

Міністерство освіти і науки України Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» Фізико-технічний інститут

КРИПТОГРАФІЯ КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1 Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

Студенти групи ФБ-73 Мухамедзянов Артем Лень 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 Н1 та Н2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення Н1 та Н2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 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, опис труднощів, що виникали, та шляхи їх розв'язання

Першим кроком до виконання комп'ютерного практикуму було отримання важливої інформації у методичці та на лекціях, щодо того, що таке ентропія, як її рахувати та для чого вона використовується. Отримавши завдання практикуму, перш за все, ми вирішили, на якій мові будемо писати нашу програму та які основні засоби ми використаємо. Розбивши наше завдання по етапам, ми прийнялись їх втілювати у життя, а саме написання циклів підрахунку частот букв. Для виконання нашої роботи ми використовували знайому всім і діючу методику вирішення задач будь-якої складності «Step by step». Але, як, мабуть, і у всіх, у нас почали виникати труднощі в написанні робочого коду, бо кількість помилок на перших етапах вражала =)

Найбільшою проблемою стало написання циклу запису підрахунку «не перекрестних» біграм до txt файлу. Вирішенням цієї проблеми стало «Гуглення» розв'язку близько двух годин, бо всі запропоновані засоби вирішення подібної проблеми, як у нас, або не працювали, або працювали неккоректно. Але, з часом, ми знайшли потрібний спосіб вирішення і здивувалися, як самі до цього не додумались.

Код программы

```
#include "locale.h"
#include "stdlib.h"
#include <fstream>
#include <ctype.h>
#include <algorithm>
#include <iostream>
#include<math.h>
#include<string.h>
using namespace std;
char AlphaZ[]="абвгдеёжзийклмнопрстуфхцчшщъыьэюя ";
int sizeAlpha = 34;
int mas, cc=0;
int **bigram;
long double count monograms, count bigrams = 0;
long double enthropy = 0.0;
long double chastota = 0.0;
int monograms [34] = \{0\};
void Sort()
{
     int temp;
     char ctemp;
      for (int i = 0; i < sizeAlpha; i++)</pre>
        for (int j = 0; j < sizeAlpha - i - 1; j++) {
            if (monograms[j] < monograms[j + 1]) {</pre>
                temp = monograms[j];
                monograms[j] = monograms[j + 1];
                monograms[j + 1] = temp;
                ctemp = AlphaZ[j];
                AlphaZ[j] = AlphaZ[j + 1];
                AlphaZ[j + 1] = ctemp;
            }
        }
```

```
void CountMonograms(char *text, char *AlphaZ)
{
    char temp;
    while(*text)
         temp = tolower(*text);
          for(int i=0;i<sizeAlpha;i++)</pre>
          if(temp==AlphaZ[i])
           monograms[i]++;
           count_monograms++;
          }
         text++;
    }
}
void CountBigrams(char *text, char *AlphaZ)
{
  char prev=0;
 while(*text)
   char curr = tolower(*text);
   char *prev_ = find(AlphaZ, AlphaZ+mas, prev);
   char *curr_ = find(AlphaZ, AlphaZ+mas, curr);
   {
     bigram[prev_ - AlphaZ][curr_ - AlphaZ]++;
     count_bigrams++;
      cc++;
```

