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

ЛАБОРАТОРНА РОБОТА №1

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконав: студенти групи ФБ-04 Ляденко Максим Петриковець Віталій Перевірив: Чорний О.

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

Хід роботи

Спочатку знайшли текст. Далі за допомогою регулярних виразів відфільтрували потрібні символи від інших. За допомогою словника руthon порахували частоти монограм і біграм з вказаними вимогами. Далі підключили до роботи бібліотеку pandas, щоб відсортувати біграми і монограми за спаданням величин частот та зберегти в ексель. Наостанок порахували ентропію та надлишковість для кожного випадку. Особливих проблем не виникло, якщо не рахувати технічні помилки в використанні рапdas при збереженні в ексель.

Результати

Монограми з пробілами

ентропія 4.407131655643124 надлишковість 0.1185736688713751

Монограми без пробілів

ентропія 4.476594054485446

надлишковість 0.09640357910325381

Біграми з пробілами (з перетинами)

ентропія 4.007893101584284

надлишковість 0.1984213796831431

Біграми з пробілами (без перетинів)

ентропія 4.007966872024695

надлишковість 0.19840662559506106

Біграми без пробілів (з перетинами)

ентропія 4.163370899401399

надлишковість 0.159627386853332

Біграми без пробілів (без перетинів)

ентропія 4.162466357700262

надлишковість 0.1598099677694821

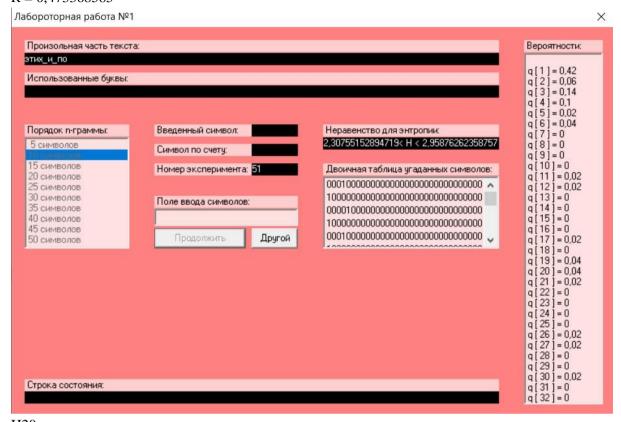
Таблиці

Монограми з пробіла	МИ	Монограми без пробілів			
	0,153914	'o'	0,112398		
'o'	0,095098	'e'	0,08464		
'e'	0,071613	ʻa'	0,077791		
ʻa'	0,065818	'и'	0,066571		
'и'	0,056325	'н'	0,066508		
'н'	0,056272	' †'	0,058236		
' †'	0,049273	'c'	0,050959		
'c'	0,043116	'л'	0,049293		
'л'	0,041706	'р'	0,04426		
'p'	0,037448	'в'	0,041753		
'в'	0,035326	'к'	0,036865		
'к'	0,031191	'M'	0,036617		
'm'	0,030981	'A'	0,029633		
'A'	0,025072		0,027397		
'y'	0,023181	'п'	0,026751		
'п'	0,022634		0,022653		
'я'	0,019166				
'ы'	0,017373	'ь'	0,020534		
'ь'	0,015876	· · · · ·	0,018764		
'r'	0,015042	·6'	0,017778		
'6'	0,014601		0,017257		
'3'	0,013899		0,016427		
' 4'	0,013011		0,015378		
'й'	0,012743	"й"	0,015061		
'ж'	0,00812	′ж′	0,009597		
ʻx'	0,007618	'x'	0,009004		
'ш'	0,007071	'ш'	0,008357		
'ю'	0,005875	'ю'	0,006944		
'ц'	0,003371	'ц'	0,003984		
'щ'	0,003313	'щ'	0,003916		
'a'	0,002254	'ə'	0,002664		
'ф'	0,0017	'ф'	0,002009		

