

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ  
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»  
Дисциплина: «Алгоритмы и структуры данных»

Контрольное Домашнее Задание

Исследование алгоритмов архивации Хаффмана, Шеннона-Фано и LZ77

Графики и таблицы

Выполнил: Шакин Кирилл,  
студент БПИ163.

Преподаватель: Мицюк А.А.,  
старший преподаватель департамента  
программной инженерии  
факультета компьютерных наук

## Оглавление

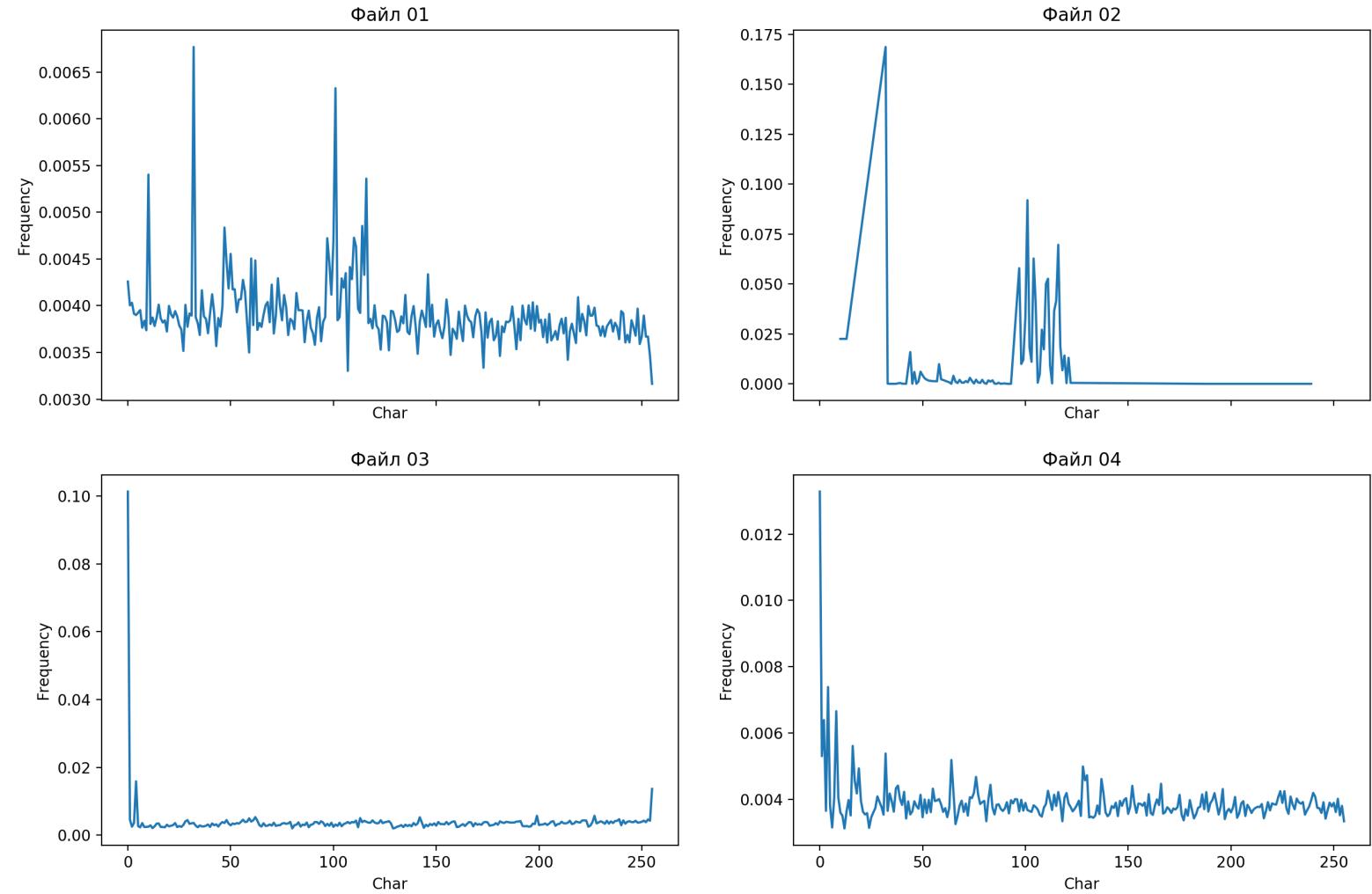
ТАБЛИЦА СРАВНЕНИЯ АЛГОРИТМОВ .....	3
ДИАГРАММЫ РАСПРЕДЕЛЕНИЯ ЧАСТОТ БАЙТОВ В 36 ТЕСТОВЫХ ФАЙЛАХ.....	4
КОЭФФИЦИЕНТЫ СЖАТИЯ КАЖДОГО ФАЙЛА ДЛЯ КАЖДОГО АЛГОРИТМА.....	13
ВРЕМЯ СЖАТИЯ ДЛЯ КАЖДОГО ФАЙЛА ДЛЯ КАЖДОГО АЛГОРИТМА .....	16
ВРЕМЯ РАЗЖАТИЯ ДЛЯ КАЖДОГО ФАЙЛА ДЛЯ КАЖДОГО АЛГОРИТМА.....	21
СРАВНЕНИЕ РАСПРЕДЕЛЕНИЙ ЧАСТОТ ФАЙЛОВ И ЭНТРОПИИ .....	25
ЗАВИСИМОСТИ ЭНТРОПИИ, КОЭФФИЦИЕНТА СЖАТИЯ, ВРЕМЕНИ СЖАТИЯ И РАЗМЕРА ФАЙЛА.....	26
ЗАВИСИМОСТИ КОЭФФИЦИЕНТА СЖАТИЯ ОТ ЭНТРОПИИ ДЛЯ КАЖДОГО ФАЙЛА ДЛЯ КАЖДОГО АЛГОРИТМА .....	27

