

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

Лабораторна робота № 1

**«Експериментальна оцінка ентропії на символ джерела відкритого
тексту»**

Виконали:

Студенти 3 курсу:

Групи ФБ-04

Осіпчук Антон

Подима Катерина

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 джерела.

Хід роботи

Для виконання був взятий текст із роману “Злочин і кара” Федора Достоевського. Текст був переведений у нижній регістр, “ё” та “ь” були видалені.

Потім була порахована кількість монограм та біграм, визначені частоти та ентропії.

Всі дані записані у excel файл.

Результати

Текст з пробілами:

монограми:

Ентропія : 4.372861176229328

Надлішковість: 0.1254277647541343

біграми з перетином:

Ентропія: 1.7725260764486034e-05

Надлишковість: 0.9979547691686357

біграми без перетину:

Ентропія: 0.010001063419265328

Надлішковість: 0.9979997873161469

Текст без пробілів:

монограми:

Ентропія для монограм: 4.45631984231334

Надлішковість: 0.10049591031136418

біграми з перетином:

Ентропія: 0.010372789798715037

Надлішковість: 0.9979062618538213

біграми без перетину:

Ентропія: 0.010113284422322623

Надлішковість: 0.9979586427770092

Найчастіша поява монограм

монограми з пробілом

Буква	Частота
-------	---------

о	0.11575178997613365
е	0.07159904534606205
и	0.05489260143198091
н	0.05369928400954654
а	0.0513126491646778
л	0.0441527446300716
р	0.041766109785202864
к	0.0405727923627685
т	0.03699284009546539
в	0.034606205250596656
с	0.03341288782816229
м	0.029832935560859187
у	0.026252983293556086
д	0.025059665871121718
й	0.021479713603818614
п	0.017899761336515514
я	0.017899761336515514
г	0.013126491646778043
ч	0.013126491646778043
х	0.011933174224343675
ы	0.011933174224343675
з	0.011933174224343675
б	0.01282051282051282
ь	0.01282051282051282
ж	0.01282051282051282
ц	0.008547008547008548
ю	0.008547008547008548
щ	0.005698005698005698
ш	0.004273504273504274
э	0.002849002849002849
ф	0.001424501424501424

Монограми без пробілу

Буква	Частота
о	0.11575178997613365
е	0.07159904534606205
и	0.05489260143198091
н	0.05369928400954654
а	0.0513126491646778
р	0.041766109785202864
к	0.0405727923627685
т	0.03699284009546539
в	0.034606205250596656
с	0.03341288782816229
м	0.029832935560859187
у	0.026252983293556086
д	0.025059665871121718
й	0.021479713603818614
п	0.017899761336515514
я	0.017899761336515514
г	0.013126491646778043
ч	0.013126491646778043
х	0.011933174224343675
ы	0.011933174224343675
з	0.011933174224343675
б	0.010739856801909307
ь	0.010739856801909307
ж	0.010739856801909307
ц	0.007159904534606206
ю	0.007159904534606206
щ	0.00477326968973747
ш	0.003579952267303103
э	0.002386634844868735
ф	0.0011933174224343676

Біграми

Біграми без перетину без пробілу	Частота	Біграми без перетину з пробілом	Частота
го	0.008353221957040573	о	0.01207692009459931
ст	0.008353221957040573	е	0.009545225448765168
на	0.008353221957040573	и	0.00879696041561664
мо	0.008353221957040573	а	0.008513937890125228
пр	0.007159904534606206	в	0.008358857054239522
ни	0.007159904534606206	п	0.00815466728699001
от	0.007159904534606206	н	0.008068080486953825
од	0.007159904534606206	с	0.008064203466056682
но	0.0059665871121718375	то	0.007265537161245299

