ПИН-32, команда 5: Надёжкина Анна, Трусов Михаил

Отчет

Лабораторная работа 2

Исследование статистических характеристик исходных текстов (как бинарных файлов, так и файлов в формате простого текста). Работа с кодовыми таблицами русского языка

Операционная система ОС Windows, 64-разрядная ОС

Задание Л2.1.

Разработайте программу или используйте набор программ (в частности, это может быть набор скриптов - однострочников, использующий стандартные утилиты GNU/Linux), который по заданному файлу Q рассчитывает:

— длину n файла Q в символах первичного алфавита A_1;

— $count(a_j)$ — общее количество вхождений каждого из символов $a_j \in A1$ в Q (ненормированную целочисленную частоту $\nu_{\text{ненорм}}(a_j)$);

и оценивает, строя модель источника символов по файлу $Q = c1 \ldots cn$ в виде источника без памяти (модель для сжатия без учёта контекста):

— вероятность $p_{\text{БП}}(a_j)$ каждого из символов

— количество информации $I_{\text{БП}}(a_j)$ = [бит] в каждом $a_j \in A_1$;

— суммарное количество информации $I_{\text{БП}}$

Сравните:

а) длину файла Q в битах $n \cdot 8$ и оценку IБП(Q) в битах

б) длину файла Q в октетах n и оценку в октетах:

а также оценки снизу длин в октетах для сжатия без учёта контекста:

— длины E = [IБП(Q) [октетов]] только сжатого текста, без информации, необходимой для его декодирования;

— длины $G_{64} = E + 256 \cdot 8$ архива, где к сжатому тексту добавляется таблица (ν(00), . . . , ν(FF)) из 256 ненормированных 64-битных частот $\nu_{\text{ненорм}}(j) = count(j)$;

— длины $G_8 = E + 256 \cdot 1$ архива, где к сжатому тексту добавляется таблица из 256 нормированных 8-битных частот $\nu_{\text{норм}}(j)$.

Все полученные величины необходимо напечатать на экране или сохранить в виде текстового файла-отчёта; причём таблица характеристик символов алфавита $j \in A_1$ должна либо печататься дважды:

— отсортированной по алфавиту (по значению j);

— отсортированной по убыванию $count(j)$;

либо в программе необходимо предусмотреть пересортировку.

Код (https://pastebin.com/YEF3mSc7):

```
import math
from collections import Counter
from math import log2

import pandas as pd

FILE_PATH = r"C:\Users\1\Downloads\otik-master\otik-master\labs-files\Файлы в формате простого текста — utf8\The Secret Adversary, by Agatha Christie.txt"
USE_UNICODE = False
FILE_OPEN_MODE = {'mode': 'r', 'encoding': 'mac-cyrillic'} if USE_UNICODE else {'mode': 'rb'}


def file_length(file_path: str) -> int:
    """
    Вычисление длины n файла Q в символах первичного алфавита A1
    :param file_path: путь до файла
    :return: длина n файла
    """
    with open(file_path, **FILE_OPEN_MODE) as file:
        return len(file.read())


def symbol_amounts(file_path: str) -> Counter:
    """
    Вычисление общего количество вхождений каждого из символов в файл (ненормированную целочисленную частоту)
    :param file_path: путь до файла
    :return: объект Counter с общим количеством вхождений
    """
    with open(file_path, **FILE_OPEN_MODE) as file:
        return Counter(file.read())


def total_information(sym_amounts: Counter) -> float:
    """
    Вычисление суммарного количества информации в файле
    :param sym_amounts: объект Counter с общим количеством вхождений
    :return: суммарное количество информации в файле
    """
    total_symbols = sum(sym_amounts.values())
    total_info_bits = sum(
        symbol_information(probability(amount, total_symbols)) * amount for amount in sym_amounts.values())
```