## Таблица сравнения алгоритмов

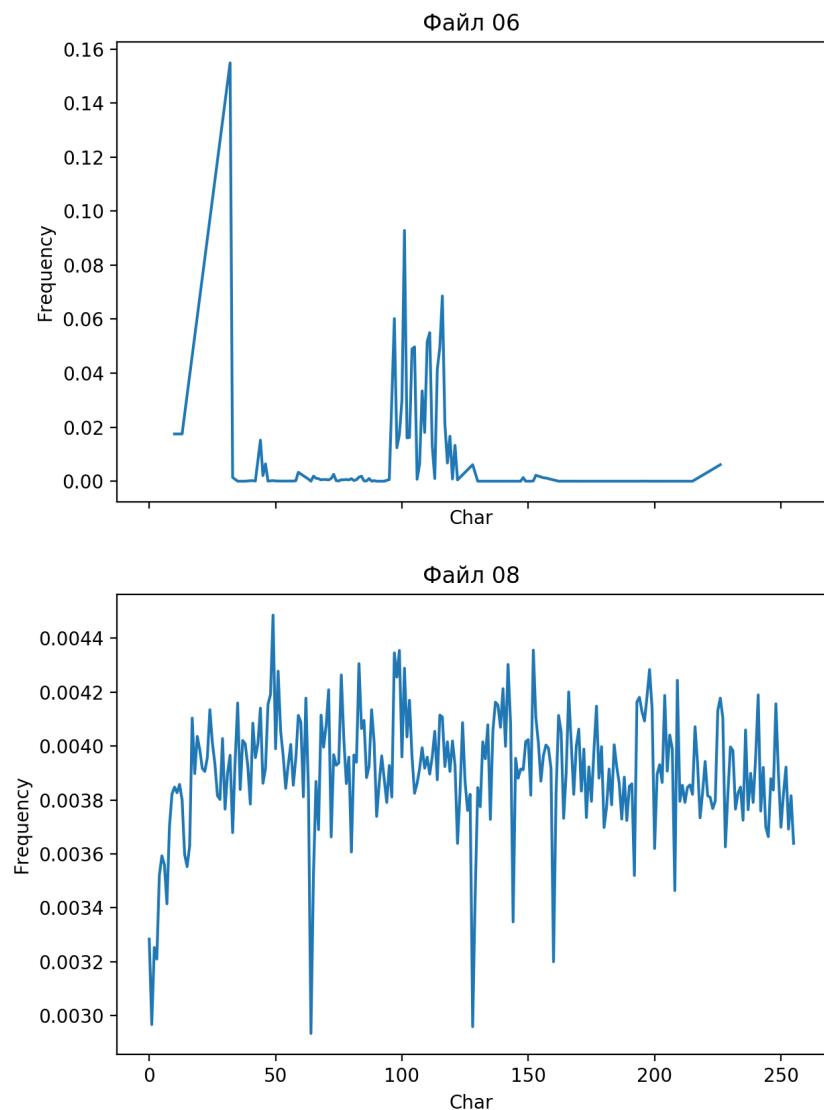
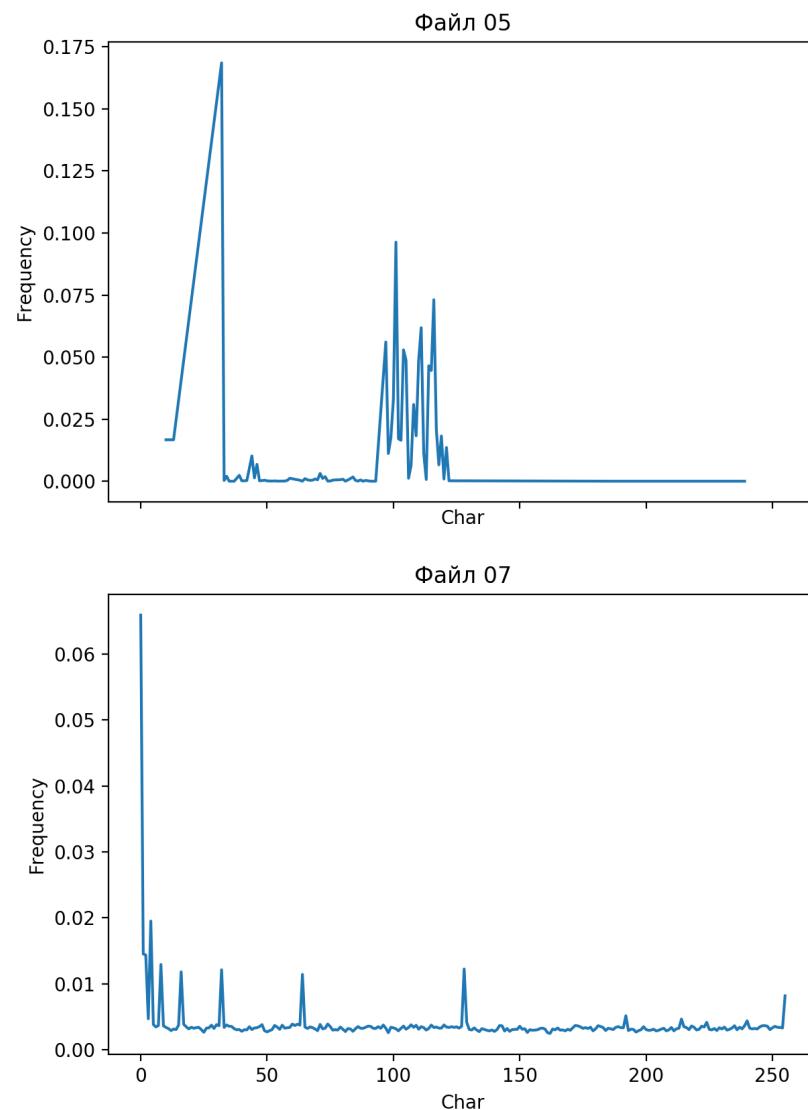
мя	Н	Хаффман			Шеннон-Фано			LZ77 Окно 4кб			LZ77 Окно 8кб			LZ77 Окно 16кб		
		K	tp	tu	K	tp	tu	K	tp	tu	K	tp	tu	K	tp	tu
1	0.999	1.0	4873085698.333	49783613170.333	1.002	911698183.333	19945845395.667	2.369	68047599792.0	536134206.333	2.296	130027180707.667	541035174.0	2.189	242030698116.667	498111684.333
2	0.713	0.579	3090383664.0	6489439118.0	0.579	464896202.333	3001774970.0	0.814	26199001934.0	271714886.333	0.735	46449807304.667	262879615.0	0.668	83587052284.0	254545385.667
3	0.951	0.972	143523332.333	1183407508.0	0.977	26785936.667	505789191.333	1.891	1608560551.333	12541615.0	1.785	2940317026.0	12837548.667	1.699	5325501118.0	12564186.667
4	0.998	1.008	274434280.333	2745887871.667	1.010	54029340.333	1124778770.333	2.350	3892673923.667	30681936.0	2.276	7274674791.333	30111902.667	2.168	13646754995.0	26393641.333
5	0.694	0.569	99230471.0	196343572.0	0.57	15345954.0	91579525.333	0.933	898626184.667	8701010.0	0.846	1613250742.667	8451577.0	0.774	2796946421.667	16383259.0
6	0.676	0.58	887264461.333	1817129315.0	0.582	140528359.0	882162638.0	1.023	9172423981.333	80525194.0	0.933	16405293759.0	78041392.0	0.856	29462879310.0	75673083.333
7	0.964	0.971	776278473.0	7062880589.667	0.974	138913989.667	2858115536.0	1.936	9312255283.0	78719091.667	1.866	17499500271.333	74209153.667	1.779	32529333202.667	69347104.667
8	1.0	1.0	5676597674.333	58241957604.0	1.002	1073351830.0	23190072125.667	2.377	79772509014.667	597271853.333	2.298	152260347520.0	594979181.333	2.182	282180905820.0	570722700.667
9	0.999	1.0	46280801496.333	462085679153.667	1.003	8744282165.333	187649438778.333	2.381	639801137914.0	4905229017.333	2.304	1235327933769.0	5056985868.667	2.184	2281209573644.0	4873027055.0
10	0.884	0.897	245174641.0	1645978979.667	0.9	45082460.0	658255679.0	1.425	2139675105.333	19642096.667	1.361	4135033639.0	19433587.333	1.300	7566312960.667	18613811.333
11	0.955	0.961	1332506193.0	10466997226.0	0.964	230582441.0	4419267934.333	1.956	15169786552.667	130848098.333	1.862	29222249411.667	111526219.333	1.757	53615058417.333	110833017.333
12	0.597	0.619	122439385.333	446135073.667	0.644	18261724.333	191858523.0	0.46	660284831.0	8503446.667	0.438	1160310708.0	7669608.0	0.422	1922138102.333	7689539.667
13	0.579	0.523	2244426985.667	4095498106.667	0.524	293930492.0	1894198529.0	0.667	18300670152.667	160944741.333	0.597	33209608105.333	164294141.667	0.531	58003619219.333	157532677.667
14	0.574	0.52	2484334097.0	4511847201.667	0.52	315927088.667	2046052530.667	0.673	20536484813.0	175981627.333	0.602	37153275540.0	179128790.333	0.541	65371029214.0	167299507.0
15	0.986	1.003	169867254.333	1573052120.0	1.007	33135644.0	652275904.0	2.312	2216950402.0	16193491.0	2.239	4458252741.667	16965518.333	2.131	7930016096.333	16038550.667
16	0.795	0.799	2245415612.333	12872623280.0	0.8	324577566.667	5140879161.0	0.912	14628293507.667	154853227.0	0.853	27252463491.0	157383400.667	0.794	50844719245.333	155631808.0
17	0.941	0.947	1873580331.0	13674344745.667	0.949	264872219.667	5604837845.333	0.846	10782166559.667	132585393.333	0.778	19523326718.667	126608067.667	0.701	35803512104.333	127885381.667
18	0.794	0.798	2571316758.333	14299308468.333	0.8	370046393.333	5787550787.0	1.206	21648271171.667	200930590.0	1.118	39558885490.0	195970774.333	1.036	73803477189.667	195971269.333
19	0.991	1.001	355787417.0	3130559677.667	1.004	66823622.333	1330789775.0	2.395	4942401880.333	35671384.333	2.317	9272761050.333	34605403.667	2.202	17765397723.0	37773713.333
20	0.998	1.017	141752772.333	1362390308.667	1.021	28970372.0	566111988.333	2.403	1970760832.667	14761121.0	2.333	3697370913.667	14804682.333	2.225	6851760718.333	13238949.333
21	0.998	1.017	144990109.667	1366769313.667	1.005	28628504.667	569192626.0	2.397	1927756568.333	54600578.333	2.327	3675392656.333	14550949.0	2.219	6871618763.333	13362216.333
22	0.999	1.002	1349321082.333	13526950881.0	1.020	250997368.0	5572215597.0	2.399	19125917453.667	144497840.0	2.225	36243230766.0	137201111.667	2.200	69482813619.667	138081839.667
23	0.94	0.945	1940028015.0	13928423781.333	0.948	310803814.333	5723394330.0	1.525	19645913068.667	173092045.667	1.421	35555057133.667	166690409.667	1.298	65730878667.333	166414244.333
24	0.975	0.986	371389663.333	3010539493.333	0.989	67938090.667	1259175345.0	2.244	4812965974.333	36028417.667	2.152	9071720647.0	34917236.667	1.984	16634815535.667	36216213.0
25	0.003	0.125	929561857.333	536678065.0	0.125	88603351.667	233888911.667	0.005	41246279016.667	79044211.0	0.002	53985722896.0	77853903.667	0.001	56886731813.667	82358905.333
26	0.98	2.101	3276704.333	18378159.333	2.105	3046604.0	7843610.333	2.225	9427113.0	328567.333	2.225	9383582.667	388667.333	2.225	9535579.333	339200.0
27	0.082	0.234	2409103.333	1501702.333	0.237	370034.0	738934.333	0.102	48731935.0	292400.0	0.099	35189793.333	455467.333	0.098	8102743.667	386034.0
28	0.562	0.77	2406837.0	3815172.0	0.795	583200.667	1910469.333	0.19	1080501.667	320500.333	0.19	1037868.333	304267.333	0.19	1081068.0	187733.667
29	0.914	0.887	25148555485.333	142309901466.0	0.89	3470090622.0	59557125063.333	0.624	110099357872.333	1778330515.667	0.55	189871981647.0	173029991.0	0.391	274818315952.667	1652025680.0
30	0.943	0.946	25533636621.333	184142731672.0	0.949	3657301848.333	73933038517.333	0.973	166733230671.0	2098645153.0	0.875	295961802912.333	1993871238.0	0.722	490138402104.333	1940194838.0
31	0.815	0.818	36048464380.667	202691416369.333	0.821	5293431916.0	82159773811.0	1.457	349955353995.667	3200222538.0	1.393	677150598112.667	3201076392.333	1.317	1265017294848.333	3195644686.0
32	0.772	0.775	144572764394.0	647338107757.333	0.777	20974663939.0	272139959150.333	1.278	1307366019496.0	13076023043.667	1.226	2481823348270.666	12762788950.333	1.159	4619994311719.0	12551641506.3
33	0.811	0.814	7457493214.0	42603938416.0	0.816	1081756645.667	17067439277.333	1.259	66997280095.667	681924217.0	1.072	111692819018.0	588547945.333	0.995	204570319805.0	606549944.333
34	0.818	0.823	18111272095.667	95345729453.333	0.828	2805228726.333	40044967382.667	1.197	171717186724.0	1597131675.667	1.054	319427645300.333	1589589157.333	1.054	598207098019.333	1545676127.0
35	0.728	0.732	3669196590.333	16404726076.0	0.733	529001883.667	6570774116.0	1.012	46290363995.667	298669246.667	0.947	87578159510.333	273307509.333	0.892	166898041169.667	269619685.333
36	0.742	0.748	1226956882.0	6941347911.0	0.752	172596244.333	2872681437.333	1.031	10692473964.0	92787333.0	0.986	19469453488.667	95647150.333	0.941	34931746207.0	91700323.0

