Вариант 5

Файл 4

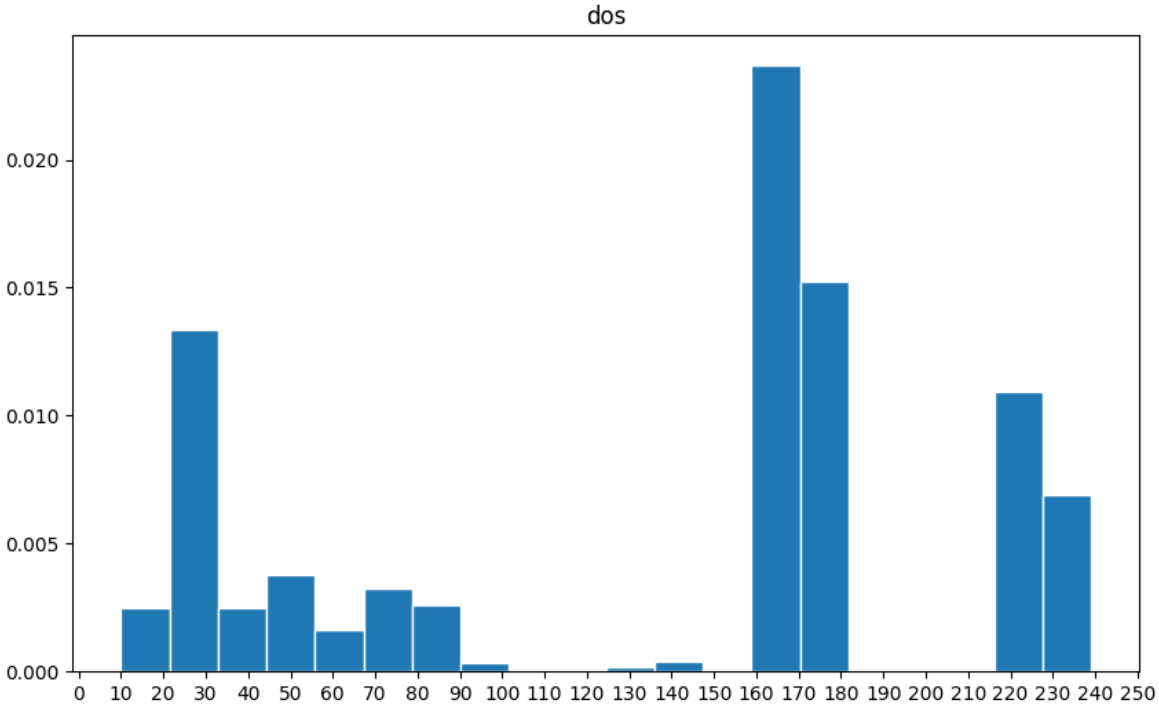
In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
from collections import Counter
import io
```

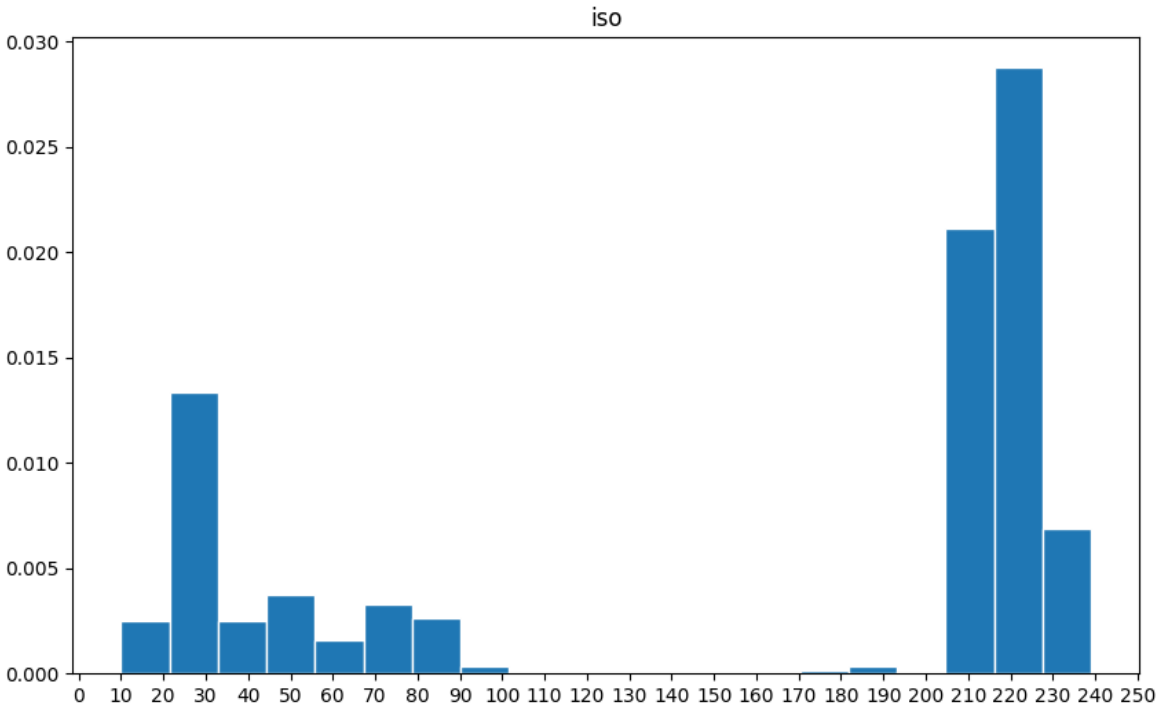
In [2]:
```
def print_most_freq_symbols(text):
    return max(set(text), key = text.count)