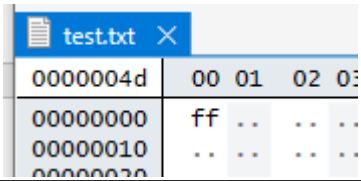
```
    return total_info_bits


def probability(symbol_frequency: int, total_symbols: int) -> float:
    """
    Вычисление вероятности символа в файле
    :param symbol_frequency: частота символа в файле
    :param total_symbols: число символов в файле
    :return: вероятность символа в сообщении
    """
    return symbol_frequency / total_symbols


def symbol_information(symbol_probability: float) -> float:
    """
    Вычисление количества информации по частоте символа в файле
    :param symbol_probability: частота символа в файле
    :return: количество информации в символе
    """
    return -log2(symbol_probability)


def display_tables(sym_amounts: Counter) -> None:
    """
    Вывод таблиц с информацией о символах файла
    :param sym_amounts: объект Counter с общим количеством вхождений
    """
    total_symbols = sum(sym_amounts.values())
    df = pd.DataFrame([(
        repr(symbol) if USE_UNICODE else f'{hex(symbol)[2:]:0>2}',
        freq,
        probability(freq, total_symbols),
        symbol_information(probability(freq, total_symbols))) for symbol, freq in sym_amounts.items()],
        columns=["Символ", "Частота", "Вероятность", "Количество информации"])

    print('\nОтсортировано по алфавиту:')
    print(df.sort_values("Символ").to_string(index=False))
    print('\nОтсортировано по частоте:')
    print(df.sort_values("Частота", ascending=False).to_string(index=False))


sym_amounts = symbol_amounts(FILE_PATH)

print(f'Длина файла Q в битах: {file_length(FILE_PATH) * 8}')
print(f'Суммарное количество информации в файле в битах: {total_information(sym_amounts):.2f}')
print(f'Дробная часть в экспон. форме: {total_information(sym_amounts) -
int(total_information(sym_amounts)):.2e}')

print(f'\nДлина файла Q в октетах: {file_length(FILE_PATH)}')
print(f'Суммарное количество информации в файле в октетах: {total_information(sym_amounts) / 8:.2f}')
E = math.ceil(total_information(sym_amounts) / 8)
print(f'Длина только сжатого текста, без информации, необходимой для его декодирования в октетах
(E): {E}')
print(
    f'Длина архива, где к сжатому тексту добавляется таблица из 256 ненормированных 64-битных
```

*частот в октетах (G64): {E + 256 * 8}')*
print(
*f'Длина архива, где к сжатому тексту добавляется таблица из 256 нормированных 8-битных частот в октетах (G8): {E + 256 * 1}')*

display_tables(sym_amounts)

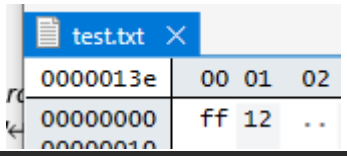
Для проверки корректности используйте файлы с заранее известным IБП(Q): для четырёх разных произвольных октетов a, b, c, d верно $I_{\text{БП}}(a)$ = 0 бит, $I_{\text{БП}}(ab)$ = 2 бита, $I_{\text{БП}}(abcd)$ = 4 · 2 = 8 бит (1 байт x86) и т.п.

1) $I_{\text{БП}}(a)$ = 0 бит



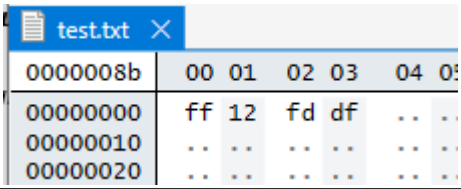

```
Длина файла Q в битах: 8
Суммарное количество информации в файле в битах: 0.00
Дробная часть в экспон. форме: 0.00e+00
```

2) $I_{\text{БП}}(ab)$ = 2 бита




```
Длина файла Q в битах: 16
Суммарное количество информации в файле в битах: 2.00
Дробная часть в экспон. форме: 0.00e+00
```

3) $I_{\text{БП}}(abcd)$ = 8 бит




```
C:\Users\1\AppData\Local\Programs\Python\Python311\python.exe D:/pyprj/transport-bot/main.py
Длина файла Q в битах: 32
Суммарное количество информации в файле в битах: 8.00
Дробная часть в экспон. форме: 0.00e+00
```

Проверьте разработанную программу на файлах различного формата (не только простом тексте; в том числе и на бинарных).