## Диаграммы распределения частот байтов в 36 тестовых файлах

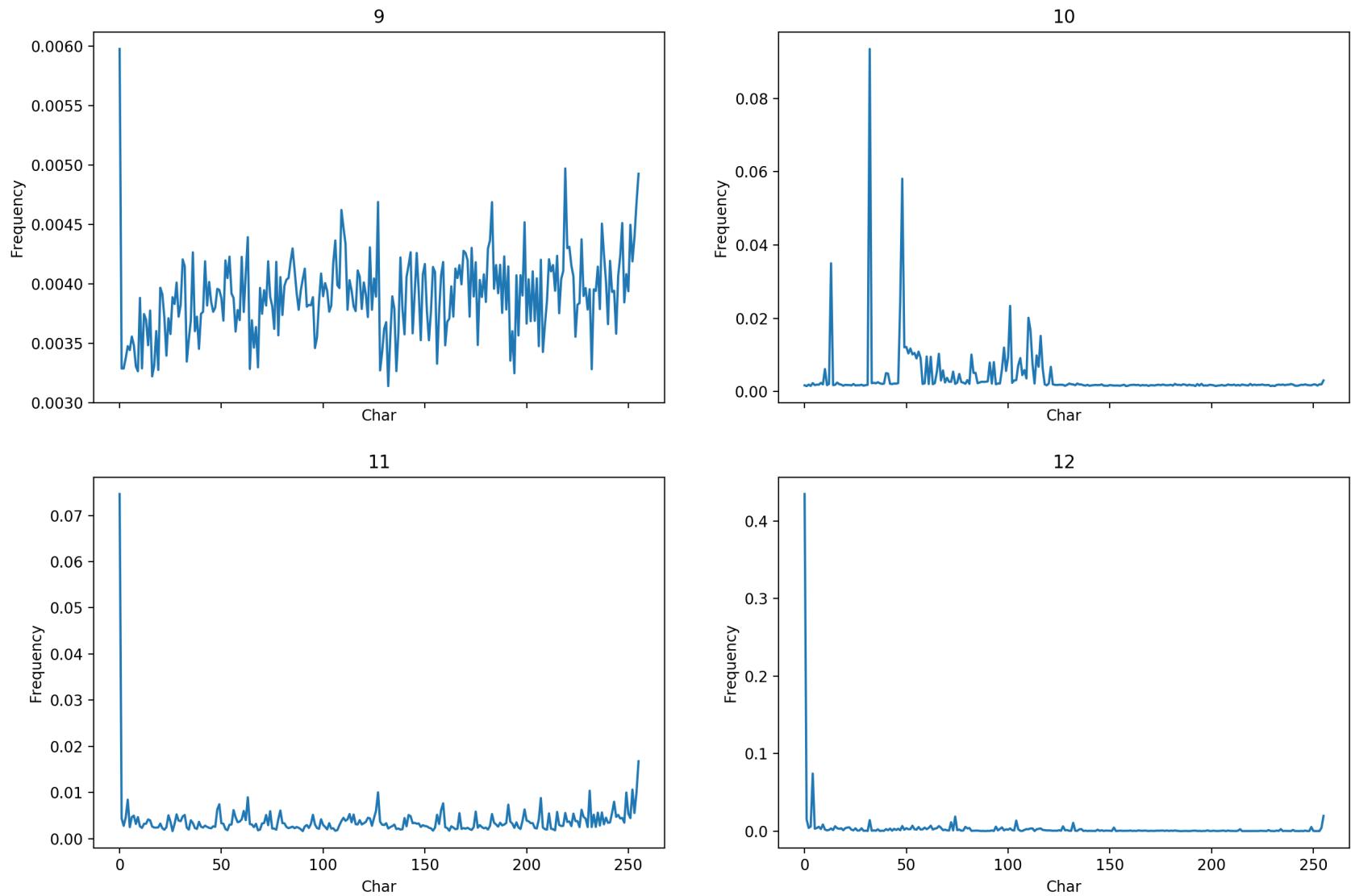
1, 3 и 4й файлы имеют почти равномерное распределение. У первого и третьего хоть график и выглядит достаточно изменяющимся, однако эти изменения только в 3м знаке после запятой — то есть по сути равномерное. В 3м файле очень популярный первые байты, что снижает энтропию на 5 сотых, однако все равно распределение можно считать равномерным. Что не скажешь про 2й файл, где заметны явно выделяющиеся по популярности байты, а энтропия уже примерно 0.7.



