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

КРИПТОГРАФІЯ КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав: Андреєв Д.Ю. Група: ФБ-06

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

- 1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку Н1 та Н2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення Н1 та Н2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 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.

Хід роботи

В роботі було використано випадково згенерований текст. Його версії з пробілами та без можна знайти відповідно у файлах sample_text.txt та sample_text_no_spaces.txt.

Було реалізовано програму підрахунку частот символів та біграм у заданому тексті мовою Python 3. Вихідний код програми знаходиться у файлі таіп.ру.

У програмі реалізовано функції char_count та calc_h1 для підрахунку частот символів та визначення ентропії відповідно.

Виникли деякі труднощі з написанням функцій count_bigrams та calc_h2, отримані значення не співпадали з очікуваними. Тому довелося частково переписати дані функції з використанням вбудованої функції count та попереднім розбиттям тексту на біграми. В результаті були отримані коректні значення h2.

Нижче наведено таблиці, що містять значення частот для символів, підрахованих програмою для заданого тексту:

Текст з пробілами

character	frequency		
character			
	0.143		
0	0.093		
a	0.074		
е	0.074		
И	0.063		
н	0.059		
Т	0.048		
С	0.045		
В	0.043		
р	0.041		
л	0.037		
κ	0.028		
м	0.026		
п	0.023		
у	0.021		
я	0.02		
д	0.02		
3	0.017		
ы	0.016		
6	0.015		
г	0.015		
ь	0.014		
й	0.012		
ч	0.011		
ш	0.01		
x	0.008		
ж	0.008		
ю	0.004		
ц	0.003		
щ	0.003		
э	0.002		
Ф	0.002		

Текст без пробілів

character	frequency
0	0.109
a	0.086
e	0.086
И	0.074
н	0.069
т	0.056
С	0.052
В	0.05
p	0.048
л	0.043
κ	0.033
М	0.03
п	0.027
у	0.025
я	0.024
д	0.023
3	0.019
ы	0.019
б	0.018
Г	0.017
ь	0.017
й	0.014
ч	0.013
ш	0.011
x	0.01
ж	0.009
ю	0.005
ц	0.004
щ	0.003
э	0.003
Ф	0.003

Найчастіші букви - о, а, е, и, н, т, с, в, р, л.

Нижче наведено таблицю найчастіших біграм у тексті для різних моделей джерела. Повні таблиці частот біграм розташовані у відповідних файлах (slide bigrams with spaces.csv, block bigrams with spaces.csv...).

Перехресні з пробілами		1 -	Не перехресні з пробілами		Перехресні без пробілів			Не перехресні без пробілів	
bigram	frequency	bigram	frequency	t	olgram	frequency	bigram	frequency	
e""	0.0181	e""	0.0182	c	т	0.0164	ст	0.0163	
""п	0.0170	""п	0.0170	p	oa	0.0142	pa	0.0142	
и""	0.0157	и""	0.0158	c	В	0.0138	ов	0.0138	
a""	0.0147	a""	0.0147	e	ен	0.0133	ен	0.0136	
""c	0.0143	""c	0.0142	н	10	0.0120	но	0.0118	
ст	0.0139	ст	0.0139	н	4a	0.0117	на	0.0116	
""В	0.0131	""В	0.0133	т	го	0.0112	то	0.0113	
0""	0.0127	o""	0.0126	н	ни	0.0110	ни	0.0108	
""o	0.0121	pa	0.0123	г	10	0.0102	по	0.0101	
pa	0.0121	я""	0.0121	a	эл	0.0099	ал	0.0100	

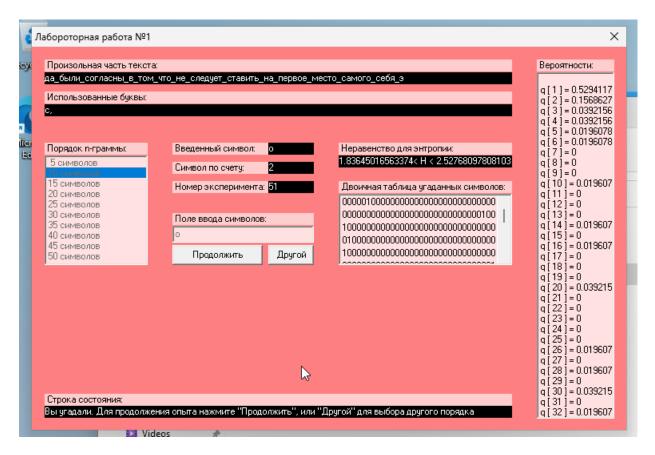
Тепер можемо обрахувати значення надлишковості R для кожного випадку.