1) Бинарный файл: otik-master\labs-files\Файлы в разных форматах\hexdump

```
Длина файла Q в битах: 208992
Суммарное количество информации в файле в битах: 164217.14
Дробная часть в экспон. форме: 1.41e-01

Длина файла Q в октетах: 26124
Суммарное количество информации в файле в октетах: 20527.14
Длина только сжатого текста, без информации, необходимой для его декодирования в октетах (E): 20528
Длина архива, где к сжатому тексту добавляется таблица из 256 ненормированных 64-битных частот в октетах (G64): 22576
Длина архива, где к сжатому тексту добавляется таблица из 256 нормированных 8-битных частот в октетах (G8): 20784
```

Отсортировано по алфавиту:

Символ	Частота	Вероятность	Количество информации
00	4730	0.181060	2.465464
01	436	0.016690	5.904904
02	191	0.007311	7.095659
03	87	0.003330	8.230145
04	1207	0.046203	4.435878
05	73	0.002794	8.483264
06	42	0.001608	9.280771
07	79	0.003024	8.369307
08	1112	0.042566	4.554147
09	34	0.001301	9.585625
0a	85	0.003254	8.263697
0b	77	0.002947	8.406302
0c	193	0.007388	7.080631
0d	30	0.001148	9.766198
0e	711	0.027216	5.199382
0f	350	0.013398	6.221877
10	280	0.010718	6.543805
11	23	0.000880	10.149526

Отсортировано по частоте:

Символ	Частота	Вероятность	Количество информации
00	4730	0.181060	2.465464
ff	1214	0.046471	4.427535
04	1207	0.046203	4.435878
08	1112	0.042566	4.554147
0e	711	0.027216	5.199382
20	616	0.023580	5.406302
01	436	0.016690	5.904904
83	424	0.016230	5.945168
74	387	0.014814	6.076898
24	354	0.013551	6.205483
0f	350	0.013398	6.221877
8b	301	0.011522	6.439469
5f	283	0.010833	6.528430
10	280	0.010718	6.543805
89	260	0.009953	6.650720
e8	254	0.009723	6.684404
73	237	0.009072	6.784345
8d	221	0.008460	6.885186
41	221	0.008460	6.885186
6e	215	0.008230	6.924895

2) Простой текст: otik-master\labs-files\Файлы в формате простого текста — utf8\The Secret Adversary, by Agatha Christie.txt

```
Длина файла Q в битах: 3563952
Суммарное количество информации в файле в битах: 2057918.11
Дробная часть в экспон. форме: 1.15e-01

Длина файла Q в октетах: 445494
Суммарное количество информации в файле в октетах: 257239.76
Длина только сжатого текста, без информации, необходимой для его декодирования в октетах (E): 257240
Длина архива, где к сжатому тексту добавляется таблица из 256 ненормированных 64-битных частот в октетах (G64): 259288
Длина архива, где к сжатому тексту добавляется таблица из 256 нормированных 8-битных частот в октетах (G8): 257496
```

Отсортировано по алфавиту:

Символ	Частота	Вероятность	Количество информации
0a	11303	0.025372	5.300628
20	70946	0.159252	2.650613
21	648	0.001455	9.425196
22	5624	0.012624	6.307666
23	1	0.000002	18.765046
24	2	0.000004	17.765046
25	1	0.000002	18.765046
27	2494	0.005598	7.480801
28	33	0.000074	13.720652
29	33	0.000074	13.720652
2a	28	0.000063	13.957692
2c	4590	0.010303	6.600768
2d	1965	0.004411	7.824733
2e	7462	0.016750	5.899700
2f	26	0.000058	14.064607
30	41	0.000092	13.407494

Отсортировано по частоте:

Символ	Частота	Вероятность	Количество информации
20	70946	0.159252	2.650613
65	41989	0.094253	3.407323
74	29304	0.065779	3.926236
6f	25420	0.057060	4.131370
61	24580	0.055175	4.179849
6e	21863	0.049076	4.348843
69	20047	0.044999	4.473948
73	19513	0.043801	4.512898
68	19300	0.043323	4.528733
72	18693	0.041960	4.574836
64	14356	0.032225	4.955680
6c	13311	0.029879	5.064715
75	11751	0.026377	5.244551
0a	11303	0.025372	5.300628

3) Простой текст: otik-master\labs-files\Файлы в формате простого текста — utf8\ Лев Николаевич Толстой. Война и мир 1.txt

```
Длина файла Q в битах: 10198496
Суммарное количество информации в файле в битах: 5394161.66
Дробная часть в экспон. форме: 6.57e-01

Длина файла Q в октетах: 1274812
Суммарное количество информации в файле в октетах: 674270.21
Длина только сжатого текста, без информации, необходимой для его декодирования в октетах (E): 674271
Длина архива, где к сжатому тексту добавляется таблица из 256 ненормированных 64-битных частот в октетах (G64): 676319
Длина архива, где к сжатому тексту добавляется таблица из 256 нормированных 8-битных частот в октетах (G8): 674527
```

