Криптографія. Лабораторна робота 1.

Експерементальна оцінка ентропії на символ джерела

Текст та форматування

Як середньостатистичний текст середньостатистичною російською мовою середньостатистичного росіянина ми взяли ~500000 символів (~1Мб) тексту ниття терориста та військового злочинця _ігоря _стрєлкова—_гіркіна з його телеграм-каналу.

Зчитуємо текст в змінну-рядок:

```
text = fileread('girkin.txt')

text =
    'Хитроумный Одиссей между Сциллой и Харибдой

В моей далекой уже молодости мною предполагалось (совершенно наивно), что каждый гуманитарий хотя-бы в конспектий "ликбез":
    Одиссей был царем на острове Итака (мелкий островок на Адриатическом море в нынешней Греции). Участвовал в оса, Подтвердить своё прозвище Одиссей смог самым наилучшим образом. - В то время, как все остальные греческие цари Ну так вот, последние спутники Одиссея (вместе с его последним кораблем) погибли в проливе между современной Си Но - все по порядку: Итак, пройти в те мифические времена пролив нельзя было, не угодив к одному из чудовищ. Х.

Так вот к чему это я всё? - А к тому, что наш президент (не критикую!) и его ближайшее окружение - находятся стольные пролив нельзя было, не угодив к одному из чудовищ. Х.
```

Форматуваємо текст (видалення зайвих символів та зайвих пробілів):

```
formated_text = blanks(length(text));

for i = 1:length(text)
    c = text(i);
    if (c <= '9' && c >= 'a') || c == ' '
        formated_text(i) = c;
    elseif (c <= '9' && c >= 'A')
        formated_text(i) = char(c + 0x0020);
    elseif c == 'Ë' || c == 'ë'
        formated_text(i) = 'e';
    else
        formated_text(i) = ' ';
    end
end

mask = (formated_text == ' ');
mask = ~(mask & [0 mask(1:end-1)]);
```

```
formated_text2 = formated_text(mask); % 3 пробілами formated_text3 = formated_text(formated_text ~= ' '); % Без пробілів
```

Оригінальний (файл girkin.txt) та кінцевий форматований текст (змінна formated_text2) мають бути десь разом з цим файлом.

Робота з форматованим текстом

Створюємо два відображення типу "символ -> кількість входжень в тексті" та "біграма -> кількість входжень в тексті":

```
bigrams = dictionary(string([]), []);
bigrams_wo_spaces = dictionary(string([]), []);
bigrams_wo_spaces_i = dictionary(string([]), []);
bigrams_i = dictionary(string([]), []);
chars = dictionary(string([]), []);
```

Заповнюємо дані про окремі символи :

```
for i = 1:length(formated_text2)
    c1 = formated_text2(i);
    if chars.isKey(c1)
        chars(c1) = chars(c1) + 1;
    else
        chars(c1) = 1;
    end
end
tttt = chars.keys'
```

```
"x" "u" "T" "p" "o" "y" "m···

ttt = chars.values'

ttt = 1×33
5356 37250 31194 24918 56742 12712···
```

Заповнюємо дані про <u>біграми</u> :

tttt = 1×33 string

```
for i = 1:(length(formated_text2)-1)
    c2 = [formated_text2(i) formated_text2(i+1)];

if bigrams.isKey(c2)
    bigrams(c2) = bigrams(c2) + 1;
else
    bigrams(c2) = 1;
```

```
end
end
for i = 1:2:(length(formated_text2)-1)
    c2 = [formated_text2(i) formated_text2(i+1)];
    if bigrams i.isKey(c2)
        bigrams_i(c2) = bigrams_i(c2) + 1;
    else
        bigrams i(c2) = 1;
    end
end
for i = 1:2:(length(formated text3)-1)
    c2 = [formated_text2(i) formated_text2(i+1)];
    if bigrams_wo_spaces_i.isKey(c2)
        bigrams_wo_spaces_i(c2) = bigrams_wo_spaces_i(c2) + 1;
    else
        bigrams_wo_spaces_i(c2) = 1;
    end
end
for i = 1:(length(formated_text3)-1)
    c2 = [formated_text3(i) formated_text3(i+1)];
    if bigrams wo spaces.isKey(c2)
        bigrams_wo_spaces(c2) = bigrams_wo_spaces(c2) + 1;
    else
        bigrams_wo_spaces(c2) = 1;
    end
end
```

