

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

Лабораторна робота № 1
з предмету «Криптографія»

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали:
Студент 3 курсу,
ФТІ, групи ФБ-05
Супрун Максим

Київ - 2021

Мета: 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.

Постановка задачі:

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 $(10) H$, $(20) H$, $(30) H$.
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи

Для виконання даного завдання був використаний текст книги “віднесені вітром”, який був попередньо оброблений з метою виключення всіх символів, що не входять в початковий алфавіт. Текст називається sample_text.txt .

Основною проблемою, яка зайняла найбільше часу, стали рішення при проектуванні програми, що були покликані зменшити час виконання програми.

Повний код можна переглянути у відповідному [репозиторію на Гітхабі](#).

Нижче наведено таблиці, які демонструють виконання коду.

Результати:

В даному прикладі я продемонстрував 10 найчастіших біграм, повну таблицю можна переглянути в репозиторії github

З пробілами:

“ _ ” - пробіл

Частота літер у тексті:

	0,171524
а	0,071128
б	0,015203
в	0,034257
г	0,014619
д	0,025419
е	0,073094
ж	0,009436
з	0,013002

и	0,05478
й	0,008239
к	0,028774
л	0,044022
м	0,026241
н	0,054306
о	0,089762
п	0,020338
р	0,033058
с	0,043354
т	0,055138
у	0,023214
ф	0,000639
х	0,007016
ц	0,002094
ч	0,013093
ш	0,007375
щ	0,002349
ъ	0,000212
ы	0,015587
ь	0,016915
э	0,004001
ю	0,004642
я	0,017165

Частота біграм, що перетинаються:

а	0,021928
о	0,021434
и	0,020708
е	0,020441
н	0,016956
с	0,016469
в	0,015737
п	0,014443
то	0,012872
о	0,012091

В даному прикладі я продемонстрував 10 найчастіших біграм, повну таблицю можна переглянути в репозиторії [github](#)

Частота біграм, що не перетинаються:

а	0,021943
о	0,02151
и	0,020755
е	0,020361
н	0,016809

с	0,016536
в	0,015784
п	0,014402
то	0,012638
о	0,012035

Без пробілів:

Частота літер у тексті:

а	0,085855
б	0,01835
в	0,04135
г	0,017646
д	0,030682
е	0,088227
ж	0,01139
з	0,015694
и	0,066122
й	0,009945
к	0,034732
л	0,053137
м	0,031673
н	0,065549
о	0,108346
п	0,024549
р	0,039903
с	0,05233
т	0,066553
у	0,028021
ф	0,000772
х	0,008469
ц	0,002528
ч	0,015804
ш	0,008902
щ	0,002835
ъ	0,000256
ы	0,018815
ь	0,020417
э	0,004829
ю	0,005603
я	0,020719

Частота біграм, що перетинаються:

то	0,01621
на	0,013311

Лабораторная работа №1

✕

Произвольная часть текста:

о_о_чем_я_собираюсь_говорить_

Использованные буквы:

Порядок n-граммы:

5 символов

10 символов

15 символов

20 символов

25 символов

30 символов

35 символов

40 символов

45 символов

50 символов

Введенный символ:

Символ по счету:

Номер эксперимента:

52

Неравенство для энтропии:

$1,46569495757881 < H < 2,10182279672496$

Двоичная таблица угаданных символов:

01000000000000000000000000000000

10000000000000000000000000000000

00001000000000000000000000000000

10000000000000000000000000000000

00000000000000000000000000000000

Поле ввода символов:

Продолжить

Другой

Вероятности:

$q[1] = 0,5686274$

$q[2] = 0,1960784$

$q[3] = 0,0392156$

$q[4] = 0$

$q[5] = 0,0196078$

$q[6] = 0$

$q[7] = 0$

$q[8] = 0$

$q[9] = 0,0196078$

$q[10] = 0,039215$

$q[11] = 0,039215$

$q[12] = 0$

$q[13] = 0$

$q[14] = 0$

$q[15] = 0,039215$

$q[16] = 0$

$q[17] = 0$

$q[18] = 0$

$q[19] = 0,019607$

$q[20] = 0$

$q[21] = 0,019607$

$q[22] = 0$

$q[23] = 0$

$q[24] = 0$

$q[25] = 0$

$q[26] = 0$

$q[27] = 0$

$q[28] = 0$

$q[29] = 0$

$q[30] = 0$

$q[31] = 0$

$q[32] = 0$

Строка состояния:

$0,711900607 < R < 0,58686228$

Вмиснок

При виконанні роботи навчався рахувати частоту символів та біграм у тексті, оцінювати надлишковість російської мови, та рахувати ентропію. Зі збільшенням значення довжини тексту було помітно, що умовна ентропія зростає та ми маємо більше і більше інформації про наступний символ.