Отсортировано по алфавиту:

Символ	Частота	Вероятность	Количество информации
0a	12275	0.010	6.698
20	106575	0.084	3.580
21	888	0.001	10.487
22	604	0.000	11.043
27	412	0.000	11.595
28	542	0.000	11.200
29	542	0.000	11.200
2a	3	0.000	18.697
2c	16162	0.013	6.302
2d	1147	0.001	10.118
2e	8389	0.007	7.248
30	184	0.000	12.758
31	430	0.000	11.534
32	295	0.000	12.077
33	192	0.000	12.697

Отсортировано по частоте:

Символ	Частота	Вероятность	Количество информации
d0	383048	3.004741e-01	1.734687
d1	157916	1.238740e-01	3.013055
20	106575	8.360056e-02	3.580344
be	60772	4.767134e-02	4.390734
b0	44485	3.489534e-02	4.840822
b5	42293	3.317587e-02	4.913722
b8	35358	2.773585e-02	5.172104
bd	34351	2.694593e-02	5.213789
82	30146	2.364741e-02	5.402174
81	28386	2.226681e-02	5.488961
80	28322	2.221661e-02	5.492218
bb	27120	2.127373e-02	5.554784
b2	23685	1.857921e-02	5.750167
ba	18742	1.470178e-02	6.087866
2c	16162	1.267795e-02	6.301535
b4	15837	1.242301e-02	6.330842
83	15334	1.202844e-02	6.377407
bc	15327	1.202295e-02	6.378065

Выгодно ли применять к проанализированным файлам сжатие без учёта контекста? Нужно ли нормировать частоты?

Сжатие без учёта контекста выгодно применять в следующих случаях:

• Небольшие файлы. Для очень маленьких файлов (например, несколько байт) затраты на анализ контекста могут быть больше, чем выигрыш от сжатия

• Ограниченные ресурсы. Сжатие без учёта контекста не требует сложных вычислений.

• Известный статистический состав данных.

• Файлы с низкой энтропией (файлы только из одной буквы)

Нормировка частот

Нормировка частот не является обязательной для статического сжатия, но она может улучшить эффективность сжатия в некоторых случаях. Нормировка позволяет более точно оценить вероятность появления символов и использовать более оптимальные коды для их представления.

Задание Л2.№2. Бонус +3 балла для пар и НБ, обязательное для ПИН -троек. Разработайте программу, аналогичную Л2.№1, но считающую символом кодирования печатный или управляющий символ Unicode, а первичным алфавитом A_1 — множество символов Unicode (строчных букв, заглавных букв, цифр, различных пробельных символов, знаков препинания и т. п) в файле Q.

1) Бинарный файл: otik-master\labs-files\Файлы в разных форматах\hexdump

```
Длина файла Q в битах: 208992
Суммарное количество информации в файле в битах: 164121.91
Дробная часть в экспон. форме: 9.14e-01

Длина файла Q в октетах: 26124
Суммарное количество информации в файле в октетах: 20515.24
Длина только сжатого текста, без информации, необходимой для его декодирования в октетах (E): 20516
Длина архива, где к сжатому тексту добавляется таблица из 256 ненормированных 64-битных частот в октетах (G64): 22564
Длина архива, где к сжатому тексту добавляется таблица из 256 нормированных 8-битных частот в октетах (G8): 20772
```

Отсортировано по алфавиту:

Символ	Частота	Вероятность	Количество информации
"'"	42	0.001608	9.280771
' '	616	0.023580	5.406302
'!'	9	0.000345	11.503163
'"'	124	0.004747	7.718892
'#'	10	0.000383	11.351160
'$'	354	0.013551	6.205483
'%'	142	0.005436	7.523341
'&'	102	0.003904	8.000663
'('	177	0.006775	7.205483
')'	154	0.005895	7.406302
'*'	49	0.001876	9.058378
'+'	23	0.000880	10.149526
','	36	0.001378	9.503163
'-'	64	0.002450	8.673088
'.'	77	0.002947	8.406302