def print_most_freq_bytes(freqs, n):
    freq = freqs.copy()
    tmp = sorted(freq)
    print(f"{n} most common bytes:")
    for i in range(n):
        index = freq.index(tmp[-1 - i])
        percent = round(tmp[-1 - i] / sum(freqs) * 100, 2)
        if percent > 0:
            print(index, f"({percent}%)", end=" ")
            freq[index] = 0
    print("\n")

    freq = freqs[:32] + [freqs[127]]
    tmp = sorted(freq)
    print(f"{n} most common bytes of nonprintable ASCII table:")
    for i in range(n):
        index = freq.index(tmp[-1 - i])
        percent = round(tmp[-1 - i] / sum(freqs) * 100, 2)
        if percent > 0:
            print(index, f"({percent}%)", end=" ")
            freq[index] = 0
    print()

def get_byte_freqs(text):
    freqs = [0] * 256
    l = []
    for i in range(len(text)):
        freqs[text[i]] += 1
        l.append(text[i])

    return freqs

def plot_freqs(freqs, bin_number, title):
    l = []
    for i in range(len(freqs)):
        l += [i] * freqs[i]
    plt.figure(figsize=(10, 6))
    plt.xticks(np.arange(0, 256, 10))
    plt.hist(l, edgecolor="white", bins=bin_number, density=True)
    plt.title(title)
```

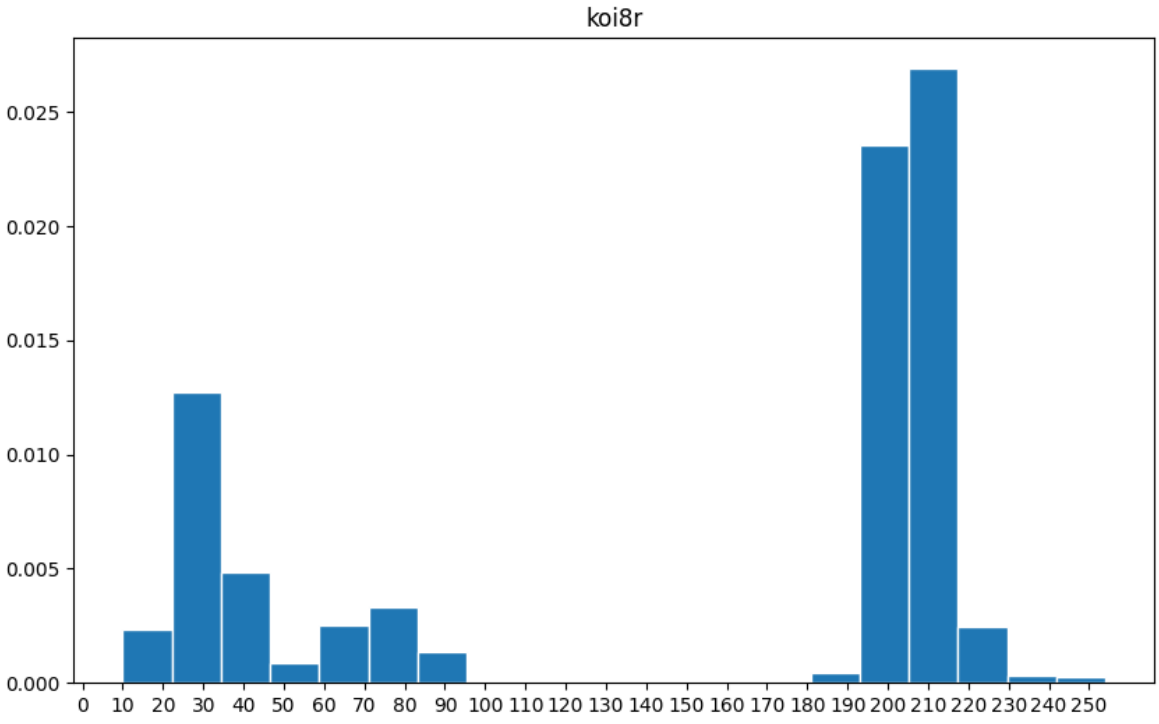
In [3]:
```
template = "Керниган, Ричи. Язык С – "
extentions = ["dos", "iso", "koi8r", "maccyrillic", "utf8", "utf16", "utf

