

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 1Мб), 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 python 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 pandas 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 пробілами		Монограми без пробілів	
' '	0,153914	'о'	0,112398
'о'	0,095098	'е'	0,08464
'е'	0,071613	'а'	0,077791
'а'	0,065818	'и'	0,066571
'и'	0,056325	'н'	0,066508
'н'	0,056272	'т'	0,058236
'т'	0,049273	'с'	0,050959
'с'	0,043116	'л'	0,049293
'л'	0,041706	'р'	0,04426
'р'	0,037448	'в'	0,041753
'в'	0,035326	'к'	0,036865
'к'	0,031191	'м'	0,036617
'м'	0,030981	'д'	0,029633
'д'	0,025072	'у'	0,027397
'у'	0,023181	'п'	0,026751
'п'	0,022634	'я'	0,022653
'я'	0,019166	'ы'	0,020534
'ы'	0,017373	'ь'	0,018764
'ь'	0,015876	'г'	0,017778
'г'	0,015042	'б'	0,017257
'б'	0,014601	'з'	0,016427
'з'	0,013899	'ч'	0,015378
'ч'	0,013011	'й'	0,015061
'й'	0,012743	'ж'	0,009597
'ж'	0,00812	'х'	0,009004
'х'	0,007618	'ш'	0,008357
'ш'	0,007071	'ю'	0,006944
'ю'	0,005875	'ц'	0,003984
'ц'	0,003371	'щ'	0,003916
'щ'	0,003313	'э'	0,002664
'э'	0,002254	'ф'	0,002009
'ф'	0,0017		

Перехресні біграми з пробілами		Неперехресні біграми з пробілами		Перехресні біграми без пробілів		Неперехресні біграми без пробілів	
'о '	0,017815	'о '	0,017662	'то'	0,014357	'то'	0,014486
'е '	0,016913	'е '	0,016928	'но'	0,013452	'но'	0,013568
'и '	0,015907	'п'	0,015826	'ст'	0,012807	'ст'	0,012945
'а '	0,015771	'а '	0,0158	'на'	0,012062	'на'	0,012186
'п'	0,015676	'и '	0,015797	'ен'	0,011441	'ен'	0,011717
'с'	0,014109	'с'	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	'ос'	0,009351	'ос'	0,009512

$$R = 0,473368585$$

Лабораторная работа №1

[illegible]

H20

$$R = 0,5514246$$

Лабораторная работа №1

[illegible]

H30

$$R = 0,613354718$$

[illegible]

## 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.