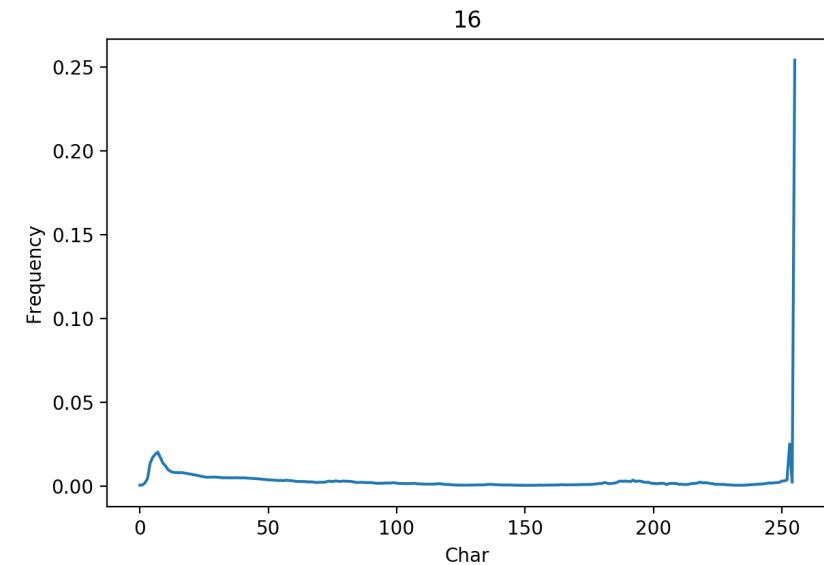
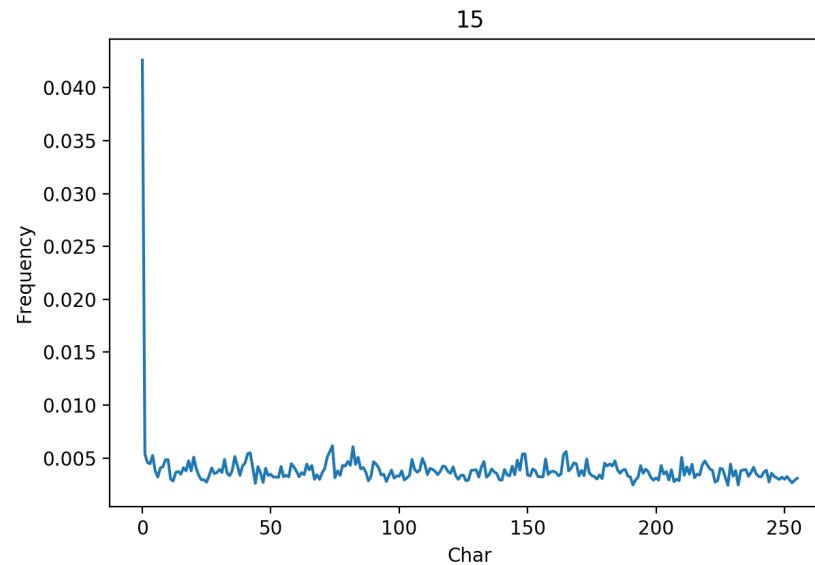
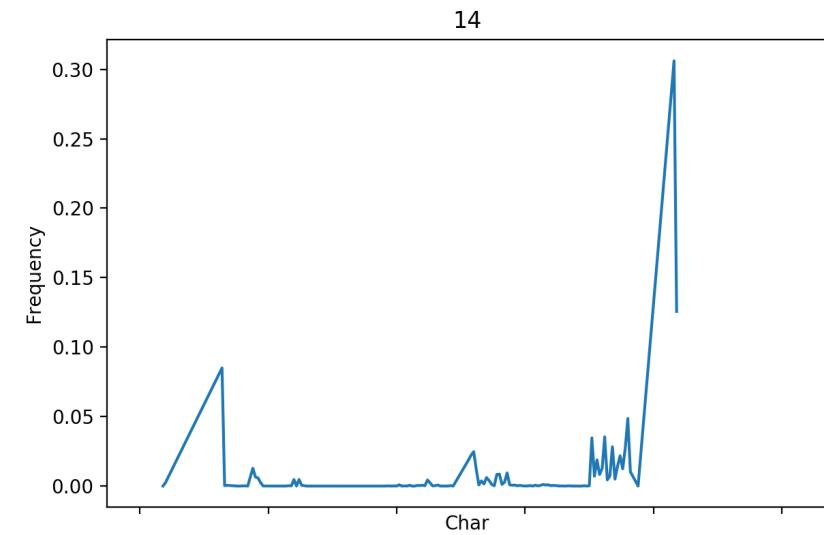
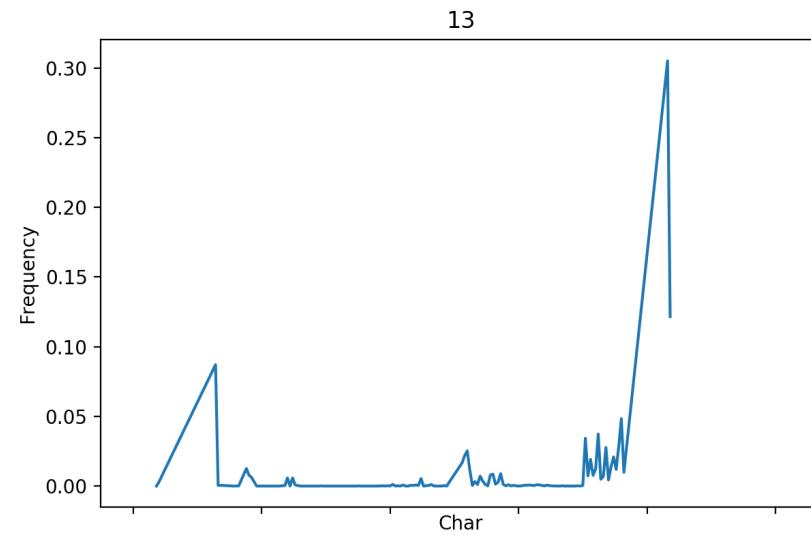
Как можно заметить у файлов 5 и 6 распределение совсем не похоже на равномерное и энтропия у них в районе 0.7. Восьмой файл схож с первым, потому что частоты колеблются в очень маленьких диапазонах. И у 7го файла тоже практически равномерное распределение.



