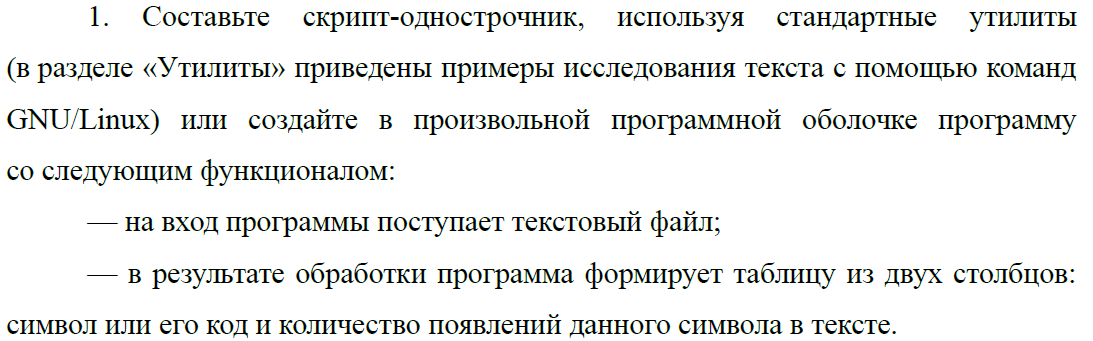
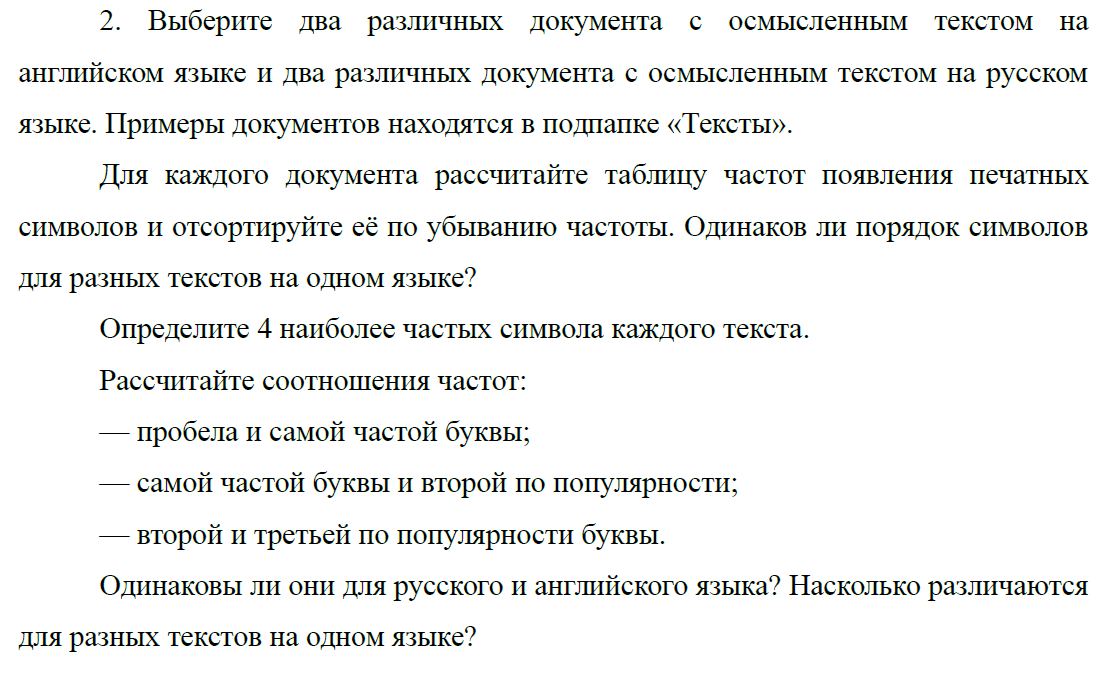
**Лабораторная работа №2**

**«Исследование статистических характеристик текстового документа»**

****

****

****

**Михаил Лермонтов. Герой нашего времени**

Таблица частот прилагается в отдельном файле «2. Михаил Лермонтов. Герой нашего времени.txt»

4 наиболее частых: « », «о», «е», «а»