Ці зібрані дані прикладені окремими файлами (див. нижче)

Обчислення ентропій H_1 (символів з пробілами), H_2 (біграм з пробілами з перетином), H_3 (символів без пробілів), H_4 (біграм без пробілів з перетином), H_5 (біграм з пробілами без перетину) та H_6 (біграм без пробілів без перетину)

```
t1 = chars.values / sum(chars.values);
H1 = -sum(t1 .* log2(t1))

H1 = 4.3851

chars(' ') = [];
t3 = chars.values / sum(chars.values);
```

```
H3 = -sum(t3 \cdot start log2(t3))
```

```
H3 = 4.4392
```

```
t_b = bigrams.values / sum(bigrams.values);
H2 = -sum(t_b .* log2(t_b)) * 0.5
```

H2 = 3.9881

```
t_b4 = bigrams_wo_spaces.values / sum(bigrams_wo_spaces.values);
H4 = -sum(t_b4 .* log2(t_b4)) * 0.5
```

H4 = 4.1241

```
t_b5 = bigrams_i.values / sum(bigrams_i.values);
H5 = -sum(t_b5 .* log2(t_b5)) * 0.5
```

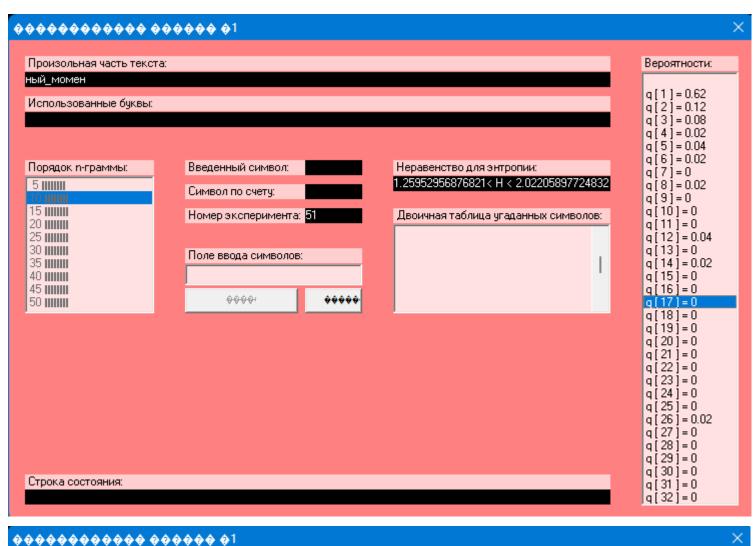
H5 = 3.9881

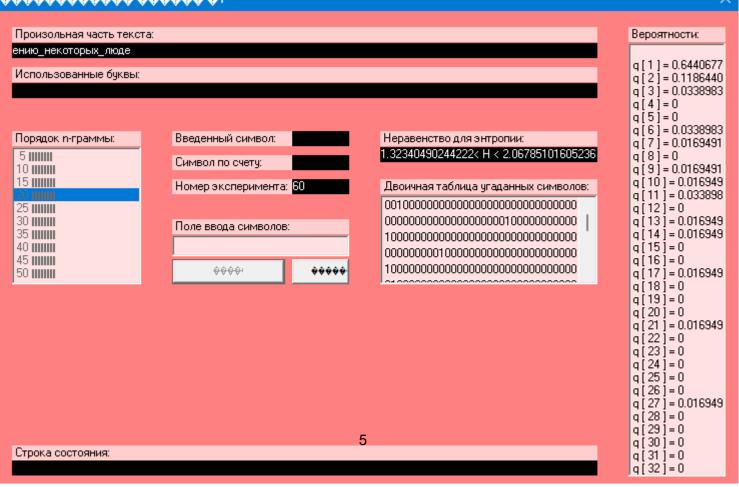
```
t_b6 = bigrams_wo_spaces_i.values / sum(bigrams_wo_spaces_i.values);
H6 = -sum(t_b6 .* log2(t_b6)) * 0.5
```