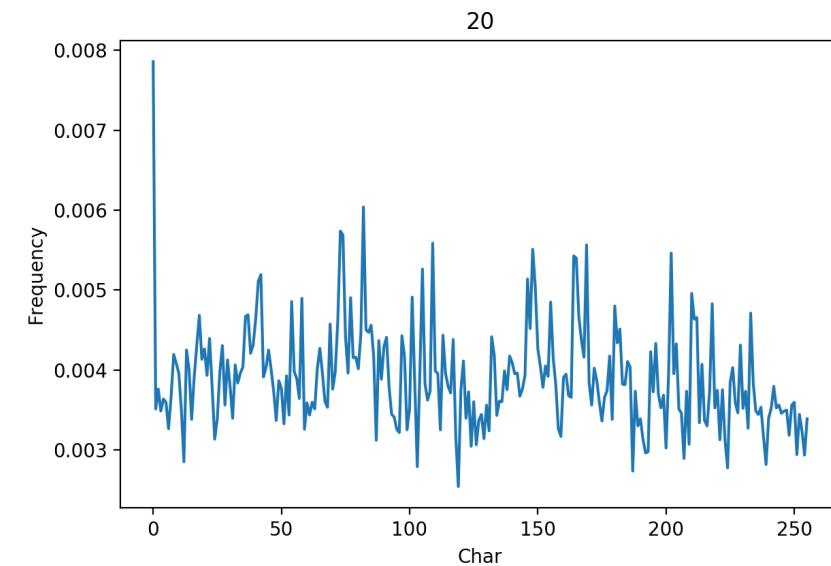
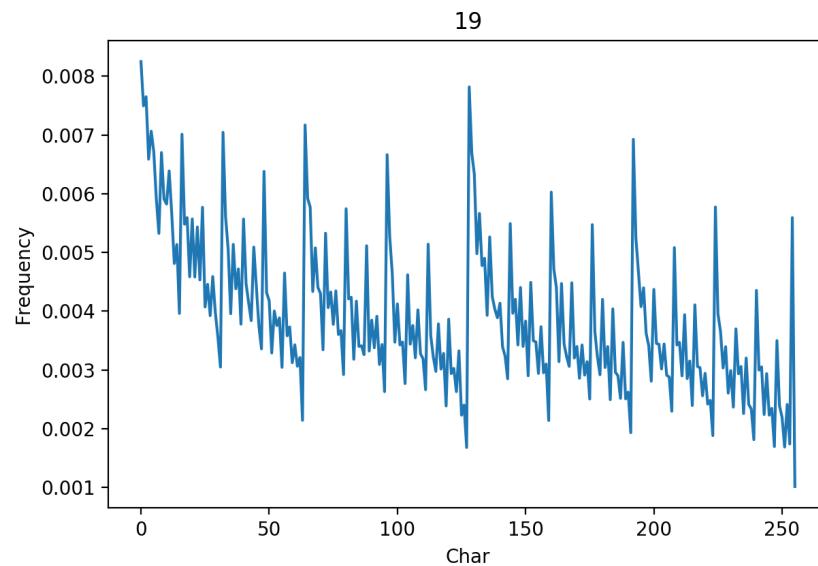
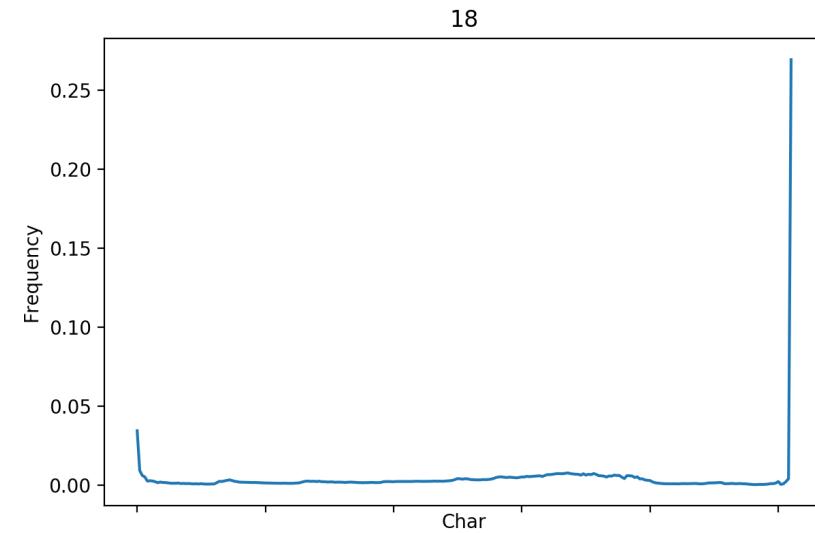
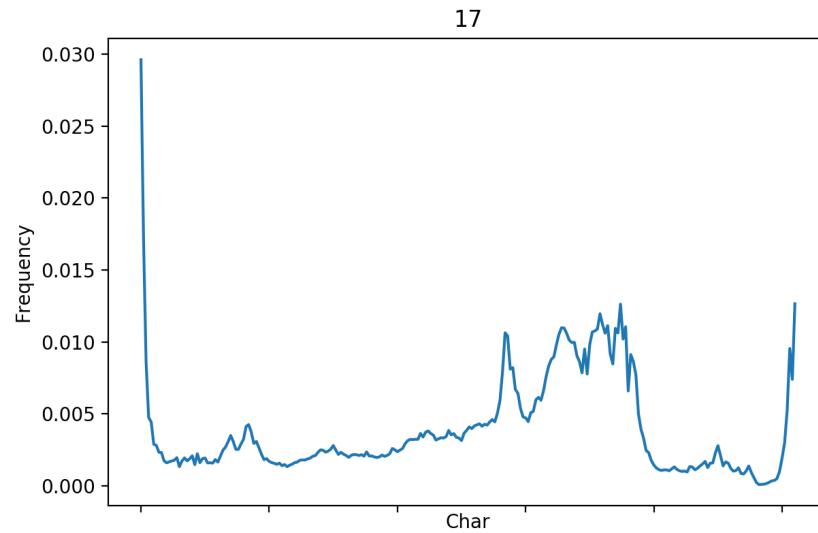
Тут можно выделить 10й и 12й, в 10м преобладают байты < 150, а в 12м сильно выделяются байты < 10.



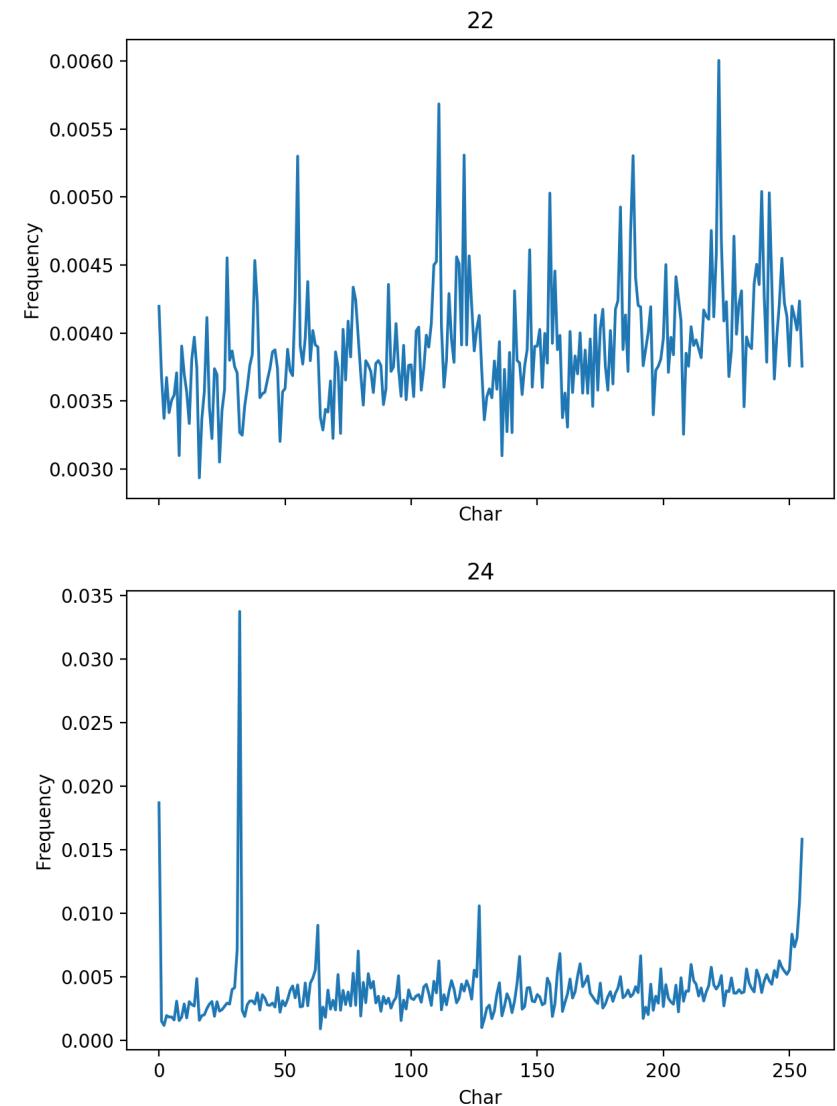
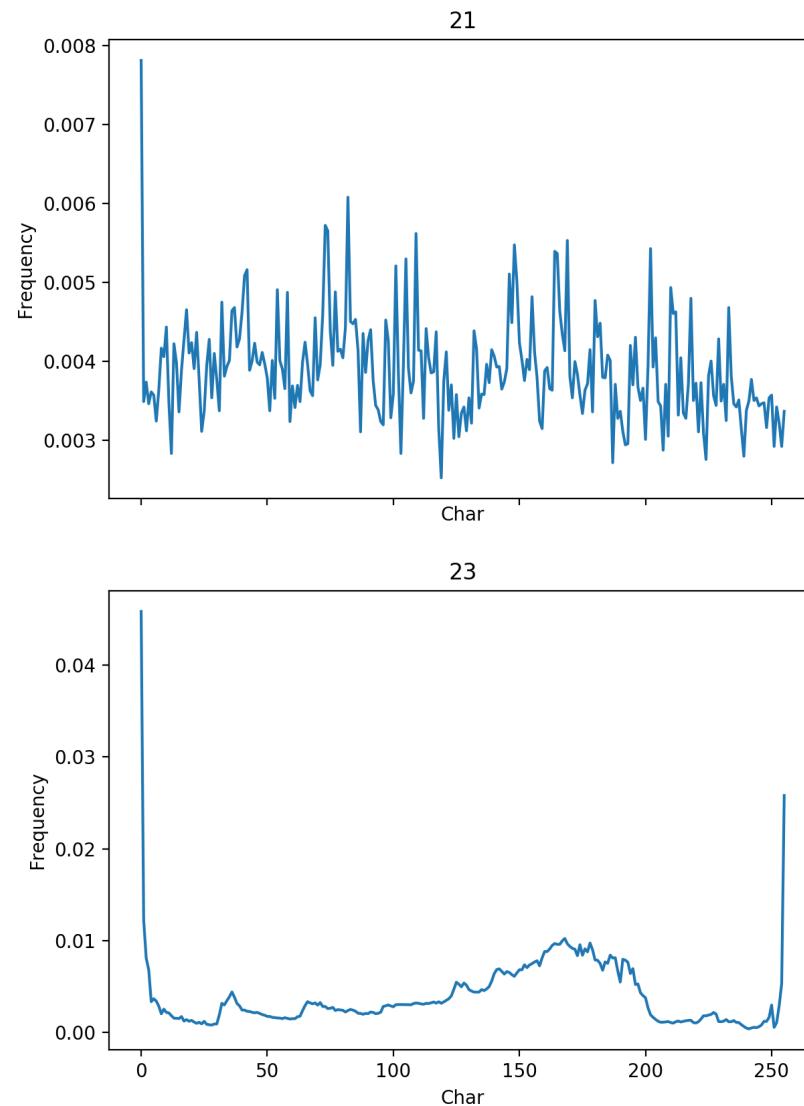
Можно отметить 13й и 14 файлы, у которых очень схожие распределения(но все же отличаются в районе 200го байта), и которые показывают энтропию в районе 0.6.



