

КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела
відкритого тексту

Виконав

ФБ-12 Сущенко Олександр

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 1 Н та 2 Н за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1 Н та 2 Н на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1 Н та 2 Н 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

Створюємо програму, яка спочатку буде фільтрувати за критеріями текст, створюючи два нових з пробілами та без відповідно. Далі обчислюємо значення частот символів та їх ентропію, використовуючи формулу:

$$H_1 = - \sum_{i=1}^n p(i) \log_2 p(i)$$

Далі рахуємо частоти біграм з перетинами та без; з пробілами та без. Наступний етап – обчислення ентропії, використовуємо наступну формулу:

$$H_2 = - \sum_{i,j} p(i,j) \log_2 p(i,j)/2$$

Потім рахуємо надлишковість за допомогою іншої формули:

$$R = 1 - \frac{H_{\infty}}{H_0}$$

Результати виконання:

Частота символів без пробілів:

о 0.11223058354016847
е 0.08425898160360601
а 0.07951647652602405
и 0.07499076263817471
н 0.06438223609826496
т 0.05865863220583453
с 0.05533634678361787
л 0.04760465515563471
р 0.04738786396590207
в 0.04736729253913912
м 0.032533711425607785
к 0.030935469807870788
д 0.03041722809518874
п 0.02819630521351163
у 0.02523560217708992
я 0.019903646601877222

ы 0.01914487743781297
г 0.01809256983801584
ь 0.01760281163931326
б 0.016867778736898573
з 0.016728526001887824
ч 0.012957625234494485
й 0.011432174819149467
х 0.010305493599517046
ж 0.009618724429123126
ш 0.007281493865363177
ю 0.006091515947998598
ц 0.005409494029934591
щ 0.0037978018639295064
э 0.0035580656212689564
ф 0.0018625053307687622
ъ 0.0002650549217534135
ё 1.582417443303961e-06

0.15449078722822107
о 0.09489191725865267
е 0.07124168883757118
а 0.06723186027547
и 0.06340533051650793
н 0.054435730689905926
т 0.04959637469754025
с 0.04678735398812037
л 0.04025014265864879
р 0.0400668438555097
в 0.04004945053842351
м 0.027507530971811453
к 0.02615620402896121
д 0.02571802623313601
п 0.02384021696156144
у 0.021336917248618067
я 0.016828703254624114

Ы 0.016187157443363729
Г 0.0152974223773052
Ь 0.014883327635907032
Б 0.014261851036942736
З 0.014144111659743903
Ч 0.01095578284309823
Й 0.009666001483783743
Х 0.00871338288644773
Ж 0.008132713685262576
Ш 0.006156565274777617
Ю 0.005150428778714861
Ц 0.004573773419934213
Щ 0.0032110739236045375
Э 0.003008374882177001
Ф 0.001574764170034392
Ъ 0.00022410620091823334
Ё 1.3379474681685574e-06

[illegible]

Лабораторная работа №1

 $H^{(30)}:$

[illegible]

$$0,432443812140248 < R < 0.574843682564354$$

Висновки

Під час виконання цієї лабораторної роботи, я набув навичок оцінки частот букв та біграм. За допомогою практичної частини я навчився розраховувати ентропію та надлишковість. Також цікавою виявилася можливість застосування цих знань при роботі з CoolPinkProgram. Я використовував ці дані для передбачення наступного символу, вибираючи його на основі ймовірності, хоча в деяких випадках всеодно виникали складності.