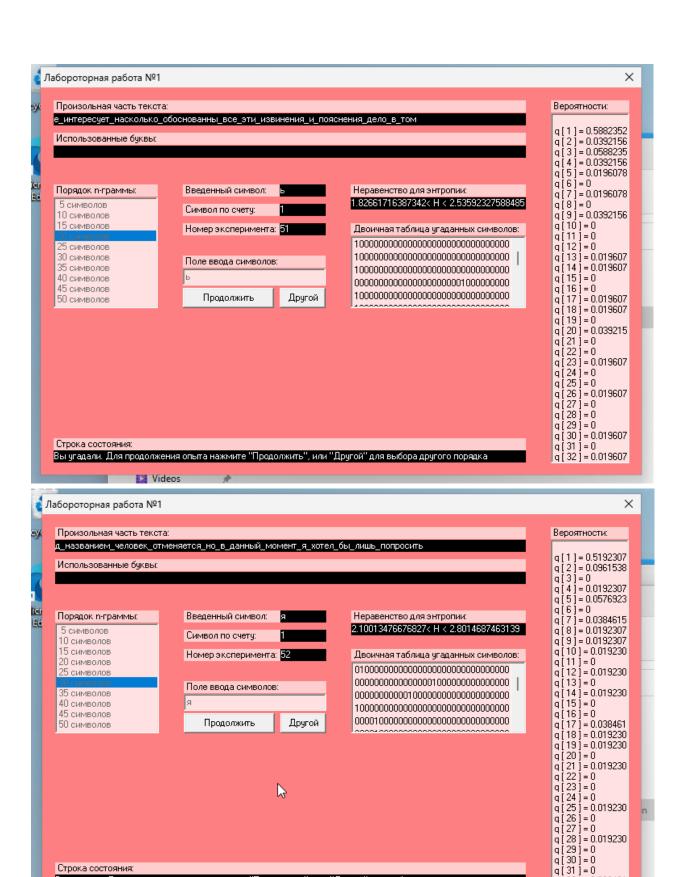
	H1 для тексту з пробіла ми	H1 для тексту без пробілі в	Н2 для тексту з пробілами (перехресн о)	Н2 для тексту з пробілами (НЕ перехресно)	H2 для тексту без пробілів (перехресно)	Н2 для тексту без пробілів (НЕ перехресно)
Н	4.395	4.452	3.952	3.951	4.078	4.079
R	0.121	0.1096	0.2096	0.2098	0.1844	0.1842

CoolPinkProgram

Далі починаємо працювати з програмою CoolPinkProgram. При використанні виникла низка труднощів. По перше, програма не запускалась на macOS, з під якої йшла робота, навіть за допомогою сторонніх утиліт. Шлях вирішення - встановлення Windows 11 на віртуальну машину. Далі в програмі замість тексту з'являлися ієрогліфи. Шлях вирішення - зміна мови системи. Далі Windows не дозволяла запуск програми, бо ідентифікувала її як шкідливе ПО. Шлях вирішення - вимкнення перевірки файлів у налаштуваннях. З точки зору безпеки це не найкращий хід, але після закінчення роботи перевірку було увімкнено назад.

Тепер, коли всі труднощі подолано, починаємо працювати. Визначаємо значення H(10), H(20), H(30).





Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

q[32]=0.038461

Строка состояния:

IVIUSIC

Отримані результати:

1.8364 < H(10) < 2.5277 1.8266 < H(20) < 2.5359 2.1001 < H(30) < 2.8014

Оцінка надлишковості російської мови

Використовуючи результати, отримані за допомогою CoolPinkProgram, можемо порахувати значення надлишковості мови:

	Н	R
H(10)	2.182	0.5636
H(20)	2.181	0.5638
H(30)	2.451	0.5098

Висновки

В ході виконання лабораторної роботи було засвоєно поняття ентропії та надлишковості, написано програму мовою Python 3 для підрахунку частот символів та біграм у тексті перехресно та не перехресно, а також обчислення значень H1 та H2 за отриманими частотами для різних моделей джерела. Також за допомогою програми CoolPinkProgram було обраховано значення ентропій надлишковість мови яка становить 0.5636 для H(10), 0.5638 для H(20) та 0.5098 для H(30).