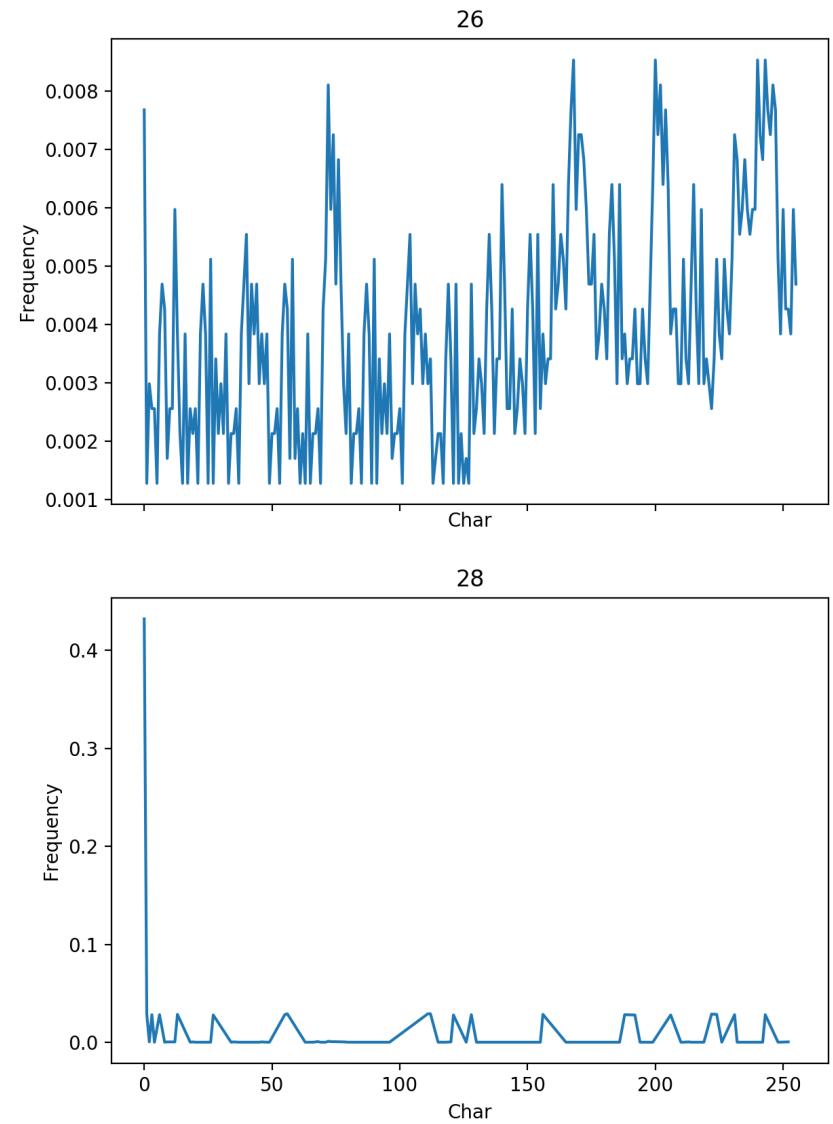
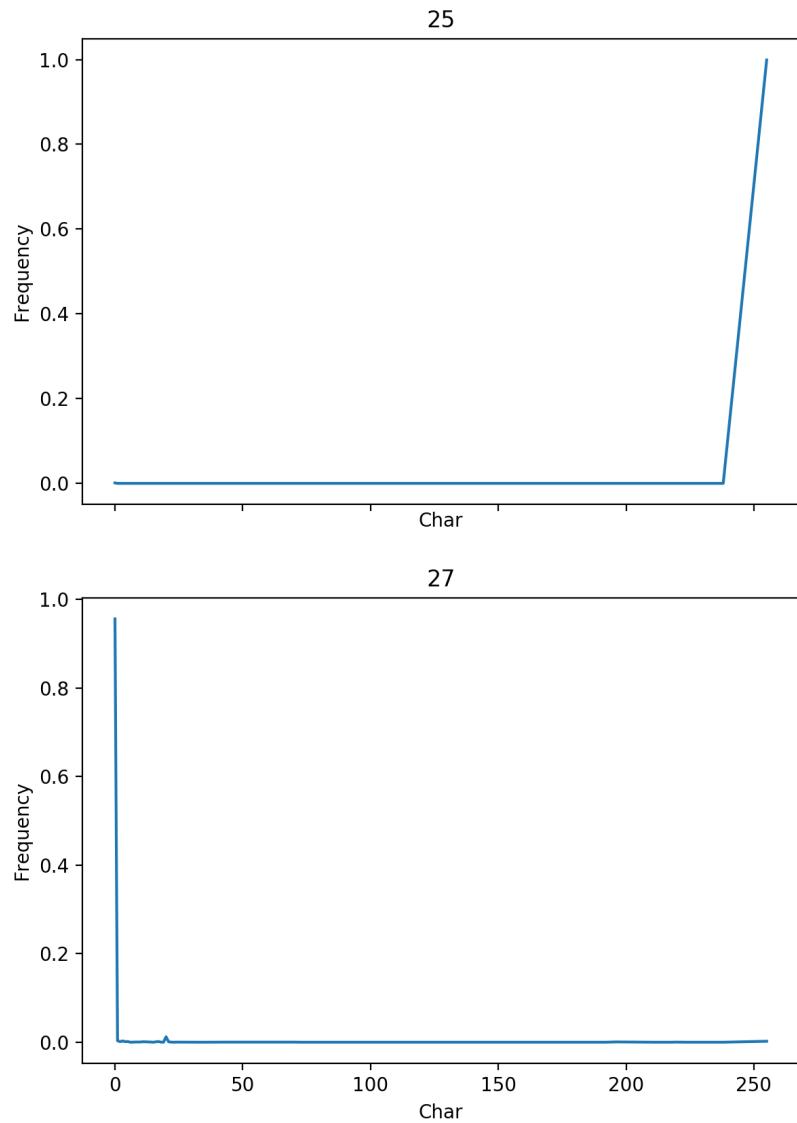
Стоить отметить, что тут на 18м файле, в отличие от ранее появлявшихся графиков, сильно преобладают не начальные байты, а те, что  $> 250$  (энтропия 0.8).