N = 5
```

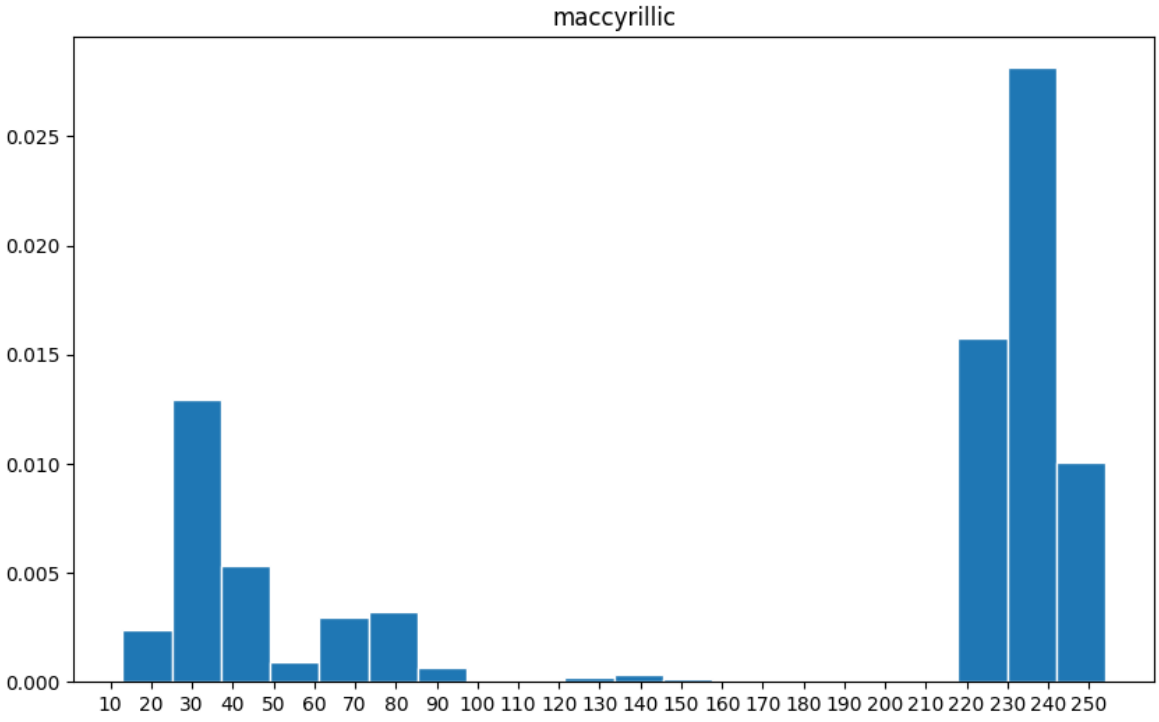
```
for ext in extentions:
    print("Encoding:", ext)
    byte_file = open(template + ext + ".txt", "rb")
    byte_text = byte_file.read()
    byte_file.close()

    byte_freqs = get_byte_freqs(byte_text)
    print_most_freq_bytes(byte_freqs, N)
    plot_freqs(byte_freqs, 20, ext)
    print("-" * 50)

print("Most common symbols in text:")
file = open(template + "utf8" + ".txt", "r")
text = file.read()
file.close()
counter = Counter(text)
[print(f"\"{counter.most_common(N)[i][0]}\" ({round(counter.most_common(N
```

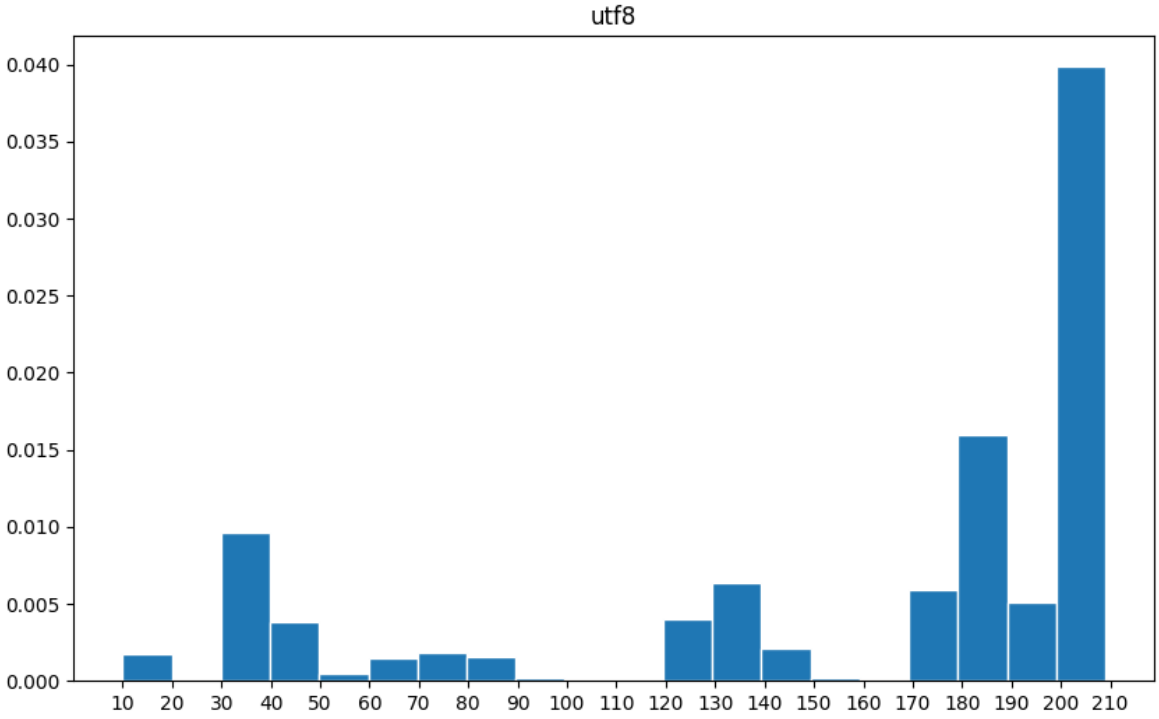
```
Encoding: dos
5 most common bytes:
32 (15.27%) 174 (6.4%) 165 (5.96%) 168 (5.27%) 160 (5.21%)

