

КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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 1Мб), 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.

Хід Роботи

1. Розрахунок частот та ентропії для літер

Маємо два випадки: з пробілом та без. Далі наведені таблиці з отриманими значеннями:

З пробілом
 $H_1 = 4.354029047$, $R = 0.129194190592$

Літера	Частота
	0,16821
п	0,02283
р	0,0348
е	0,07245
с	0,04403
т	0,05386
у	0,02467
л	0,03823
н	0,05414
и	0,05394
а	0,06627
к	0,02747
з	0,01281
г	0,01405
ь	0,01932
ы	0,01374
й	0,00833
о	0,09543
м	0,02615
в	0,03848
ж	0,00949
ю	0,00467
б	0,01447
д	0,02663
ч	0,01506
э	0,00293
ц	0,00231
я	0,01777
х	0,00708
ш	0,00685
ф	0,00104
щ	0,00249

Без пробілу
 $H_1 = 4.44876864953$, $R = 0.12919419$

Літера	Частота
п	0,02744
р	0,04184
е	0,0871
с	0,05294
т	0,06476
у	0,02965
л	0,04597
н	0,06509
и	0,06485
а	0,07967
к	0,03303
з	0,0154
г	0,01689
ь	0,02323
ы	0,01651
й	0,01001
о	0,11473
м	0,03144
в	0,04627
ж	0,01141
ю	0,00562
б	0,01739
д	0,03202
ч	0,0181
э	0,00353
ц	0,00277
я	0,02137
х	0,00851
ш	0,00823
ф	0,00124
щ	0,00299

2. Розрахунок частот та ентропій для біграм

Загалом маємо 4 випадки, а саме: біграми з перетином з пробілом, біграми з перетином без пробілу, біграми без перетину з пробілом, біграми без перетину без пробілу. Далі будуть наведені невеликі частини таблиць з частотами біграм, бо тих біграм багато і не хочеться засоряти і так кривий протокол. Усі частоти можна буде переглянути у csv файлах, як мають бути у цій же папці)

З перетином біграм:

З пробілом

$H_2 = 3.9432864$, $R = 0.1749084$

Біграма	Частота
пр	0,00666
ре	0,00508
ес	0,00432
ст	0,0094
ту	0,00163
уп	0,00065
пл	0,00062
ле	0,00368
ен	0,00647
ни	0,00752
ие	0,00166
є	0,019
и	0,01165
и	0,01748
н	0,01592
на	0,01002
ак	0,00552
ка	0,00685
аз	0,00338
за	0,00448

Без пробілу

$H_2 = 4.126525$, $R = 0.1508674$

Біграма	Частота
пр	0,00801
ре	0,00611
ес	0,00751
ст	0,01164
ту	0,00216
уп	0,0015
пл	0,00075
ле	0,00482
ен	0,00962
ни	0,00936
ие	0,00245
єи	0,00163
ин	0,00576
на	0,0121
ак	0,00768
ка	0,00829
аз	0,00457
за	0,0054
ан	0,00576
ег	0,00437

3. Оцінка значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$ за допомогою програми CoolPinkProgram

Труднощі

Під час виконання даної лабораторної роботи ми зіткнулися з наступними труднощами:

1. На етапі розробки коду для виконання підрахунків ентропії для літер та біграм виникла проблема підрахунку для випадків, коли якась літера (в нас це була літера “ё”), та деякі біграми не були наявними у тексті, тому під час підрахунку за формулою $\log_2 0$ отримували помилку виконання коду. Після виявлення причини помилки, ми прийшли до висновку: необхідно виконати об'єднання літер “ё” з “е”, та “ь” з “ъ”, а також трохи змінити код так, що перед виконанням обрахунків перевірялося умова не рівності частоти нулю.
2. Після виконання програми та аналізу отриманих результатів, довелося визначитися з оформленням даних для біграм, оскільки досить складно надати дані для всіх біграм (враховуючи їх загальну кількість), тому ми вирішили надати як приклад 20 довільно вибраних біграм й прикріпити csv-файл з повним списком отриманих результатів.

Висновок

В результаті виконання даної лабораторної роботи ми ознайомилися з поняттям ентропії, отримали навички її підрахунку для подальшого розгляду надлишковості вибраної мови. Також провели експерименти з наданою програмою CoolPinkProgram, де спробували вгадати зашифровану літеру по попереднім n-грамам. Загалом поставлене завдання допомогло нам розібратися у значенні криптографії для захисту інформації, розвинути навички роботи у команді та автоматизації роботи з даними за допомогою програмування.