Отсортировано по частоте:

Символ	Частота	Вероятность	Количество информации
'\x00'	4730	0.181060	2.465464
'€'	1214	0.046471	4.427535
'\x04'	1207	0.046203	4.435878
'\x08'	1112	0.042566	4.554147
'\x0e'	711	0.027216	5.199382
' '	616	0.023580	5.406302
'\x01'	436	0.016690	5.904904
'Г'	424	0.016230	5.945168
't'	387	0.014814	6.076898
'$'	354	0.013551	6.205483
'\x0f'	350	0.013398	6.221877
'Л'	301	0.011522	6.439469
'_'	283	0.010833	6.528430
'\x10'	280	0.010718	6.543805
'Й'	260	0.009953	6.650720

2) Простой текст: otik-master\labs-files\Файлы в формате простого текста — utf8\The Secret Adversary, by Agatha Christie.txt

```
Длина файла Q в битах: 3563952
Суммарное количество информации в файле в битах: 2057918.11
Дробная часть в экспон. форме: 1.15e-01

Длина файла Q в октетах: 445494
Суммарное количество информации в файле в октетах: 257239.76
Длина только сжатого текста, без информации, необходимой для его декодирования в октетах (E): 257240
Длина архива, где к сжатому тексту добавляется таблица из 256 ненормированных 64-битных частот в октетах (G64): 259288
Длина архива, где к сжатому тексту добавляется таблица из 256 нормированных 8-битных частот в октетах (G8): 257496
```

Отсортировано по алфавиту:

Символ	Частота	Вероятность	Количество информации
'"'	2494	0.005598	7.480801
' '	70946	0.159252	2.650613
'!'	648	0.001455	9.425196
'"'	5624	0.012624	6.307666
'#'	1	0.000002	18.765046
'$'	2	0.000004	17.765046
'%'	1	0.000002	18.765046
'('	33	0.000074	13.720652
')'	33	0.000074	13.720652
'*'	28	0.000063	13.957692
','	4590	0.010303	6.600768
'-'	1965	0.004411	7.824733
'.'	7462	0.016750	5.899700

Отсортировано по частоте:

Символ	Частота	Вероятность	Количество информации
' '	70946	0.159252	2.650613
'e'	41989	0.094253	3.407323
't'	29304	0.065779	3.926236
'o'	25420	0.057060	4.131370
'a'	24580	0.055175	4.179849
'n'	21863	0.049076	4.348843
'i'	20047	0.044999	4.473948
's'	19513	0.043801	4.512898
'h'	19300	0.043323	4.528733
'r'	18693	0.041960	4.574836
'd'	14356	0.032225	4.955680

3) Простой текст: otik-master\labs-files\Файлы в формате простого текста — utf8\ Лев Николаевич Толстой. Война и мир 1.txt

```
Длина файла Q в битах: 10198496
Суммарное количество информации в файле в битах: 5394161.66
Дробная часть в экспон. форме: 6.57e-01

Длина файла Q в октетах: 1274812
Суммарное количество информации в файле в октетах: 674270.21
Длина только сжатого текста, без информации, необходимой для его декодирования в октетах (E): 674271
Длина архива, где к сжатому тексту добавляется таблица из 256 ненормированных 64-битных частот в октетах (G64): 676319
Длина архива, где к сжатому тексту добавляется таблица из 256 нормированных 8-битных частот в октетах (G8): 674527
```

Отсортировано по алфавиту:

Символ	Частота	Вероятность	Количество информации
"'"	412	3.231849e-04	11.595353
' '	106575	8.360056e-02	3.580344
'!'	888	6.965733e-04	10.487437
'"'	604	4.737954e-04	11.043448
'('	542	4.251607e-04	11.199704
')'	542	4.251607e-04	11.199704
'*'	3	2.353288e-06	18.696891
','	16162	1.267795e-02	6.301535
'-'	1147	8.997405e-04	10.118203
'.'	8389	6.580578e-03	7.247570

Отсортировано по частоте:

Символ	Частота	Вероятность	Количество информации
' '	106575	0.147075	2.765375
'о'	60772	0.083866	3.575766
'а'	44485	0.061390	4.025854
'е'	42293	0.058365	4.098754
'и'	35358	0.048795	4.357136
'н'	34351	0.047405	4.398821
'т'	30146	0.041602	4.587206
'с'	28386	0.039173	4.673993
'л'	27118	0.037423	4.739922
'р'	24091	0.033246	4.910678

Какой будет длина в октетах этих данных для исследуемых файлов?