5 most common bytes of nonprintable ASCII table:
10 (2.85%)
--------------------------------------------------
Encoding: iso
5 most common bytes:
32 (15.27%) 222 (6.4%) 213 (5.96%) 216 (5.27%) 208 (5.21%)

5 most common bytes of nonprintable ASCII table:
10 (2.85%)
--------------------------------------------------
Encoding: koi8r
5 most common bytes:
32 (15.27%) 207 (6.4%) 197 (5.96%) 201 (5.27%) 193 (5.21%)

5 most common bytes of nonprintable ASCII table:
10 (2.85%)
--------------------------------------------------
Encoding: maccyrillic
5 most common bytes:
32 (15.27%) 238 (6.4%) 229 (5.96%) 232 (5.27%) 224 (5.21%)

5 most common bytes of nonprintable ASCII table:
13 (2.85%)
--------------------------------------------------
Encoding: utf8
5 most common bytes:
208 (27.33%) 209 (12.32%) 32 (9.21%) 190 (3.86%) 181 (3.6%)

5 most common bytes of nonprintable ASCII table:
10 (1.72%)
--------------------------------------------------
Encoding: utf16
5 most common bytes:
4 (32.86%) 0 (17.14%) 32 (7.64%) 62 (3.23%) 53 (3.0%)

5 most common bytes of nonprintable ASCII table:
4 (32.86%) 0 (17.14%) 10 (1.43%) 18 (0.05%) 31 (0.04%)
--------------------------------------------------
Encoding: utf32
5 most common bytes:
0 (58.57%) 4 (16.43%) 32 (3.82%) 62 (1.62%) 53 (1.5%)

5 most common bytes of nonprintable ASCII table:
0 (58.57%) 4 (16.43%) 10 (0.71%) 18 (0.02%) 31 (0.02%)
--------------------------------------------------
Encoding: windows
5 most common bytes:
32 (15.27%) 238 (6.4%) 229 (5.96%) 232 (5.27%) 224 (5.21%)