Соотношение частот пробела и самой частой буквы: 56825 / 22921 = 2,48

Соотношение частот самой частой и второй по популярности: 22921 / 18026 = 1,27

Соотношение частот второй и третьей по популярности букв: 18026 / 17400 = 1,04

**Лев Николаевич Толстой. Война и мир 4+эпилог**

Таблица частот прилагается в отдельном файле «2. Лев Николаевич Толстой. Война и мир 4+эпилог.txt»

4 наиболее частых: « », «о», «е», «а»

Соотношение частот пробела и самой частой буквы: 107504 / 63765 = 1,69

Соотношение частот самой частой и второй по популярности: 63765 / 46555 = 1,37

Соотношение частот второй и третьей по популярности букв: 46555 / 42268 = 1,10

**Three Men in a Boat, by Jerome K. Jerome**

Таблица частот прилагается в отдельном файле «2. Three Men in a Boat, by Jerome K. Jerome.txt»

4 наиболее частых: « », «e», «t», «a»

Соотношение частот пробела и самой частой буквы: 67083 / 36155 = 1.86

Соотношение частот самой частой и второй по популярности: 36155 / 27981 = 1.29

Соотношение частот второй и третьей по популярности букв: 27981 / 23743 = 1.17

**Treasure Island, by Robert Louis Stevenson**

Таблица частот прилагается в отдельном файле «2. Treasure Island, by Robert Louis Stevenson.txt»

4 наиболее частых: « », «e», «t», «a»

Соотношение частот пробела и самой частой буквы: 67026 / 35696 = 1.88

Соотношение частот самой частой и второй по популярности: 35696 / 25840 = 1.38

Соотношение частот второй и третьей по популярности букв: 25840 / 23981 = 1.08

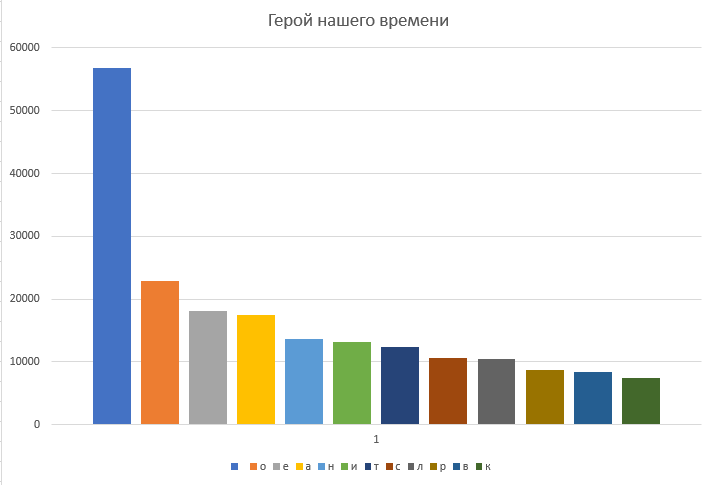
Порядок символов в целом не одинаков. Но существует закономерность расположения символов. Так в списках совпадают первые 4 символа (для русских текстов) и первые 6 символов (для английских текстов). А далее символы можно объединить в поименованные группы, порядок которых будет одинаков для разных текстов, но внутри группы порядок символов будет несколько отличаться.

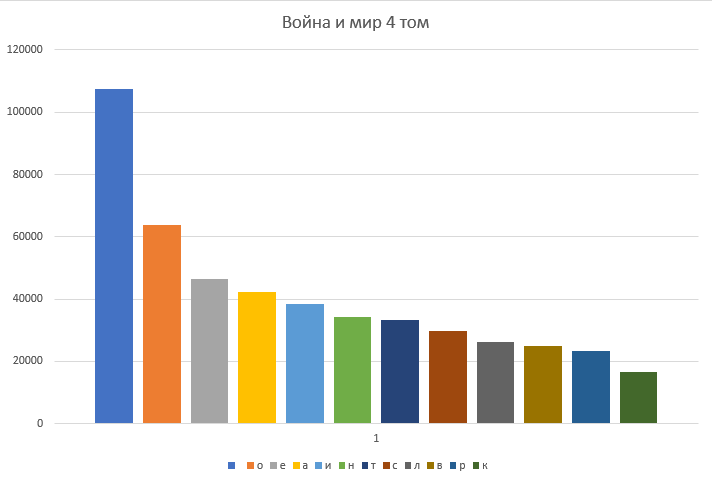
Например:

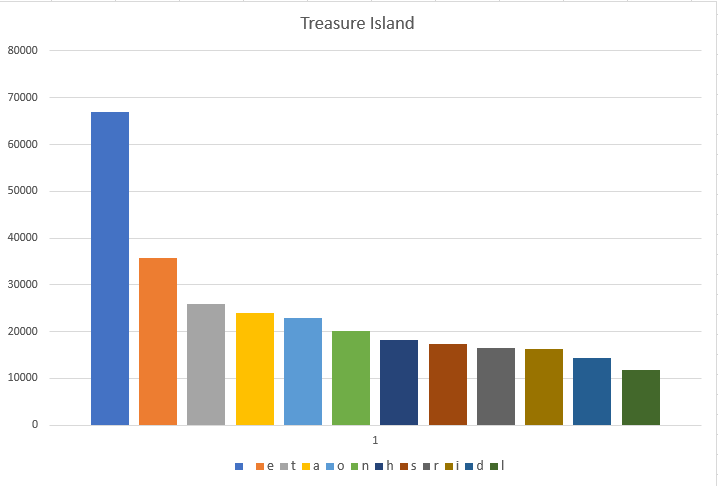
В случае с текстами на русском языке первые 12 символов в списке (самые часто встречающиеся) – одни и те же в обоих документах. Лишь их порядок незначительно меняется (а именно символы «н» «и» поменялись местами, и аналогично «в» «р»). Затем «м» «д» «у» «,» можно объединить в группу..

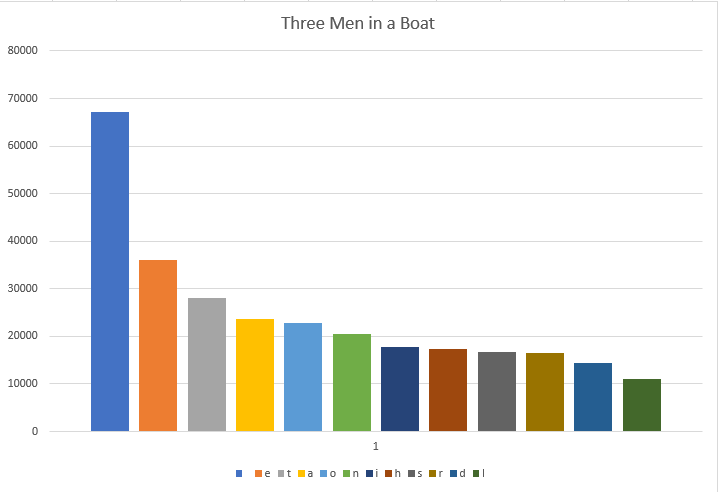
Соотношения частот не одинаковы, но наверняка попадают в одни и те же диапазоны, которые можно определить, собрав достаточную статистику.

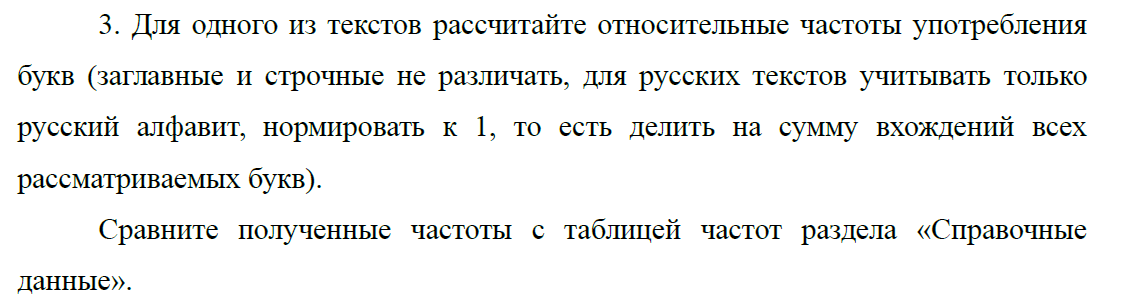
Гистограммы частот построим только для самых часто встречающихся символов (чтобы они были читабельны):