Простой текст: otik-master\labs-files\Файлы в формате простого текста — utf8\Лев Николаевич Толстой. Война и мир 1.txt

```
Длина файла Q в битах: 10198496
Суммарное количество информации в файле в битах: 5394161.66
Дробная часть в экспон. форме: 6.57e-01

Длина файла Q в октетах: 1274812
Суммарное количество информации в файле в октетах: 674270.21
Длина только сжатого текста, без информации, необходимой для его декодирования в октетах (E): 674271
Длина архива, где к сжатому тексту добавляется таблица из 256 ненормированных 64-битных частот в октетах (G64): 676319
Длина архива, где к сжатому тексту добавляется таблица из 256 нормированных 8-битных частот в октетах (G8): 674527
```

1. Преобразование бит в байты: 10198496 бит / 8 бит/байт = 1274812 байт

2. Преобразование байт в октеты: 1274812 байт = 1274812 октетов

Итого: Длина данных в исследуемом файле "Война и мир" будет 1274812 октетов.

Какой была бы длина в октетах, если бы сохранялись частоты не для символов файла $A1 \subseteq$ Unicode, а для всего Unicode (длина и состав алфавита постоянны, сохраняются только частоты ($\nu(0)$, ..., $\nu(|\text{Unicode}| - 1)$))?

• Размер Unicode: 1,114,112 символов

• Размер одного символа: 4 байта (предположим, мы используем integer)

• Итого: 1,114,112 символов * 4 байта/символ = 4,456,448 байт (4,35 МБ)

Исследуйте один и тот же длинный текстовый файл в кодировке UTF-8 программами Л2.№1 и Л2.№2. Как выбор первичного алфавита влияет на:

— оценку $I_{\text{БП}}(Q)$ без учёта контекста;

Количество информации, связанное с появлением символа в тексте. Она вычисляется как −∑ p(x) log2 p(x), где p(x) - вероятность появления символа x.

Большой алфавит: Вероятность появления каждого символа будет меньше. Это приведет к более высокой оценке $I_{\text{БП}}(Q)$, т.к. более редкие символы несут больше информации.

Маленький алфавит: Вероятность появления каждого символа будет больше. Это приведет к более низкой оценке $I_{\text{БП}}(Q)$, т.к. более частые символы несут меньше информации.

— оценку снизу длин архива для сжатия без учёта контекста?

Длина архива это минимальное количество бит, необходимых для записи сжатых данных. Оценка снизу не может быть меньше, чем $I_{\text{БП}}(Q)$ умноженное на количество символов.

Большой алфавит: $I_{\text{БП}}(Q)$ будет выше, что приведет к более высокой оценке снизу длин архива.

Маленький алфавит: $I_{\text{БП}}(Q)$ будет ниже, что приведет к более низкой оценке снизу длин архива.

Задание Л2.№3. (Вариант 5)

Рассчитайте, используя программу Л2.№1, частоты октетов в файлах, являющихся простым текстом в различных кодировках (папка labs-files/Файлы в формате простого текста — кодировки разные). Исследуйте несколько разных осмысленных русскоязычных текстов и все представленные кодировки. Определите 4 наиболее частых октета среди всех используемых и 4 наиболее частых октета, не являющихся кодами печатных символов ASCII. Обратите внимание на распределение октетов многобайтовых кодировок.

1) Льюис Кэрролл. Охота на Снарка — Кружков — dos

```
Отсортировано по частоте:
Символ  Частота  Вероятность  Количество информации
    20     5087     0.185380               2.431446
    ae     2046     0.074560               3.745455
    a5     1434     0.052258               4.258216
    a0     1429     0.052075               4.263255
    a8     1244     0.045334               4.463275
```

2) Льюис Кэрролл. Охота на Снарка — Кружков — iso

```
Отсортировано по частоте:
Символ  Частота  Вероятность  Количество информации
    20     5087     0.185380               2.431446
    de     2046     0.074560               3.745455
    d5     1434     0.052258               4.258216
    d0     1429     0.052075               4.263255
    d8     1244     0.045334               4.463275
```