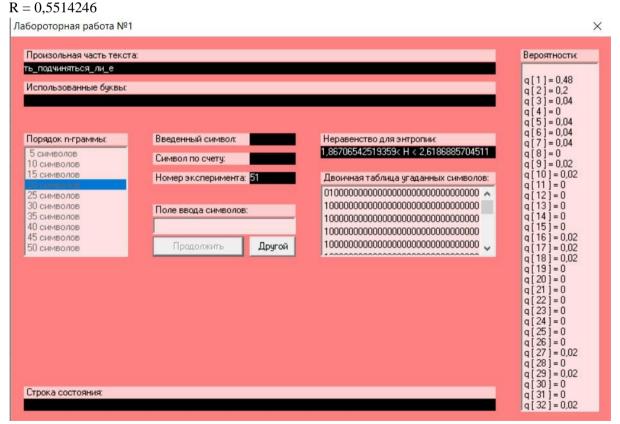
Перехре	есні	Неперехресні		Перехресні		Неперехресні	
біграми	іграми з біграми з		біграми без		біграми без		
пробілами		пробілами		пробілів		пробілів	
'o '	0,017815	'o '	0,017662	'то'	0,014357	'то'	0,014486
'e '	0,016913	'e '	0,016928	'но'	0,013452	'но'	0,013568
'и '	0,015907	'п'	0,015826	'ст'	0,012807	'ст'	0,012945
'a '	0,015771	'a '	0,0158	'на'	0,012062	'на'	0,012186
'п'	0,015676	'и '	0,015797	'ен'	0,011441	'ен'	0,011717
' c'	0,014109	' c'	0,01409	'ко'	0,011264	'ко'	0,011445
'в'	0,013238	'в'	0,013379	'не'	0,010045	'не'	0,010159
'н'	0,013211	'н'	0,013316	'по'	0,010036	'по'	0,009995
'я '	0,012716	'я '	0,012688	'ов'	0,009628	'ов'	0,00972
'то'	0,011862	'то'	0,011957	ʻoc'	0,009351	ʻoc'	0,009512

Cool Pink Program

H10 R = 0.473368585



H20



H30 R = 0,613354718 Лабороторная работа №1 X Произольная часть текста: Вероятности: днако_они_всегда_были_согласн q[1] = 0.52Использованные буквы: q[2] = 0,2 q[3] = 0,06 q[4] = 0 q[5] = 0,02 q[6] = 0,06 q[7] = 0,02 q[8] = 0 q[10] = 0 q[11] = 0,02 q[12] = 0 q[13] = 0,04 q[14] = 0 q[15] = 0 q[16] = 0 q[17] = 0,04 q[18] = 0 q[20] = 0 q[20] = 0 q[20] = 0 q[21] = 0 q[22] = 0 q[26] = 0 q[27] = 0 q[28] = 0 q[28] = 0 q[28] = 0 q[28] = 0 q[29] = 0 q[28] = 0 q[28] = 0 q[30] = 0 q[31] = 0,02 Порядок п-граммы: Введенный символ: Неравенство для энтропии: 1,60140639774841< H < 2,26504641727735 5 символов Символ по счету: 10 символов 15 символов Номер эксперимента: 51 Двоичная таблица угаданных символов: 20 символов 1000000000000000000000000000000000000 5 символов Поле ввода символов: 5 символов 40 символов 45 символов Продолжить Другой 50 символов Строка состояния:

Висновки:

В роботі ми оцінили частоту появи окремих монограм, біграм в російському тексті. Найчастіші монограми в тексті з пробілами: "", "о", "е", "а", без пробілів: "о", "е", "а", "и". Найрідші з пробілами: "э", "ф", без пробілів - теж саме. Найчастіші перехресні біграми в тексті з пробілами: "о ", "е ", "и "; неперехресні: "о ", "е ", " п"; перехресні без пробілів: "то", "но", "ст"; неперехресні - аналогічно. При шифруванні краще прибирати пробіли, щоб ускладнити взлом шифру.

q[32]=0