****

****

****

****

****

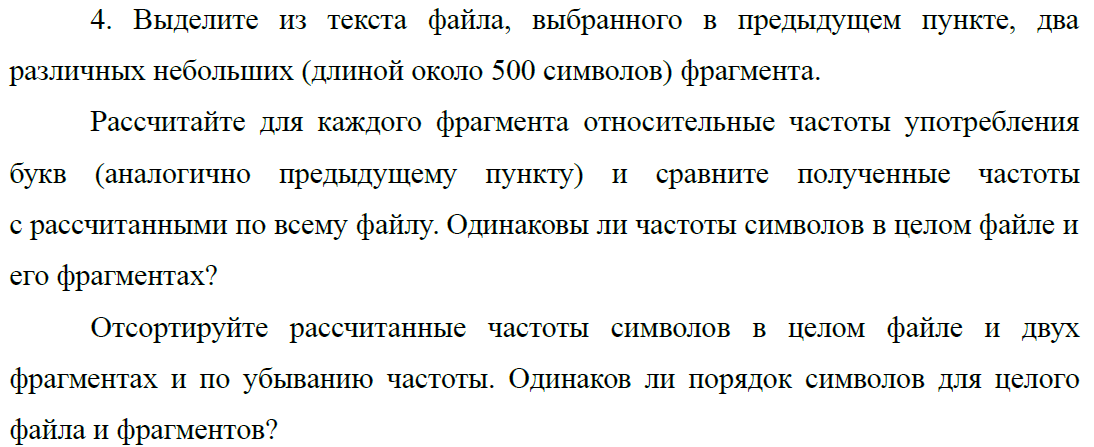
**#!/bin/bash**

**sed -r 's/[^A-Za-z]//g' $1 | sed 's/[A-Za-z]/\L&\n/g' | sort | uniq -c | tail -n +2 | sort -nr > intfreq**

**awk 'FNR==NR{s+=$1;next} {print $1/s, $2}' intfreq intfreq > rel\_freq**

|  |
| --- |
| Целый файл |
| 0.120706 e |
| 0.0907332 t |
| 0.084819 a |
| 0.0789805 o |
| 0.0692999 n |
| 0.0648703 i |
| 0.0638016 h |
| 0.0606469 s |
| 0.0562997 r |
| 0.0487739 d |
| 0.0411896 l |
| 0.0299282 u |
| 0.0258732 w |
| 0.023911 m |
| 0.0222512 c |
| 0.0209831 f |
| 0.0206773 g |
| 0.0204298 y |
| 0.0169865 p |
| 0.0164435 b |
| 0.0100689 k |
| 0.0083197 v |
| 0.00172511 j |
| 0.00101376 q |
| 0.00094503 x |
| 0.000323028 z |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ранг** | **буква** | **употреблений** | **частотность** | **ранг** | **буква** | **употреблений** | **частотность** |
| 1 | **E** | 21912 | 12.02 | 13 | **C** | 4943 | 2.71 |
| 2 | **T** | 16587 | 9.10 | 14 | **M** | 4761 | 2.61 |
| 3 | **A** | 14810 | 8.12 | 15 | **F** | 4200 | 2.30 |
| 4 | **O** | 14003 | 7.68 | 16 | **Y** | 3853 | 2.11 |
| 5 | **I** | 13318 | 7.31 | 17 | **W** | 3819 | 2.09 |
| 6 | **N** | 12666 | 6.95 | 18 | **G** | 3693 | 2.03 |
| 7 | **S** | 11450 | 6.28 | 19 | **P** | 3316 | 1.82 |
| 8 | **R** | 10977 | 6.02 | 20 | **B** | 2715 | 1.49 |
| 9 | **H** | 10795 | 5.92 | 21 | **V** | 2019 | 1.11 |
| 10 | **D** | 7874 | 4.32 | 22 | **K** | 1257 | 0.69 |
| 11 | **L** | 7253 | 3.98 | 23 | **X** | 315 | 0.17 |
| 12 | **U** | 5246 | 2.88 | 24 | **Q** | 205 | 0.11 |
| 13 | **C** | 4943 | 2.71 | 25 | **J** | 188 | 0.10 |
| 14 | **M** | 4761 | 2.61 | 26 | **Z** | 128 | 0.07 |