3) Льюис Кэрролл. Охота на Снарка — Кружков — koi8r

```
Отсортировано по частоте:
Символ  Частота  Вероятность  Количество информации
    20     5087     0.185380               2.431446
    cf     2046     0.074560               3.745455
    c5     1434     0.052258               4.258216
    c1     1429     0.052075               4.263255
    c9     1244     0.045334               4.463275
```

4) Льюис Кэрролл. Охота на Снарка — Кружков — maccyrillic

```
Отсортировано по частоте:
Символ  Частота  Вероятность  Количество информации
    20     5087     0.185380              2.431446
    ee     2046     0.074560              3.745455
    e5     1434     0.052258              4.258216
    e0     1429     0.052075              4.263255
    e8     1244     0.045334              4.463275
```

5) Льюис Кэрролл. Охота на Снарка — Кружков — utf8

```
Отсортировано по частоте:
Символ  Частота  Вероятность  Количество информации
    d0    14200     0.300612              1.734026
    d1     5596     0.118466              3.077449
    20     5087     0.107691              3.215030
    be     2046     0.043314              4.529039
    b5     1434     0.030358              5.041800
```

6) Льюис Кэрролл. Охота на Снарка — Кружков — utf16

```
Отсортировано по частоте:
Символ  Частота  Вероятность  Количество информации
    04    19796     0.360701              1.471124
    00     7645     0.139299              2.843745
    20     5117     0.093236              3.422963
    3e     2046     0.037280              4.745455
    35     1434     0.026129              5.258216
```

7) Льюис Кэрролл. Охота на Снарка — Кружков — utf32

```
Отсортировано по частоте:
Символ  Частота  Вероятность  Количество информации
    00    62529     0.569647              0.811860
    04    19796     0.180344              2.471177
    20     5117     0.046617              4.423015
    3e     2046     0.018639              5.745508
    35     1434     0.013064              6.258269
```

8) Льюис Кэрролл. Охота на Снарка — Кружков — windows

```
Отсортировано по частоте:
Символ  Частота  Вероятность  Количество информации
    20     5087     0.185380              2.431446
    ee     2046     0.074560              3.745455
    e5     1434     0.052258              4.258216
    e0     1429     0.052075              4.263255
    e8     1244     0.045334              4.463275
```

Рассчитайте частоты октетов в файле, соответствующем варианту N mod 9 в папке labs-files/Варианты 2 — определение кодировки простого текста (далее — файл W).

5.txt:

Отсортировано по частоте:

Символ	Частота	Вероятность	Количество информации
d0	5829	0.276781	1.853185
d1	2353	0.111728	3.161932
20	1972	0.093637	3.416774
be	840	0.039886	4.647972
b5	739	0.035090	4.832787

```
def find_no_ascii(df: pd.DataFrame):
    print("Непечатные символы: ")
    df = df.sort_values("Частота")
    for index, row in df.iterrows():
        if row['Символ'][0] in ("0", "1"):
            print(row)
```

```
Непечатные символы:
Символ                  0d
Частота                103
Вероятность         0.004891
Количество информации    7.675717
Символ                  0a
Частота                103
Вероятность         0.004891
Количество информации    7.675717
```

для разделения строк используется сочетание кодов возврата каретки (CARRIAGE RETURN) и перевода строки (LINE FEED) — $0D_{16}$ + $0A_{16}$
https://ru.wikipedia.org/wiki/%D0%A3%D0%BF%D1%80%D0%B0%D0%B2%D0%BB%D1%8F%D1%8E%D1%89%D0%B8%D0%B5_%D1%81%D0%B8%D0%BC%D0%B2%D0%BE%D0%BB%D1%8B

Определите, является ли W простым русскоязычным текстом в одной из стандартных кодировок (один из вариантов представляет собой нерусскоязычный текст); если да — определите кодировку.

Определить, является ли файл простым русскоязычным текстом в одной из стандартных кодировок, можно с помощью следующих шагов:

UTF-8: Большинство байтов в диапазоне 192-255

http://blog.kislenko.net/show.php?id=2045

```
def is_russian(sym_amounts):
    rus = 0
    for item in sym_amounts.keys():
        if 192 <= item <= 255:
            print(item, sym_amounts[item])
            rus += sym_amounts[item]
    print("Байтов в диапазоне 192-255: ", rus)
```

```
Байтов в диапазоне 192-255:  8312
```