

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:

Студентки 3 курсу,
групи ФБ-92

Шаповал Ольга
Прохорська Олександра

Перевірила:
Селюх П.В.

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

Спочатку був знайдений текст з яким відбувається подальша робота. Далі текст був відредагований та розбитий у двох різних .txt файлах. У одному з пробілами у іншому без. На цьому етапі виникли невеликі труднощі. Було важко видалити всі не потрібні символи. Спочатку ми намагались зробити це через replase на строку з вручну прописаними різними символами, але це не працювало коректно. Рішення було знайдене: було використано регулярні вирази та відповідну бібліотеку re. Після цього текст редагувався коректно.

Код програми можна знайти на репозиторії в гітхаб (main.py)

Результати

Ентропія:

text_with_spaces:

Entropy for monogram: 4.2909824

Redundant: 0.14180352000000007

text_without_spaces:

Entropy for monogram: 4.49173059

Redundant: 0.09334828323562339

Bigrams for text_with_spaces and with intersection:

Entropy for bigram 3.9157157783714567

Redundant: 0.21685684432570862

Bigrams for text_with_spaces and without intersection:

Entropy for bigram 3.9156052154323326

Redundant: 0.21687895691353343

Bigrams for text_without_spaces and with intersection:

Entropy for bigram 4.159820757213352

Redundant: 0.16034397981122583

Bigrams for text_without_spaces and without intersection:

Entropy for bigram 4.159920757074965

Redundant: 0.16032379493050097

Найчастіша поява літер/біграм

Монограми з пробілами	Монограми без пробілів
' ': 0.20959958012850902,	'о': 0.1121013474214812,
'о': 0.08860495207009862,	'е': 0.08523773749515316,
'е': 0.067371943505065,	'а': 0.07616324156649865,
'а': 0.06019945811293432,	'н': 0.06612301279565723,
'н': 0.052263657076855444,	'т': 0.0623218786351299,
'т': 0.049259239040386774,	'и': 0.06199229352462195,
'и': 0.0489987348306579,	'с': 0.054879556029468785,
'с': 0.043376824128053146,	'л': 0.047124612252811164,
'л': 0.037247313310903155,	'р': 0.04226686700271423,
'р': 0.033407749425597794,	'в': 0.039292119038386975,
'в': 0.031056507385581668,	'м': 0.03691474408685537,
'м': 0.029177429225699124,	'к': 0.03305787126793331,
'к': 0.026128955330232184,	'у': 0.031269387359441646,
'у': 0.024715336898026966,	'д': 0.03044905971306708,
'д': 0.02406694958190032,	'п': 0.027522780147343932,
'п': 0.02175401698449138,	'я': 0.02483399573478092,
'я': 0.019628800655857658,	'б': 0.02082808259015122,

'б': 0.016462525224373616, 'Ы': 0.014149592626964675, 'б': 0.013776075561544598, 'ч': 0.013046280679877677, 'з': 0.012364372601469742, 'Г': 0.01234234467197061, 'ж': 0.009358996828935888, 'й': 0.009145421686400819, 'ш': 0.007303695233060283, 'х': 0.007274005415039713, 'ю': 0.00481262372752792, 'э': 0.004086659790121718, 'ф': 0.003750494431243649, 'щ': 0.002930672359449838, 'ц': 0.0023387914711687926	'Ы': 0.01790180302442807, 'б': 0.017429236138038, 'ч': 0.0165059131446297, 'з': 0.01564317564947654, 'Г': 0.015615306320279177, 'ж': 0.011840829778984102, 'й': 0.011570618456766189, 'ш': 0.009240500193873595, 'х': 0.00920293718495541, 'ю': 0.006088842574641334, 'э': 0.0051703664210934475, 'ф': 0.004745056223342381, 'щ': 0.003707832493214424, 'ц': 0.002958995734780923
--	--