****

**Фрагмент 1**

**awk 'NR >= 1000 && NR <= 1500' $1 | sed -r 's/[^A-Za-z]//g' | sed 's/[A-Za-z]/\L&\n/g' | sort | uniq -c | tail -n +2 | sort -nr > intfreq**

**awk 'FNR==NR{s+=$1;next} {print $1/s, $2}' intfreq intfreq > rel\_freq\_fragment1**

**Фрагмент 2**

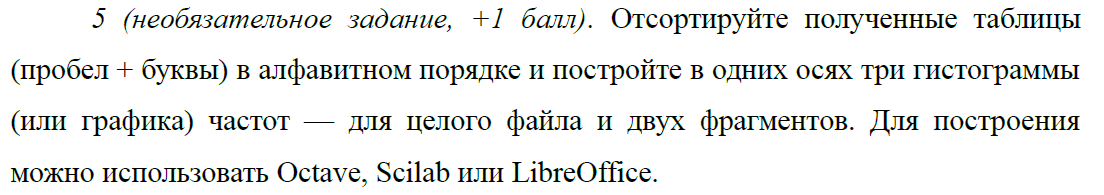
**awk 'NR >= 3000 && NR <= 3500' $1 | sed -r 's/[^A-Za-z]//g' | sed 's/[A-Za-z]/\L&\n/g' | sort | uniq -c | tail -n +2 | sort -nr > intfreq**

**awk 'FNR==NR{s+=$1;next} {print $1/s, $2}' intfreq intfreq > rel\_freq\_fragment2**

|  |  |  |
| --- | --- | --- |
| **Целый файл** | **Фрагмент 1** | **Фрагмент 2** |
| 0.120706 e | 0.120299 e | 0.12595 e |
| 0.0907332 t | 0.0887398 t | 0.0922006 t |
| 0.084819 a | 0.0874753 a | 0.0837344 a |
| 0.0789805 o | 0.0757642 o | 0.0771238 o |
| 0.0692999 n | 0.0731251 n | 0.0669759 n |
| 0.0648703 i | 0.0664724 i | 0.0658742 h |
| 0.0638016 h | 0.0639982 h | 0.0651203 i |
| 0.0606469 s | 0.0614691 s | 0.0607133 s |
| 0.0562997 r | 0.0519024 r | 0.0600174 r |
| 0.0487739 d | 0.0489883 d | 0.0519571 d |
| 0.0411896 l | 0.0424456 l | 0.0398956 l |
| 0.0299282 u | 0.0306246 u | 0.0280081 u |
| 0.0258732 w | 0.0273257 m | 0.0242969 m |
| 0.023911 m | 0.0255663 w | 0.0234271 w |
| 0.0222512 c | 0.0211678 y | 0.0219774 c |
| 0.0209831 f | 0.0210578 f | 0.0211076 f |
| 0.0206773 g | 0.020563 c | 0.0200058 g |
| 0.0204298 y | 0.0162745 b | 0.0179762 y |
| 0.0169865 p | 0.0162745 b | 0.0166425 p |
| 0.0164435 b | 0.01479 p | 0.0157727 b |
| 0.0100689 k | 0.0111062 k | 0.00927805 k |
| 0.0083197 v | 0.00720255 v | 0.00817628 v |
| 0.00172511 j | 0.00241918 j | 0.00179762 q |
| 0.00101376 q | 0.00109963 x | 0.00127573 j |
| 0.00094503 x | 0.000934682 q | 0.000637866 x |
| 0.000323028 z | 0.000164944 z | 5.79878e-05 z |

***Одинаков ли порядок символов для целого файла и фрагментов?***

Небольшие отличия в порядке следования букв заметны на самих редко используемых буквах.