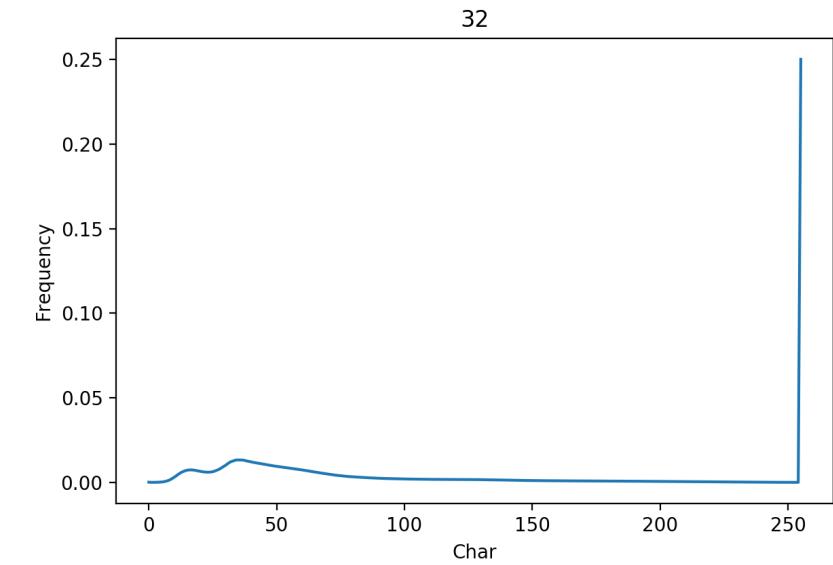
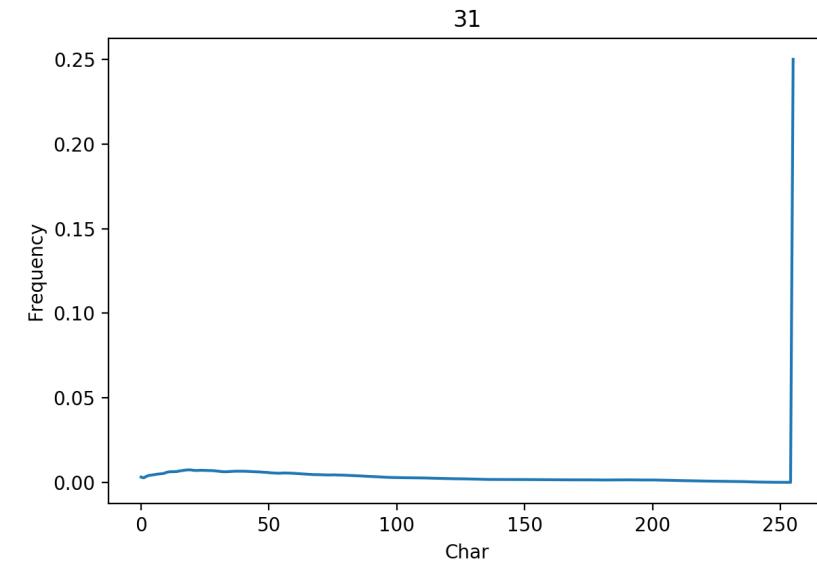
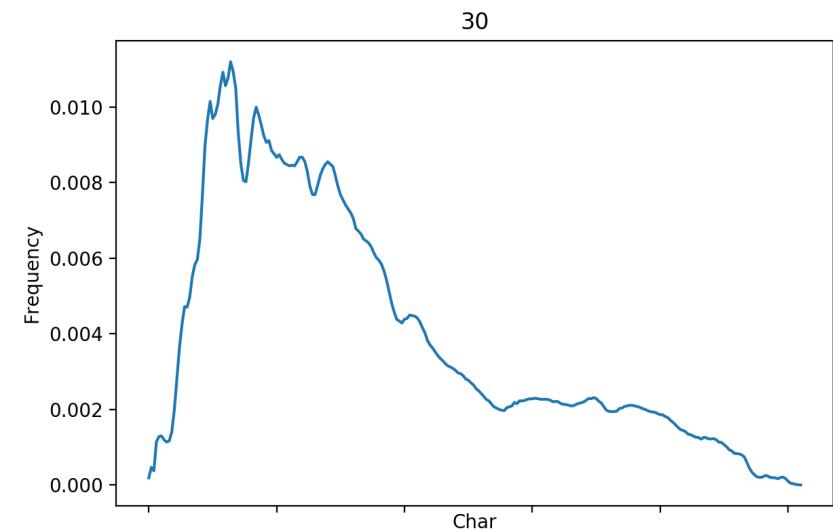
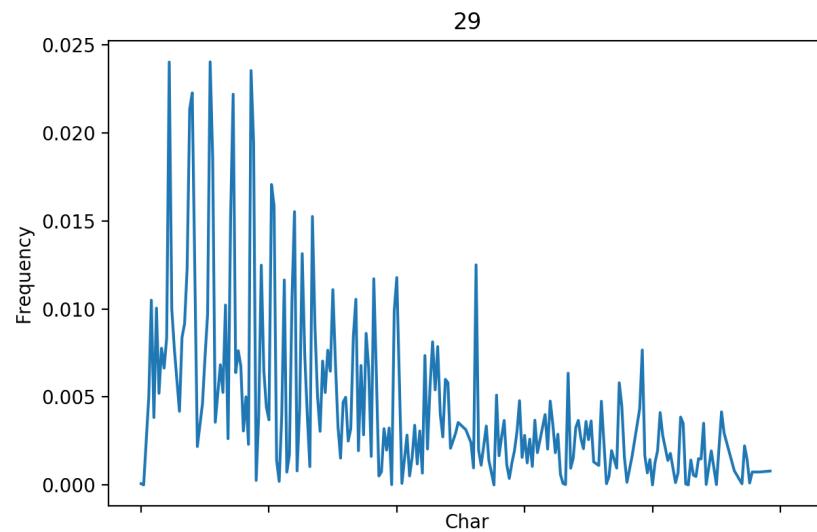
В 23м выделяются начальные(< 10) и конечные (> 250) байты, и в плане энтропии они как бы друг друга компенсируют и получается достаточно высока энтропия в итоге (0.94).