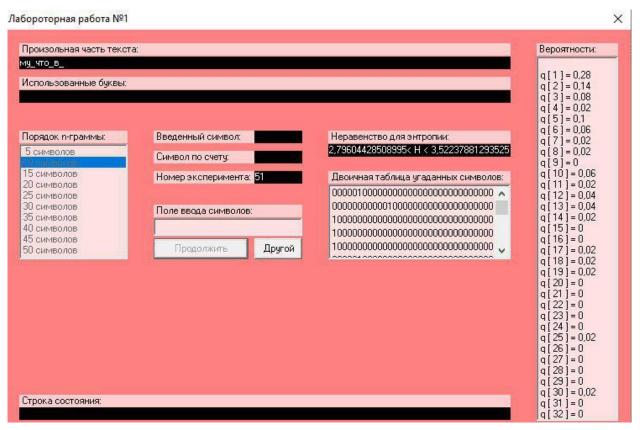
```
prev = curr;
    text++;
 }
}
void SaveBigrams(char *filename, char *AlphaZ)
 ofstream ofs;
 ofs.open (filename, std::ofstream::out);
 ofs << " ";
 for( int i = 0; i < mas; i++ )
  ofs << AlphaZ[i] << " ";
 ofs << endl;
 for( int j = 0; j < mas; j++ )
   ofs << AlphaZ[j] << " ";
   for( int i = 0; i < mas; i++ )
    ofs.width(4);
    ofs << bigram[i][j] << " ";
   ofs << endl;
 ofs.close();
}
long double EnthropyMonograms()
  chastota, enthropy = 0;
  Sort();
 for(int i=0;i<sizeAlpha;i++)</pre>
    {
          chastota = (monograms[i]/count monograms);
          if(chastota!=0)
```

```
{
             \verb|cout|<<| AlphaZ[i]<<" "<<| monograms[i]<<" "<<| chastota<<| endl|;
       enthropy -= chastota * (log(chastota)/log(2));
return enthropy;
}
long double EnthropyBigrams()
    chastota, enthropy = 0;
    for(int i = 0; i < mas; i++)
  {
      for(int j = 0; j < mas; j++)
          chastota = (bigram[i][j]/count_bigrams);
               if(chastota!=0)
           {
       enthropy -= chastota * (log(chastota)/log(2));
 return enthropy;
 }
int main(int argc, char* argv[])
{
    setlocale(LC_ALL, "rus");
    string str;
    ifstream file;
    file.open ( "text.txt" );
    getline ( file, str, '\0' );
    file.close();
 char *input_text = new char[str.length()+1];
 strcpy(input text, str.c str());
 mas = sizeof(AlphaZ) - 1;
```

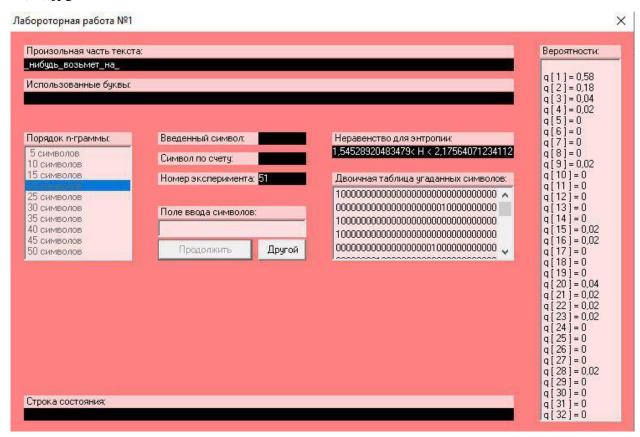
```
bigram = new int* [mas];
for(int i = 0; i < mas; i++)
{
    bigram[i] = new int[mas];
    for(int j = 0; j<mas; j++)
    bigram[i][j] = 0;
}
CountBigrams( input_text, AlphaZ );
SaveBigrams( "bigrams неперекрёстные.txt", AlphaZ);
CountMonograms( input_text, AlphaZ );
cout<<endl;
cout<<"H1 : "<< EnthropyMonograms() <<endl;
cout<<"H2 : "<< EnthropyBigrams() <<endl;
cout<<fixed<<"Bcero символов в тексте : "<<iint(count_monograms)<<endl;
cout<<fixed<<"Bcero биграм в тексте : "<<iint(count_bigrams)<<endl;
return 0;
}</pre>
```

Скріншоти до результатів розової програмки:

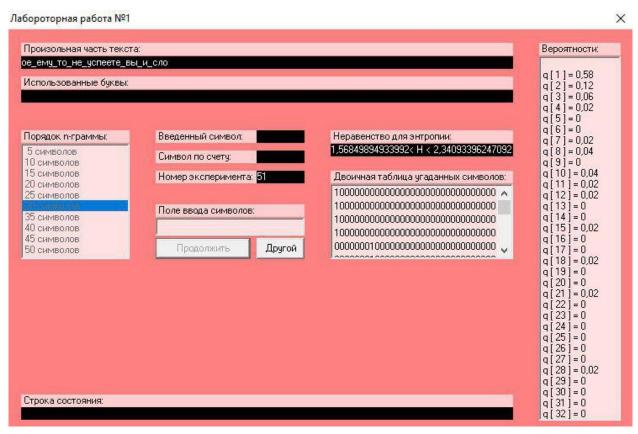
H(10).jpg



H(20).jpg



H(30).jpg



Результати роботи програми з пробілами

```
327015 0.154086
0 152292 0.0931342
e 109122 0.0667336
a 105671 0.0646231
и 90629 0.0554242
н 84617 0.0517475
т 77729 0.0475352
с 70572 0.0431583
л 64887 0.0396817
в 60707 0.0371254
р 57550 0.0351947
к 43126 0.0263737
д 40011 0.0244687
м 38363 0.0234609
п 35247 0.0215553
у 34631 0.0211786
я 27973 0.0156172
ь 25196 0.0154086
г 25027 0.0153053
6 22386 0.0136902
з 22116 0.015325
ч 19176 0.017271
й 14756 0.00902403
ж 14258 0.00871948
ш 11761 0.00719244
х 11162 0.00682612
ю 7735 0.00473034
ц 4734 0.00289508
в 3998 0.00244498
щ 3935 0.00240645
ф 2747 0.00167993
ь 523 0.000319841
Н1 : 4.28745
Н2 : 3.953165
Всего символов в тексте : 1635189
Всего биграм в тексте : 732620
```

Результати роботи програми без пробілів

```
o 152292 0.116416
  109122 0.0834155
   105671 0.0807775
  90629 0.069279
84617 0.0646833
77729 0.0594179
  70572 0.053947
64887 0.0496012
60707 0.0464059
57550 0.0439926
43126 0.0329666
  40011 0.0305854
   38363 0.0293256
35247 0.0269437
34631 0.0264728
   27973 0.0213832
25537 0.0195211
  25196 0.0192604
25027 0.0191312
22386 0.0171124
22116 0.016906
19176 0.0146586
  14756 0.0112798
14258 0.0108992
11761 0.00899039
11162 0.0085325
 7735 0.00591282
4734 0.00361878
  3998 0.00305617
3935 0.00300801
2747 0.00209987
ь 523 0.000399794
Н1 : 4.45686
12 : 3.99106
Всего символов в тексте : 1308174
Зсего биграм в тексте : 527085
```

	0	grams неперекрестные.txt			
6 449 8 2476 7 7 602 1 1376 16 6 780 16 6 780 16 6 9 32 2 552 1 8 877 1 8 449 1 1 8 266 1 1 8 266 7 8 8 9 7 7 8 9 7 7 8 9 7 7 8 9 7 7 8 9 7 8 9 7 9 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		2 0 94 33 19 437 0 2 0 212 1384 2147 671 0 3 7 228 15 10 963 0 4 0 212 245 147 1265 0 6 1839 4371 2529 3267 1800 96 6 0 0 0 0 0 0 0 0 6 0 22 23 26 686 0 7 0 0 0 0 0 0 0 7 0 0 0 0 0 0 7 0 0 0 0	0 54 159 222 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	N
bigrams перекрёстные.txt					
6 907 8 5004 1' r 1169 2' 4 2778 324 e 1576 324 H 172 104 H 172 104 H 172 104 H 172 104 H 172 104 H 172 104 H 173 104 H 174 44 X 1449 14 X 149 16 X 149 16 X 150 16	6 B	ё ж 3 и й к л м н 0 2128 7742 135 0 11145 10138 4101 16381 10 77 304 628 2 0 33 41 1 75 31. 0 0 1262 36667 1 723 4 1 47 1145 10138 4101 16381 10 115 34 732 0 1 156 78 123 742 10 10 974 2222 225 0 416 0 911 74 51 51 575 7365 4582 2 583 6700 4325 13499 38. 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	5 1887 12629 2238 8281 20 3 8 689 0 8 0 7 28 8 689 0 16 0 436 2746 4221 1307 0 14 413 31 24 1890 0 18 3712 9011 5141 6513 342 177 0 10 10 10 10 10 10 10 10 10 10 10 10 1	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2062

Посчёты.txt (ентропия)

bigrams неперекрёстные.txt

H1 = 4.45686 без пробела H1= 4.28745 c пробелом неперекрёстные без пробела H2 = 3.991915перекрёстные без пробела H2 = 3.99106перекрёстные с пробелом H2 = 3.953165неперекрёстные с пробелом

H2 = 3.952285

Результаты розовой программки.txt

```
H NP = (max+min)/2
H10 = (2.796+3.522) / 2 = 3.16

H20 = (1.545+2.175) / 2 = 1.86

H30 = (1.568+2.341) / 2 = 1.95

Избыточность R = 1 - 1.95/5 = 0.61
```

Висновки

Під час виконання даного практикуму ми навчилися рахувати ентропію на символ джерела та знаходити надлишковість; навчилися порівнювати різні моделі джерела відкритого тексту для наближеного визначення ентропії.

По результатам нашої програми ми можемо бачити, що дані, які ми отримали ϵ близькими до середньостатистичних даниї, які можна знайти в інтернеті.