****

**awk 'NR >= 3000 && NR <= 3500' $1 | sed -r 's/[^A-Za-z]//g' | sed 's/[A-Za-z]/\L&\n/g' | sort | uniq -c | tail -n +2 > intfreq**

**awk 'FNR==NR{s+=$1;next} {print $1/s, $2}' intfreq intfreq > rel\_freq\_fragment2**

**sed -i 's/.$//' rel\_freq\_fragment2**

x = 0:25;

a = [ 0.084819 0.0164435 0.0222512 0.0487739 0.120706 0.0209831 0.0206773 0.0638016 0.0648703 0.00172511 0.0100689 0.0411896 0.023911 0.0692999 0.0789805 0.0169865 0.00101376 0.0562997 0.0606469 0.0907332 0.0299282 0.0083197 0.0258732 0.00094503 0.0204298 0.000323028 ];

b = [ 0.0837344 0.0157727 0.0219774 0.0519571 0.12595 0.0211076 0.0200058 0.0658742 0.0651203 0.00127573 0.00927805 0.0398956 0.0242969 0.0669759 0.0771238 0.0166425 0.00179762 0.0600174 0.0607133 0.0922006 0.0280081 0.00817628 0.0234271 0.000637866 0.0179762 5.79878e-05 ];

c = [ 0.0874753 0.0162745 0.020563 0.0489883 0.120299 0.0210578 0.0190235 0.0639982 0.0664724 0.00241918 0.0111062 0.0424456 0.0273257 0.0731251 0.0757642 0.01479 0.000934682 0.0519024 0.0614691 0.0887398 0.0306246 0.00720255 0.0255663 0.00109963 0.0211678 0.000164944 ];