5 most common bytes of nonprintable ASCII table:
10 (2.85%)
--------------------------------------------------
Most common symbols in text:
" " (15.27%)   "о" (6.4%)   "е" (5.96%)   "и" (5.27%)   "а" (5.21%)
```

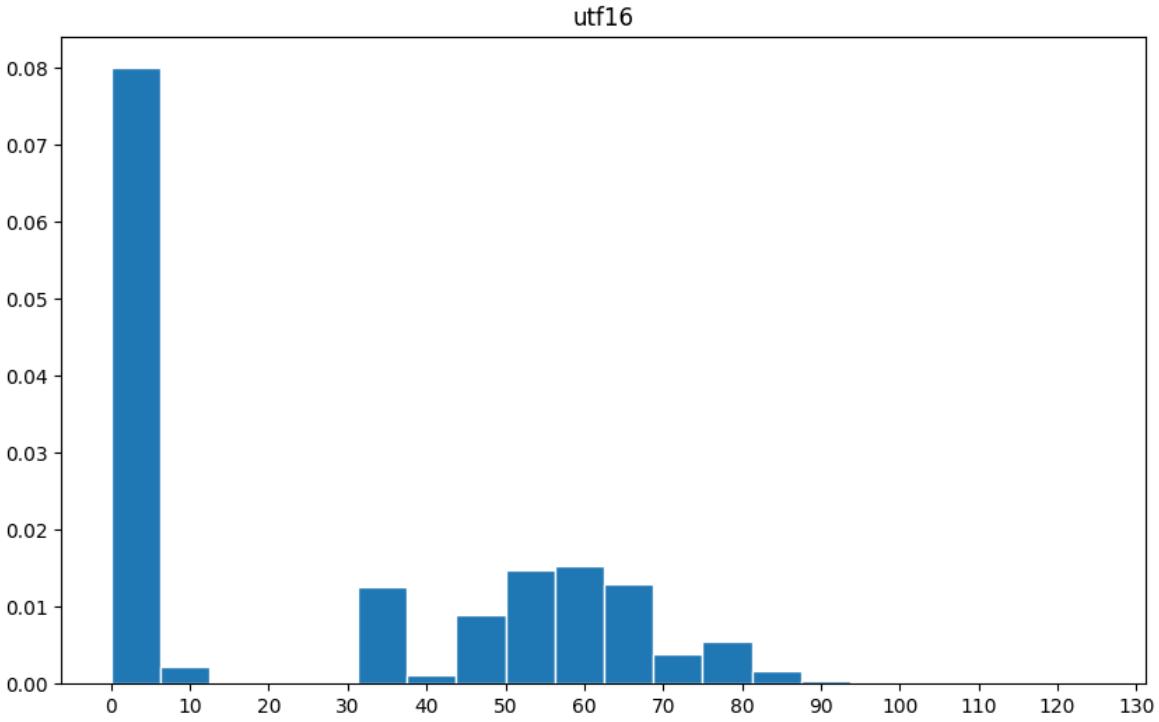

dos
0.020
0.015
0.010
0.005
0.000
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250

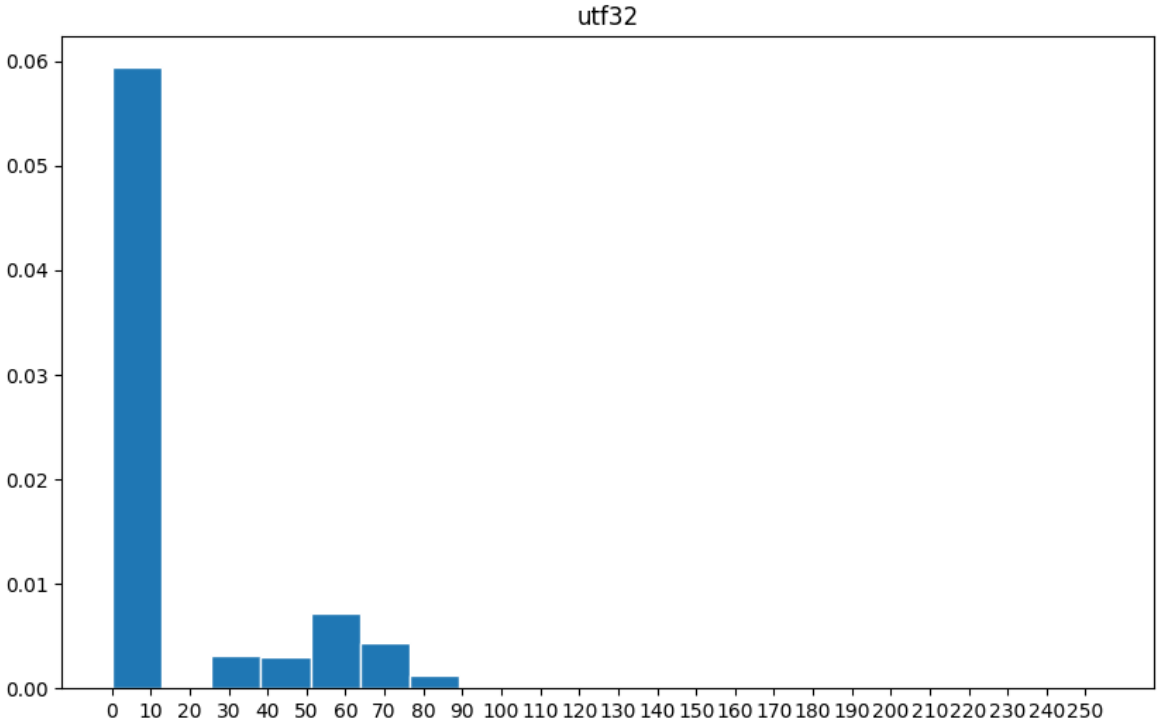

iso
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250

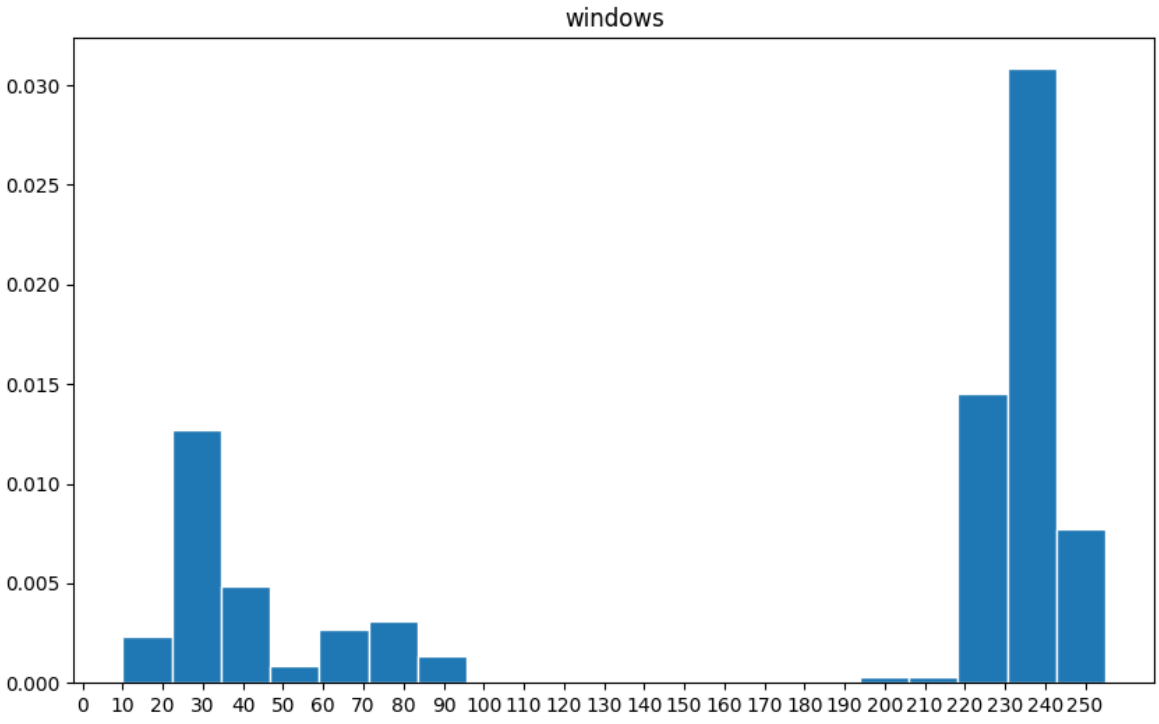

koi8r
0.025
0.020
0.015
0.010
0.005
0.000
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250


maccyrillic
0.025
0.020
0.015
0.010
0.005
0.000
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250


utf8
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210