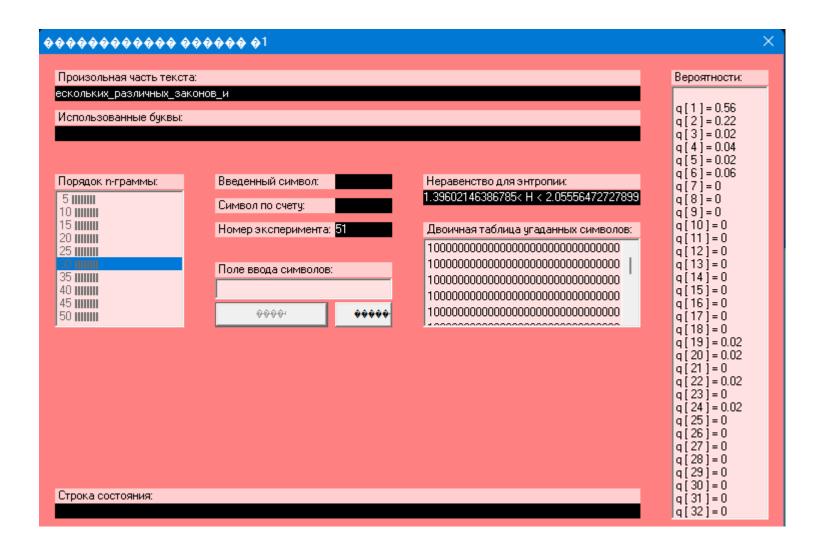
H6 = 3.9878

Оцінки для $H^{(10)},\;H^{(20)},\;H^{(30)}$ та обчислення надлишковості мови

За допомогою CoolPinkProgram.exe обчислимо $H^{(10)},\ H^{(20)},\ H^{(30)}$:







Отже, маємо такі результати:

$$1.26 \le H^{(10)} \le 2.02$$

 $1.32 \le H^{(20)} \le 2.07$
 $1.40 \le H^{(30)} \le 2.06$

Візьмемо $H^{(30)}$ як найкраще наближення для H_{∞} , тоді обчислимо надлишковість російської мови R за формулою :

$$R = 1 - \frac{H_{\infty}}{H_0}$$
, де $H_0 = log_2(32) = 5$

Тоді надлишковість R буде в таких межах :

$$0.58 \le R \le 0.72$$

Тобто маємо надлишковість російської мови в районі 65%.

Візуалізація

Наступний код потрібен лише для красивих табличок з даними (які також мають бути десь разом з цим файлом) і ніякого іншого корисного навантаження не несе.

(можна було б зробити ці таблички інтерактивними прямо тут, але це занадто складно, а результат не вартує того)

```
keys c = chars.keys;
values c = chars.values;
[sortedValues_c, sortInd_c] = sort(values_c);
sortedKeys_c = keys_c(sortInd_c);
svalues1 = zeros(32, 1);
for i = 1:32
   for j = 1:32
        b = string([char(i + 'a' - 1) char(j + 'a' - 1)]);
        if bigrams.isKey(b)
            svalues1(i) = svalues1(i) + bigrams(b);
        end
    end
end
alph = 'абвгдежзийклмнопрстуфхцчшщъыьэюя';
[~, sortInd1] = sort(svalues1);
alph = alph(sortInd1);
bfreq matrix = zeros(32, 32);
for i = 1:32
    for j = 1:32
        b = string([alph(i) alph(j)]);
        if bigrams.isKey(b)
            bfreq_matrix(i, j) = bigrams(b);
        end
    end
end
```