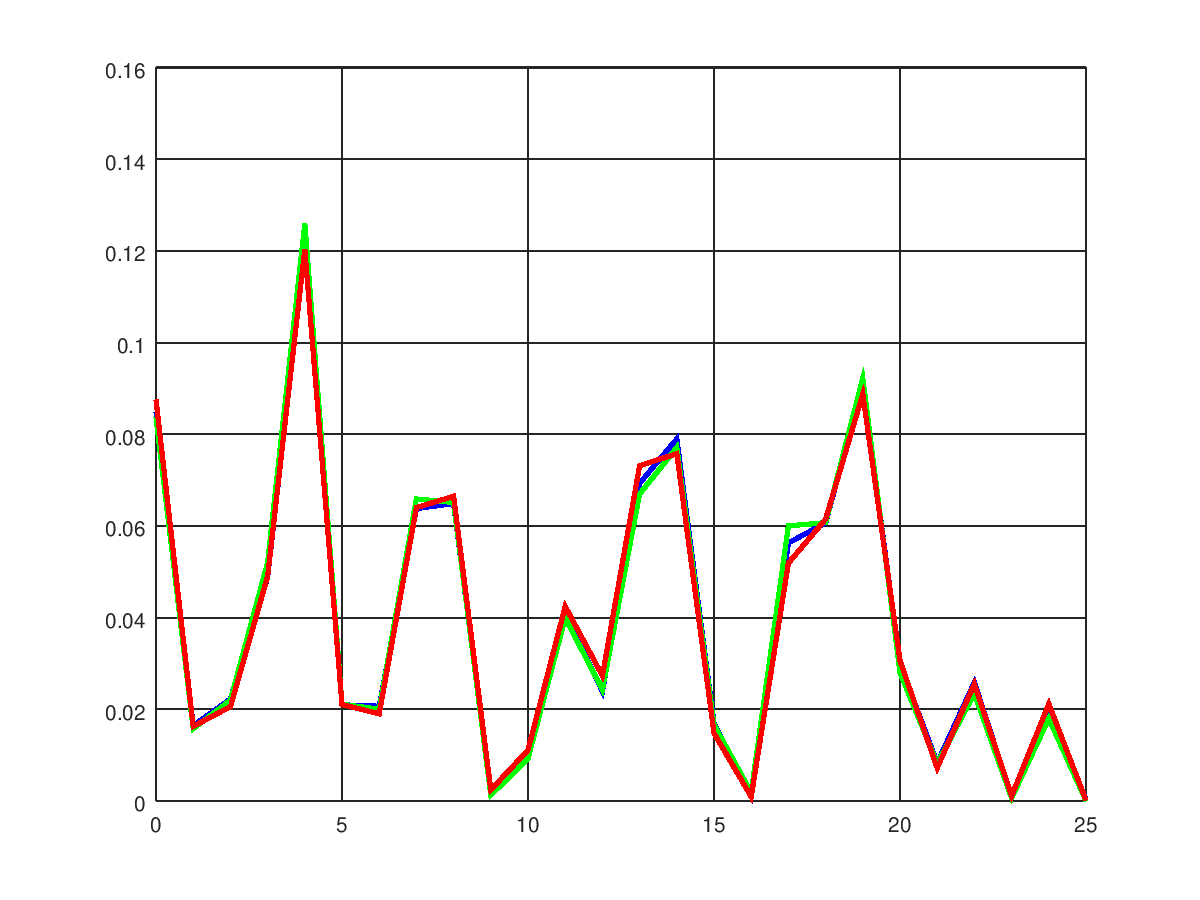
hold on

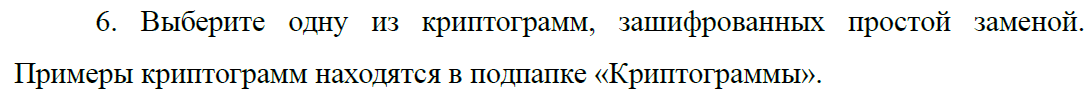
grid on

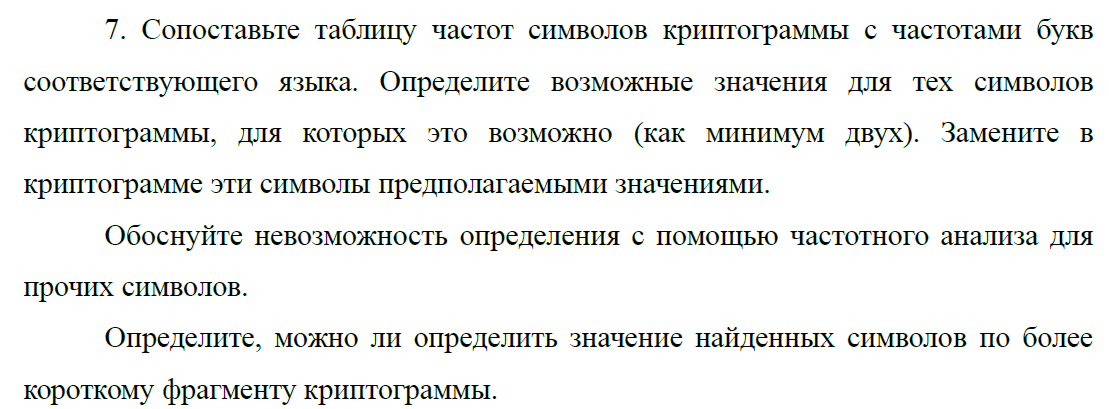
plot(x,a,'b','Linewidth',2)

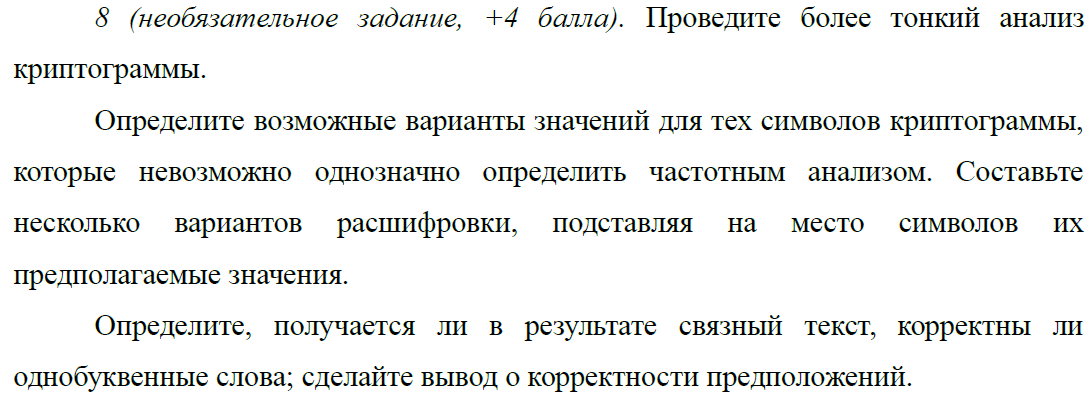
plot(x,b,'g','Linewidth',2)