Біграми

Частоти перехресних біграм у тексті з пробілами	Частоти не перехресних біграм у тексті з пробілами	Частоти перехресних біграм у тексті без пробілів	Частоти не перехресних біграм у тексті без пробілів
' ': 0.0572036049 'о ': 0.0219484164 'е ': 0.0177726911 'с ': 0.0154540143 'п ': 0.0153275933 'н ': 0.0151255112 'я ': 0.0139982569 'и ': 0.0135529101 'а ': 0.0135146007 'в ': 0.0125923017	' ': 0.0569277772 'о ': 0.021895741 'е ': 0.0179594495 'с ': 0.0154980702 'п ': 0.0154061276 'н ': 0.0151590319 'я ': 0.0141093542 'и ': 0.0135615297 'а ': 0.0135059811 'в ': 0.0125942172	'то': 0.0153499418 'но': 0.0137044397 'ст': 0.0131034316 'не': 0.0121716266 'по': 0.0120407619 'ен': 0.0110980516 'на': 0.0107551377 'ос': 0.0103177104 'ко': 0.0097288193 'ов': 0.0091908201	'то': 0.0153717526 'но': 0.0138813494 'ст': 0.0130937379 'по': 0.0121801086 'не': 0.0121170997 'ен': 0.0109514347 'на': 0.0109102365 'ос': 0.0102244087 'ко': 0.0098560489 'ов': 0.0092332299

В таблиці наведено 10 найчастіших біграм. Усі інші частоти біграм можна подивитися в текстовому файлі 1.txt у репозиторії.

Cool pink program

Значення $H(10)$:

[illegible]
$$R = 0.46001528589999996$$

Значения $H(20)$:

Лабораторная работа №1

Произвольная часть текста:
_человек_отсюда_сле

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: 51

Неравенство для энтропии:
 $2,00435220361306 < H < 2,81732065859684$

Двоичная таблица угаданных символов:

Поле ввода символов:

Продолжить Другой

Вероятности:

$q[1]$	= 0,44
$q[2]$	= 0,2
$q[3]$	= 0,04
$q[4]$	= 0,04
$q[5]$	= 0,02
$q[6]$	= 0,04
$q[7]$	= 0,02
$q[8]$	= 0,02
$q[9]$	= 0,04
$q[10]$	= 0,02
$q[11]$	= 0
$q[12]$	= 0
$q[13]$	= 0
$q[14]$	= 0
$q[15]$	= 0,04
$q[16]$	= 0
$q[17]$	= 0
$q[18]$	= 0,02
$q[19]$	= 0
$q[20]$	= 0
$q[21]$	= 0
$q[22]$	= 0
$q[23]$	= 0
$q[24]$	= 0
$q[25]$	= 0,02
$q[26]$	= 0,02
$q[27]$	= 0
$q[28]$	= 0,02
$q[29]$	= 0
$q[30]$	= 0
$q[31]$	= 0
$q[32]$	= 0

Строка состояния:

$R = 0.517832713779316$

Значения $H(30)$:

Лабораторная работа №1

Произвольная часть текста:
_на_первое_место_самого_себя_

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

Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:
 $1,06030488513371 < H < 1,6786157951538$

Двоичная таблица угаданных символов:

01000000000000000000000000000000	▲
10000000000000000000000000000000	■
10000000000000000000000000000000	
10000000000000000000000000000000	
10000000000000000000000000000000	
10000000000000000000000000000000	▼

Вероятности:

$q[1] = 0,6470588$
$q[2] = 0,1960784$
$q[3] = 0,0196078$
$q[4] = 0,0392156$
$q[5] = 0$
$q[6] = 0$
$q[7] = 0,0196078$
$q[8] = 0,0392156$
$q[9] = 0$
$q[10] = 0$
$q[11] = 0$
$q[12] = 0$
$q[13] = 0$
$q[14] = 0$
$q[15] = 0$
$q[16] = 0$
$q[17] = 0$
$q[18] = 0,019607$
$q[19] = 0$
$q[20] = 0$
$q[21] = 0$
$q[22] = 0$
$q[23] = 0$
$q[24] = 0$
$q[25] = 0$
$q[26] = 0$
$q[27] = 0$
$q[28] = 0$
$q[29] = 0$
$q[30] = 0$
$q[31] = 0$
$q[32] = 0,019607$

Строка состояния:

$R = 0.726107931971249$

Висновки: в ході лабораторної роботи ми вивчили що таке ентропія на символі джерела та надлишковість, дослідили та порівняли різні джерела відкритого тексту для наближеного визначення ентропії (з пробілом та без пробілів), а також набули практичних навичок щодо оцінки ентропії на символі джерела.

Під час аналізу тексту ми експериментально перевірили той факт, що в російському алфавіті найчастіше зустрічається пробіл, що свідчить про те, що при шифруванні потрібно його видаляти, щоб зломисник мав прикласти більше зусиль для зламу. Найбільш вживані монограми в тексті – це “о”, “е”, “а”, а рідше за все зустрічаються: “щ”, “ц”. Серед біграм у тексті з пробілом частіше всього зустрічаються: “о_”, “е_”, “_с”, а у тексті без пробілів: “то”, “ст”, “но”.