

Лабораторна робота 1

ФБ-11 Яцентюк Андрій ФБ-11 Кустов Іван (Варіант 4)

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

В цій лабі було вираховано частоти монограм та біграм для тексту "Біблія".(Text.txt)
Текстові файли з частотами знаходяться в файлах results1.txt (Текст з пробілами)
results2.txt (Текст без пробілів).

Частоти були вираховані за допомогою python скрипта, з використанням функції "**count**".

Ентропії були пораховані з використанням бібліотеки "**math**".

В цілому робота була не дуже складною, через зручність використання мови програмування **Python**.

Найскладнішою частиною було розібратись з архітектурою **git** та **github**.

Результат работы

Лабораторная работа №1

Произвольная часть текста:
де_восхищ

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: 50

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:
2,202748430524< H < 2,91457540542538

Двоичная таблица угаданных символов:
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
01000000000000000000000000000000
00000010000000000000000000000000

Вероятности:
q[1] = 0,4693877
q[2] = 0,0816326
q[3] = 0,0612244
q[4] = 0,0612244
q[5] = 0,0612244
q[6] = 0
q[7] = 0,0408163
q[8] = 0,0204081
q[9] = 0,0204081
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] = 0,020408
q[19] = 0
q[20] = 0
q[21] = 0,020408
q[22] = 0,020408
q[23] = 0,020408
q[24] = 0,040816
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0,020408
q[29] = 0,040816
q[30] = 0
q[31] = 0
q[32] = 0

Строка состояния:

Лабораторная работа №1

Произвольная часть текста:
сли_нет_никакого_за

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

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:
4,39063309650523< H < 4,12553325055685

Двоичная таблица угаданных символов:
00000000000000000000000000000000
00000000000000000000000000000000
01000000000000000000000000000000
00000000000000000000000000000000
00000000000000000000000000000000

Вероятности:
q[1] = 0,14
q[2] = 0,04
q[3] = 0
q[4] = 0,08
q[5] = 0,04
q[6] = 0,04
q[7] = 0,04
q[8] = 0
q[9] = 0
q[10] = 0
q[11] = 0,02
q[12] = 0,06
q[13] = 0,04
q[14] = 0,02
q[15] = 0
q[16] = 0,02
q[17] = 0,04
q[18] = 0,02
q[19] = 0,04
q[20] = 0
q[21] = 0,08
q[22] = 0
q[23] = 0,02
q[24] = 0,04
q[25] = 0,02
q[26] = 0
q[27] = 0
q[28] = 0,04
q[29] = 0
q[30] = 0
q[31] = 0,04
q[32] = 0,12

Строка состояния:

[illegible]

Таблиці частот біграм знаходяться в файлах **bigram_table_russian1.xlsx**(з пробілами) **bigram_table_russian2.xlsx** (без пробілів)

Частоти для монограм однакові, з пробілами, чи без.

Взагалом, ентропія з CoolPinkProgram виявилась у середньому меншою за ту що підрахував скрипт.

3 пробілами

Ентропія монограм: 4.433878685419471

Ентропія біграм: 3.930350462829067

Без пробілів

Ентропія монограм: 4.433878685419471

Ентропія біграм: 4.1144731556708365

Надлишковість:

$$R = 1 - \frac{H_1}{H_0}$$

,де H_1 – ентропія відкритого тексту, H_0 – максимальна ентропія відкритого тексту

Максимальна ентропія відкритого тексту обчислюється за формулою:

$$H_0 = \log_2 n$$

,де n – кількість букв в алфавіті

$$H_0 = \log_2(n) = \log_2(32) = 5$$

$$H_1(\text{3 пробілами}) = 4.909277396497351$$

$$H_1(\text{Без пробілів}) = 4.433878685419471$$

$$H_2(\text{3 пробілами})(\text{3 перетином}) = 3.9261968370593516$$

$$H_2(\text{3 пробілами})(\text{Без перетинів}) = 3.8500985125798097$$

$$H_2(\text{Без пробілами})(\text{3 перетином}) = 4.1144731556708365$$

$$H_2(\text{Без пробілами})(\text{Без перетинів}) = 4.018163043724456$$

$$R_1(\text{3 пробілами}) = 0.181445207005298$$

$$R_1(\text{Без пробілів}) = 0.1832242629161058$$

$$R_2(\text{3 пробілами})(\text{3 перетином}) = 0.2147603258812968$$

$$R_2(\text{3 пробілами})(\text{Без перетинів}) = 0.22998029748403806$$

$$R_2(\text{Без пробілами})(\text{3 перетином}) = 0.1771053688658327$$

$$R_2(\text{Без пробілами})(\text{Без перетинів}) = 0.1963673912551088$$

В цілому, російська мова є досить передбачуваною якщо знати велику бібліотеку слів, та знати якусь кількість букв до необхідного місця. **Крім того моменту коли останній відомий символ є пробілом. Тоді біда. Тоді вгадати першу букву слова є значно складнішою задачею.**

Частоти монограм майже співпадають з нормою наданою Вікіпедією.