utf16
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0 10 20 30 40 50 60 70 80 90 100 110 120 130





В многобайтовых кодировках можно увидеть большое преобладание байтов, которые в ASCII отвечают за непечатные символы. В принципе, оценка частот октетов таких кодировок нам мало сможет дать информации.

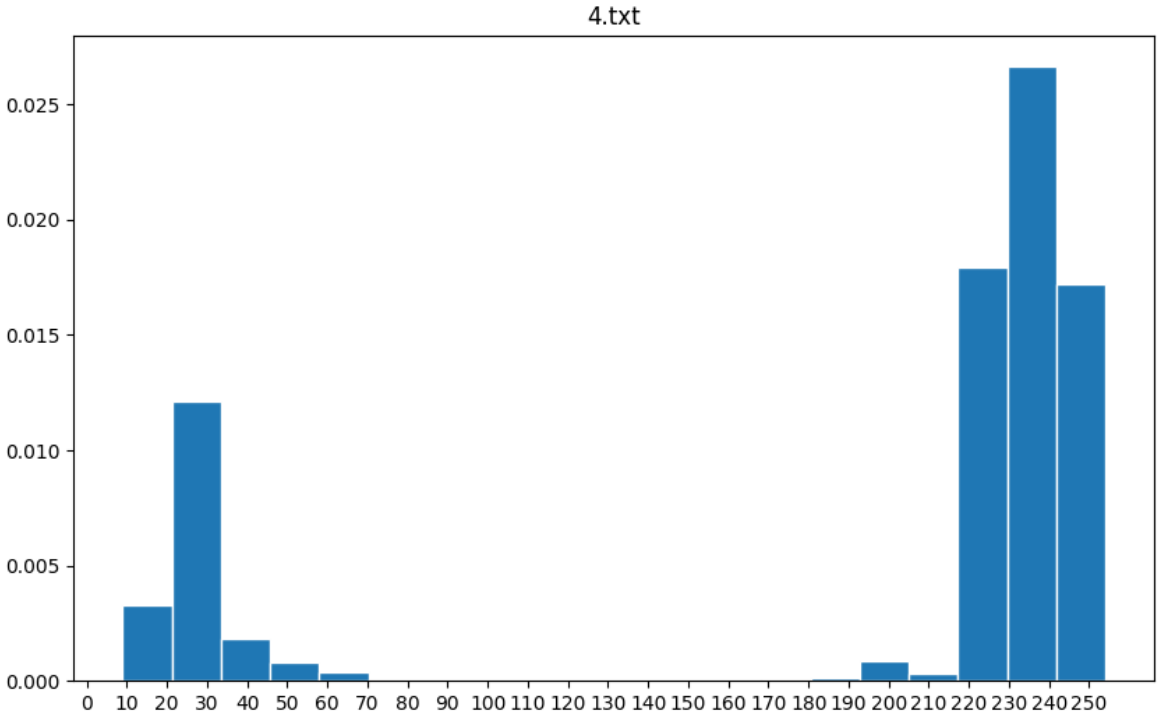
Для однобайтовых кодировок сделал проверку. Например:

1. В dos: 174_{10} - "о", 165_{10} - "е", 168_{10} - "и", 160_{10} - "а";
2. В windows-1251: 238_{10} - "о", 229_{10} - "е", 232_{10} - "и", 224_{10} - "а".

Что в точности совпадает с самыми распространенными буквами в рассмотренном тексте. Для остальных кодировок проверку опущу.

In [4]:
```
file = open('4.txt','rb')
text = file.read()
file.close()