Біграми з перетином без проб	Частота	Біграми з перетином з проб	Частота
од	0.014319809069212411	о	0.024299874642990993
на	0.014319809069212411	е	0.019095620258726527
мо	0.013126491646778043	и	0.017577120407345662
ст	0.011933174224343675	а	0.017202341720621874
ст	0.0046588867780664005	в	0.016691867302498095
ка	0.011933174224343675	п	0.01628865712919526
но	0.011933174224343675	н	0.016147792036599076
хо	0.010739856801909307	с	0.0160831750216467
го	0.010739856801909307	то	0.014629292185218211

Лабораторная работа №1

Произвольная часть текста:
 _брат_каждую_понравившуюся_женщину_вы_не_имеете_права_разного_мнения_держа

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: ж

Символ по счету: 1

Номер эксперимента: 50

Неравенство для энтропии:
 $2,28154783100973 < H < 2,96998800090043$

Двоичная таблица угаданных символов:

00000000000000000000000000000000
00000000000000000000000000000000
00010000000000000000000000000000
00100000000000000000000000000000
00000000000000000000000000000000

Поле ввода символов:
 ж

Продолжить Другой

Вероятности:

q[1] = 0,46
q[2] = 0,14
q[3] = 0,04
q[4] = 0,02
q[5] = 0,04
q[6] = 0,04
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0,02
q[14] = 0
q[15] = 0,02
q[16] = 0,02
q[17] = 0
q[18] = 0,02
q[19] = 0,02
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0,02
q[24] = 0,02
q[25] = 0
q[26] = 0,04
q[27] = 0,04
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0,02
q[32] = 0,02

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

$$R = 0,474846416808984$$

Лабораторная работа №1

Произвольная часть текста:
 ва_будет_пять_люди_расходились_во_взглядах_на_то_по_отношению_к_кому_не_сле

Использованные буквы:
 х, й, э, ц, ы, ф, ж, в, д, а, л, п, о,

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: р

Символ по счету: 14

Номер эксперимента: 50

Неравенство для энтропии:
 $2,18863043781812 < H < 2,99876262358757$

Двоичная таблица угаданных символов:

00000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
00010000000000000000000000000000

Поле ввода символов:
 р

Продолжить Другой

Вероятности:

q[1] = 0,42
q[2] = 0,14
q[3] = 0,1
q[4] = 0,06
q[5] = 0,04
q[6] = 0,02
q[7] = 0,02
q[8] = 0
q[9] = 0
q[10] = 0
q[11] = 0
q[12] = 0,02
q[13] = 0,02
q[14] = 0,02
q[15] = 0,02
q[16] = 0
q[17] = 0,02
q[18] = 0
q[19] = 0
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0,04
q[26] = 0,02
q[27] = 0,02
q[28] = 0,02
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0

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

$$R = 0,481260693859431$$

Лабораторная работа №1

Произвольная часть текста:
 законам_которые_он_разделяет_с_другими_телами_и_организмами_но_тот_закон_к

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: _ (пробел)

Символ по счету: 1

Номер эксперимента: 50

Неравенство для энтропии:
 $1.43463026388786 < H < 2.13166338028599$

Двоичная таблица угаданных символов:

10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
01000000000000000000000000000000
00000000000000000000000000000000

Вероятности:

q[1] = 0,64
q[2] = 0,1
q[3] = 0,02
q[4] = 0,02
q[5] = 0
q[6] = 0,04
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0,04
q[11] = 0,02
q[12] = 0,02
q[13] = 0
q[14] = 0,02
q[15] = 0
q[16] = 0
q[17] = 0,02
q[18] = 0
q[19] = 0
q[20] = 0,02
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0,02
q[27] = 0
q[28] = 0,02
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0

Поле ввода символов:

Продолжить Другой

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

$R = 0,643370635582615$

Висновок

Були засвоєні поняття ентропії на символ джерела та його надлишковості, вивчені та порівняні різні моделі джерела відкритого тексту для наближеного визначення ентропії, отримані практичні навички щодо оцінки ентропії на символ джерела. Після аналізу тексту визначено, що найчастіші у використанні букви російського алфавіту «о», «е», «а».