ю	0.004259091945384999	
щ	0.0029857551782080404	
ц	0.0023592194479392717	
ф	0.002127857197139214	
Э	0.0019066275996588674	

## Частота букв без пробілів:

Літера	Частота	
0	0.11360733428035955	
е	0.08551487587192068	
а	0.074242080592727	
И	0.07091555152339911	
Н	0.06371337734219877	
Т	0.060796695753900516	
С	0.055472657602272865	
р	0.05168057362730937	
Л	0.04970296005379746	
В	0.043065818322171955	
К	0.03434163771870138	
Д	0.030458034403313797	
П	0.029847242283526917	
M	0.028482410380680664	
У	0.027756224309924277	
Ы	0.020259596598663818	
Ь	0.01982985360884634	
б	0.019282727117180567	
Γ	0.01858240520784838	
Я	0.018395387425242625	
3	0.016242693374397665	
Ч	0.013089255231524035	
й	0.01116734908261809	
ж	0.010409330197801148	
X	0.009201672814378881	
ш	0.007872652827350755	
ю	0.005017647316294819	
щ	0.0035175259536912138	
ц	0.0027794025776621185	
ф	0.0025068341072686253	
Э	0.0022462029421478394	

## Частота біграм з пробілами (крок 1):

Біграма	Частота	
('o', ' ')	0.018640736037180146	
('и', ' ')	0.01848367964549164	
(' ' <i>,</i> 'п')	0.017060039449863548	
('a', ' ')	0.015031816585154963	
('e', ' ')	0.015004796130670918	
(' ', 'c')	0.014955821556918587	
(' ', 'H')	0.014503228944310843	
(' ', 'B')	0.013773676673241643	
('c', 'т')	0.012032546137426032	
(' ', 'o')	0.01158839741684455	

# Частота біграм без пробілів (крок 1):

Біграма	Частота	
('c', 't')	0.014400397910967422	
('t', 'o')	0.013509077343944293	
('н', 'a')	0.012215866699825915	
('н', 'о')	0.01177816463566277	
('л', 'и')	0.011754289977617509	
('e', 'p')	0.01141407610047252	
('p', 'o')	0.01098234270082069	
('o', 'c')	0.010827157423526485	
('e', 'н')	0.010741606565530963	
('o', 'в')	0.010397413578711764	

## Частота біграм з пробілами (крок 2):

Біграма	Частота	
('o', ' ')	0.01862047069631711	
('и', ' ')	0.01820503120862493	
(' ' <i>,</i> 'п')	0.017279580642546408	
('a', ' ')	0.015293577237969142	
('e', ' ')	0.014992974681834149	
(' ', 'c')	0.014668729228025616	
(' ' <i>,</i> 'H')	0.01420938150179686	
('', 'B')	0.013902023832040854	
('c', 'т')	0.011926153097895107	
(' ', 'o')	0.011686346564349212	

## Частота біграм без пробілів (крок 2):

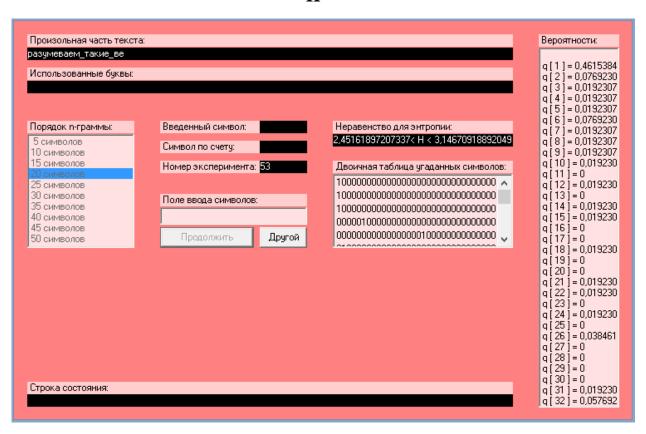
Біграма	Частота	
('c', 'T')	0.014826133148703012	
('t', 'o')	0.013580674298583837	
('н', 'a')	0.012124323055313493	
('л', 'и')	0.011921388865677462	
('н', 'о')	0.0115672488092538	
('e', 'p')	0.011376251924890475	
('p', 'o') 0.010966404443860843		
('o', 'c')	0.010795303068285366	
('е', 'н')	0.01077142845773995	
('н', 'e')	0.01055655696283121	

<sup>\*</sup>Наведені приклади 10 найуживаніших біграм

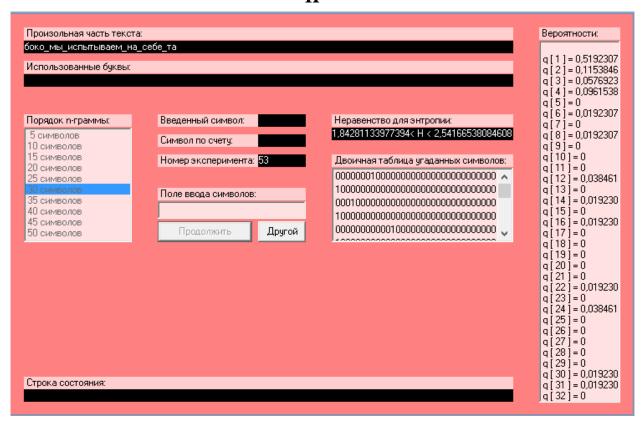
 $H^{(10)}$ 



#### $H^{(20)}$



### $H^{(30)}$



Н	R
$H_1$ з пробілами	0.12203307807851682
$H_1$ без пробілів	0.10182014462183964
$H_2$ з пробілами (крок 1)	0.20105955117398733
$H_2$ з пробілами (крок 2)	0.20095038196089476
$H_2$ без пробілів (крок 1)	0.16597892513198087
$H_2$ без пробілів (крок 2)	0.16626726561149263
$H^{(10)}$	0.414 < R < 0.546
$H^{(20)}$	0.37 < R < 0.51
$H^{(30)}$	0.492 < R < 0.632

#### Висновки

ПІд час виконання лабораторної роботи ми засвоїли поняття ентропії на символ джерела та його надлишковості, вивчили та порівняли різні моделі джерела відкритого тексту для наближеного визначення ентропії, набули практичних навичок щодо оцінки ентропії на символ джерела.

У результаті виконання лабораторної роботи, найуживанішими виявилися:

- символи пробілу, "о", "е", "а" для випадку з пробілом;
- символи "о", "е", "а" для випадку без пробілів;
- біграми ('о '),('и '),(' п') для випадку з пробілом;
- біграми ('cт'),('то'),('на') для випадку без пробілів

У російській мові можливе ущільнення тексту деякою схемою кодування символів без втрати його змісту, причому російська мова більш надлишкова, ніж англійська.