freqs = get_byte_freqs(text)
print_most_freq_bytes(freqs, N)
plot_freqs(freqs, 20, "4.txt")
```

```
5 most common bytes:
32 (14.8%) 225 (7.93%) 239 (6.58%) 244 (6.37%) 233 (4.79%)

5 most common bytes of nonprintable ASCII table:
10 (1.92%) 13 (1.92%) 9 (0.18%)
```



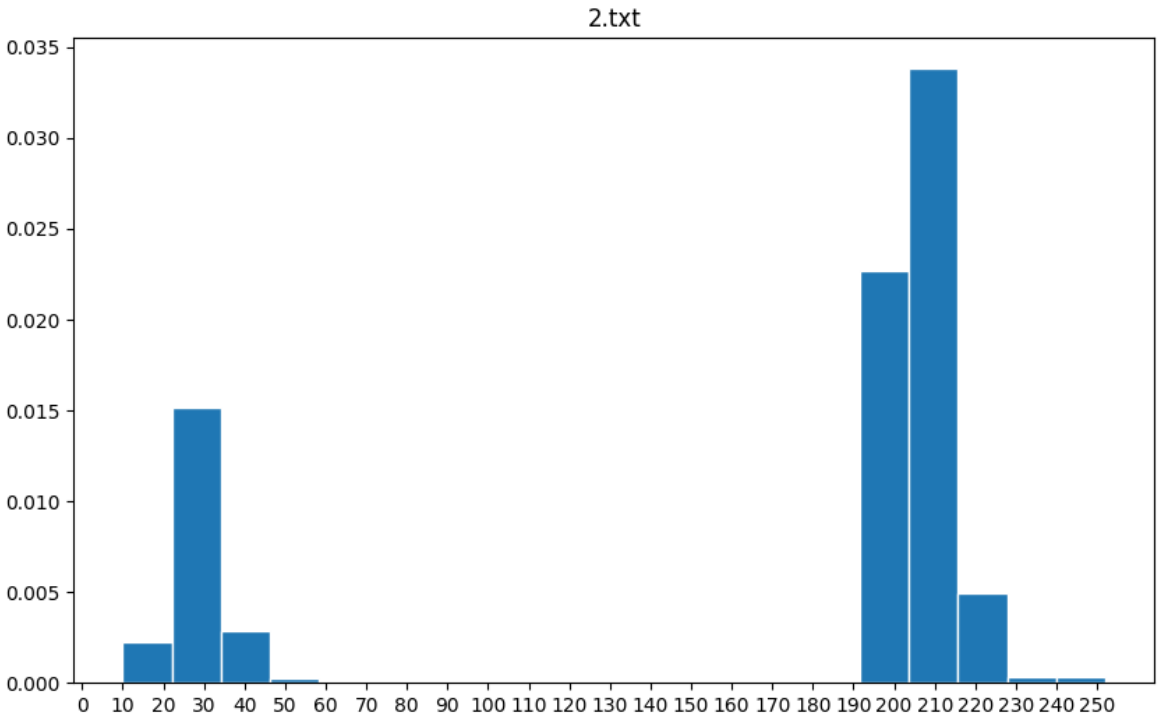
По самым распространенным байтам и их частотам можем понять, что текст нерусский, но кодировка однобайтовая.

Текст действительно написан не на русском. Например, с помощью редактора гитлаба можно понять, что текст написан на греческом языке.

Ради интереса взял файл другого варианта и решил на нем протестировать анализ.

In [5]:
```
file = open('2.txt','rb')
text = file.read()
file.close()

freqs = get_byte_freqs(text)
print_most_freq_bytes(freqs, N)
plot_freqs(freqs, 20, "2.txt")
```

```
5 most common bytes:
32 (18.16%) 207 (8.71%) 197 (6.44%) 193 (5.81%) 201 (5.45%)

5 most common bytes of nonprintable ASCII table:
10 (1.38%) 13 (1.38%)
```



Полученные байты крайне похожи на те, что можно встретить в русских текстах закодированных с помощью koi8r. Расшифруем текст:

In [6]:
```
file = io.open('2.txt',mode='r', encoding="koi8-r")
text = file.read()
file.close()