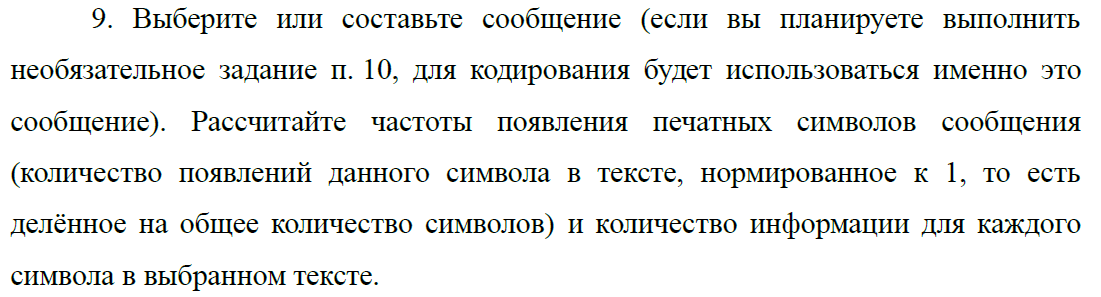
plot(x,c,'r','Linewidth',2)



****

****

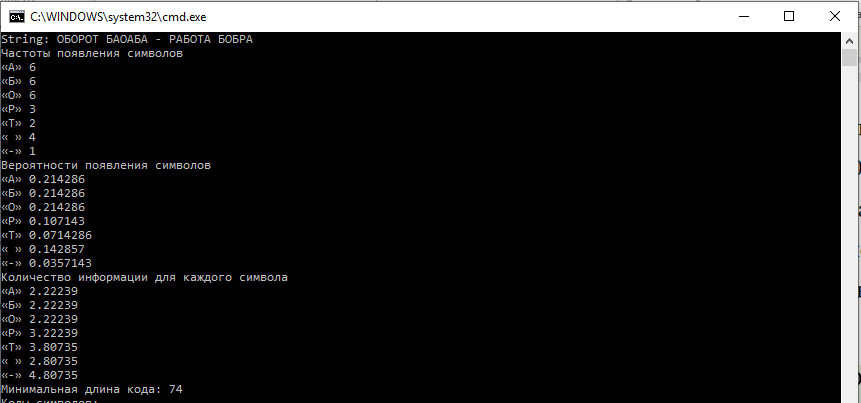
****

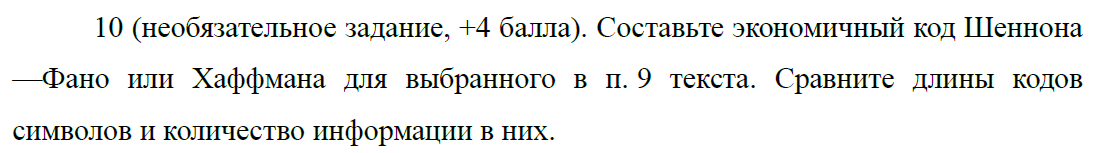
****

Сообщение (взято из кр№1 по ОТИК): ОБОРОТ БАОАБА – РАБОТА БОБРА

Программа Shannon-Fano производит расчет частот, вероятностей и кол-ва информации

Результат:

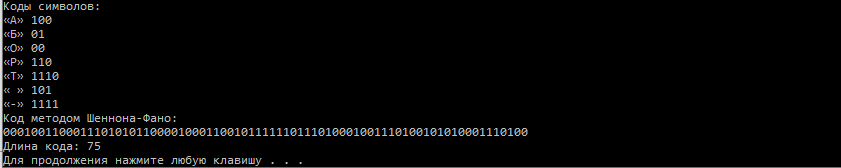


****

Сообщение (взято из кр№1 по ОТИК): ОБОРОТ БАОАБА – РАБОТА БОБРА

Программа Shannon-Fano производит расчет частот, вероятностей и кол-ва информации

Результат:



**Код Методом Хаффмана**

|  |  |  |
| --- | --- | --- |
| Буква | Частота | Код |
| О | 6 | 00 |
| Б | 6 | 01 |
| Р | 3 | 1111 |
| Т | 2 | 11101 |
|  | 4 | 110 |
| А | 6 | 10 |
| - | 1 | 11100 |