print(text)
```

Иосиф Бродский. Полторы комнаты

В полутора комнатах (если вообще по-английски эта мера пространства имеет смысл), где мы жили втроем, был паркетный пол, и моя мать решительно возражала против того, чтобы члены ее семьи, я в частности, разгуливали в носках. Она требовала от нас, чтобы мы всегда ходили в ботинках или тапочках. Выговаривая мне по этому поводу, вспоминала старое русское суеверие. "Это дурная примета, -- утверждала она, -- к смерти в доме".

Может быть, конечно, она просто считала эту привычку некультурной, обычным неумением себя вести. Мужские ноги пахнут, а эпоха дезодорантов еще не наступила. И все же я думал, что в самом деле можно легко поскользнуться и упасть на до блеска натертом паркете, особенно если ты в шерстяных носках. И что если ты хрупок и стар, последствия могут быть ужасны. Связь паркета с деревом, землей и т. д. распространялась в моем представлении на всякую поверхность под ногами близких и дальних родственников, живших с нами в одном городе. На любом расстоянии поверхность была все той же. Даже жизнь на другом берегу реки, где впоследствии я снимал квартиру или комнату, не составляла исключения, в том городе слишком много рек и каналов. И хотя некоторые из них достаточно глубоки для морских судов, смерти, я думал, они покажутся мелкими, либо в своей подземной стихии она может проползти под их руслами.

Теперь ни матери, ни отца нет в живых. Я стою на побережье Атлантики: масса воды отделяет меня от двух оставшихся теток и двоюродных братьев -- настоящая пропасть, столь великая, что ей впору смутить саму смерть. Теперь я могу расхаживать в носках сколько душе угодно, так как у меня нет родственников на этом континенте. Единственная смерть в доме, которую я теперь могу навлечь, это, по-видимому, моя собственная, что, однако, означало бы смешение приемного и передаточного устройств. Вероятность такой путаницы мала, и в этом отличие электроники от суеверия. Если я все-таки не расхаживаю в носках по широким, канадского клена половицам, то не потому, что такая возможность тем не менее существует и не из инстинкта самосохранения, но потому, что моя мать этого не одобрила бы. Вероятно, мне

хочется хранить привычки нашей семьи теперь, когда я -- это все, что от нее осталось.

Нас было трое в этих наших полутора комнатах: отец, мать и я. Семья, обычная советская семья того времени. Время было послевоенное, и очень немногие могли позволить себе иметь больше чем одного ребенка. У некоторых не было возможности даже иметь отца -- невредимого и присутствующего: большой террор и война поработали повсеместно, в моем городе -- особенно. Поэтому следовало полагать, что нам повезло, если учесть к тому же, что мы -- евреи. Втроем мы пережили войну (говорю "втроем", так как и я тоже родился до нее, в 1940 году); однако родители уцелели еще и в тридцатые.

Думаю, они считали, что им повезло, хотя никогда ничего такого не говорилось. Вообще они не слишком прислушивались к себе, только когда состарились и болезни начали осаждать их. Но и тогда они не говорили о себе и о смерти в той манере, что вселяет ужас в слушателя или побуждает его к состраданию. Они просто ворчали, безадресно жаловались на боли и ли принимались обсуждать то или иное лекарство. Ближе всего мать подходила к этой теме, когда, указывая на очень хрупкий китайский сервиз, говорила: "Он перейдет к тебе, когда ты женишься или..." -- и обрывала фразу. И еще как-то помню ее говорящей по телефону с одной своей неблизкой подругой, которая, как мне было сказано, болела: помню, мать вышла из телефонной будки на улицу, где я поджидал ее, с каким-то непривычным выражением таких знакомых глаз за стеклами очков в черепаховой оправе. Я склонился к ней (уже был значительно выше ростом) и спросил, что же такое сказала та женщина, и мать ответила, рассеянно глядя перед собой: "Она знает, что умирает, и плакала в трубку".

Они все принимали как данность: систему, собственное бессилие, нищету, своего непутевого сына. Просто пытались во всем добиваться лучшего: чтоб всегда на столе была еда -- и чем бы еда эта ни оказывалась, поделить ее на ломтики; свести концы с концами и, невзирая на то, что мы вечно перебивались от получки до получки, отложить рубль-другой на детское кино, походы в музей, книги, лакомства. Те посуда, утварь, одежда, белье, что мы имели, всегда блестели чистотой, были отутюжены, заплатаны, накрахмалены. Скатер

ть
-- всегда безупречна и хрустела, на абажуре над ней -- ни пылинки, парк
ет
был подметен и сиял.
Поразительно, что они никогда не скучали. Уставали -- да, но не
скучали. Большую часть домашнего времени они проводили на ногах: готов
я,
стирая, крутясь по квартире между коммунальной кухней и нашими полуто
ра
комнатами, возясь с какой-нибудь мелочью по хозяйству. Застать сидящими и
х,
конечно, можно было во время еды, но чаще всего я помню мать на стул
е,
склонившуюся над зингеровской швейной машинкой с комбинированным ножн
ым
приводом, штопающую наши тряпки, изнанкой пришивающую обтрепанные воротнич
ки
на рубашках, производящую починку или перелицовку старых пальто. Отец
же
сидел, только когда читал газету или за письменным столом. Иногда по вечер
ам
они смотрели фильм или концерт по нашему телевизору образца 1952 года. Тог
да
они, бывало, тоже сидели. Вот так год назад сосед нашел сидящего на стуле
в
полутора комнатах моего отца мертвым.

Гипотеза подтвердилась