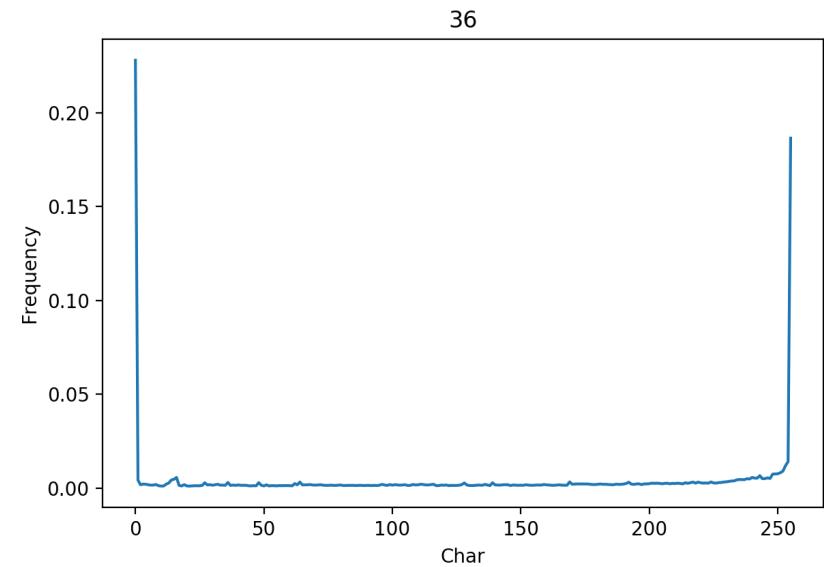
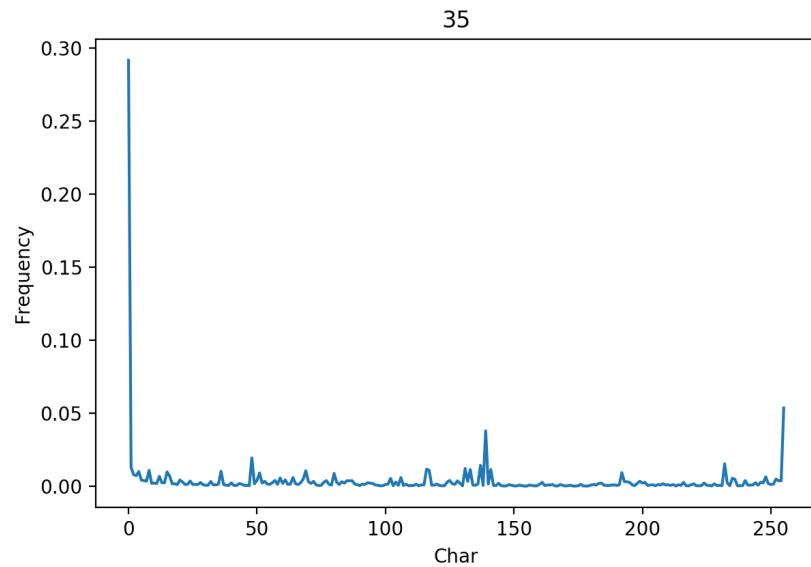
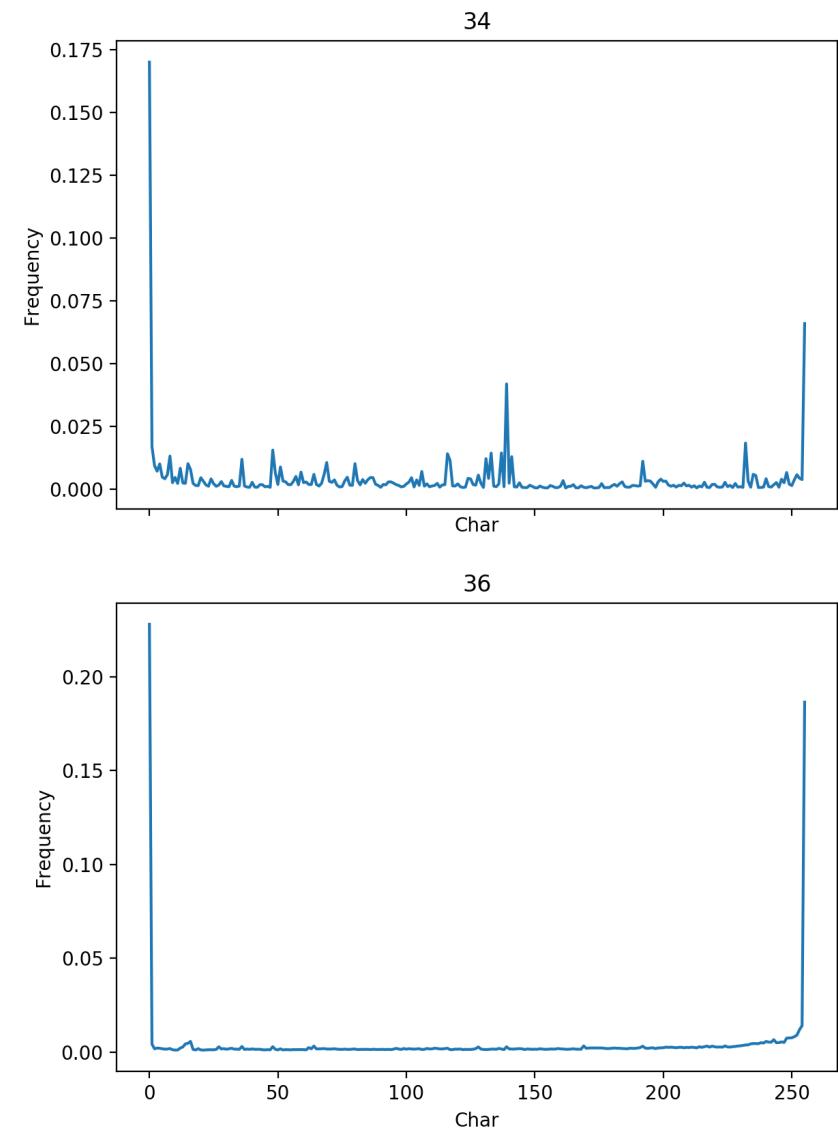
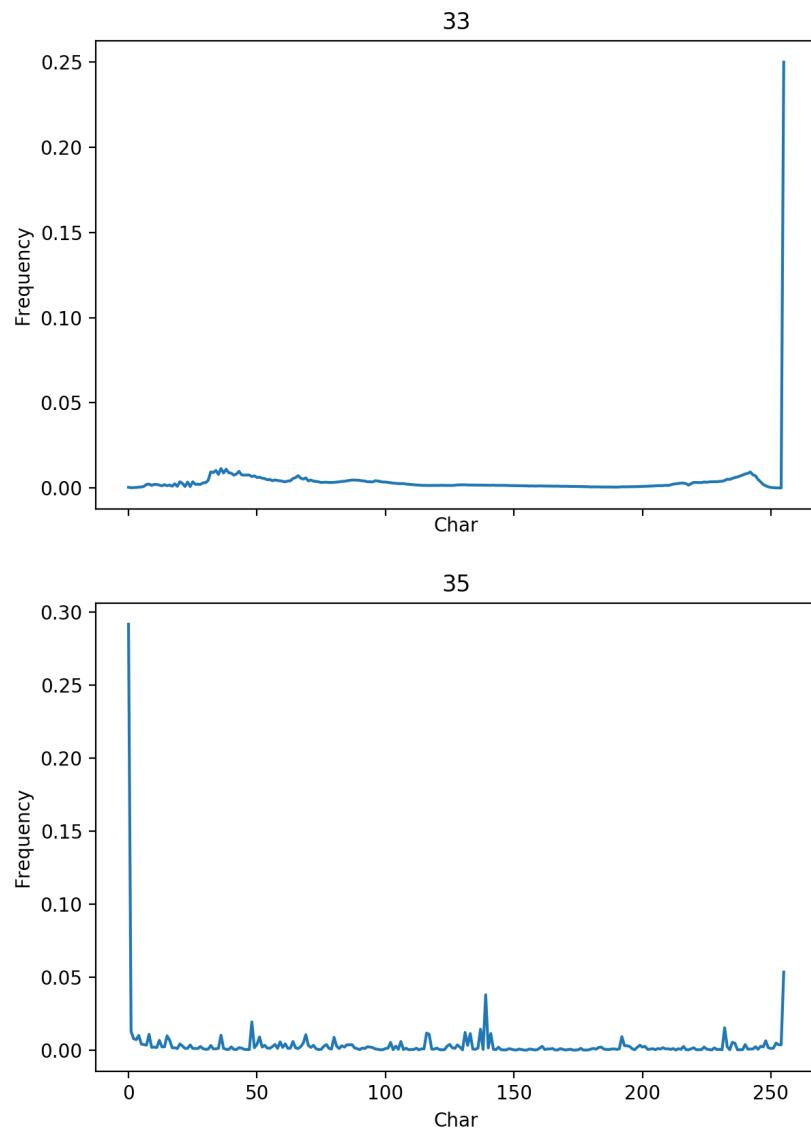
Тут очень интересны файлы 25 и 27, у которых энтропия стремится к 0, значит большая часть файла заполнена очень маленьким количеством различных байт (и скорее всего преобладает какой-то один, как евро в 25м).



Здесь отличие 31го и 32го от 25 и 27 в том, что там частота некоторых байт стремилась к 1, а тут только к 0.25 => и энтропия выше(около 0.8).

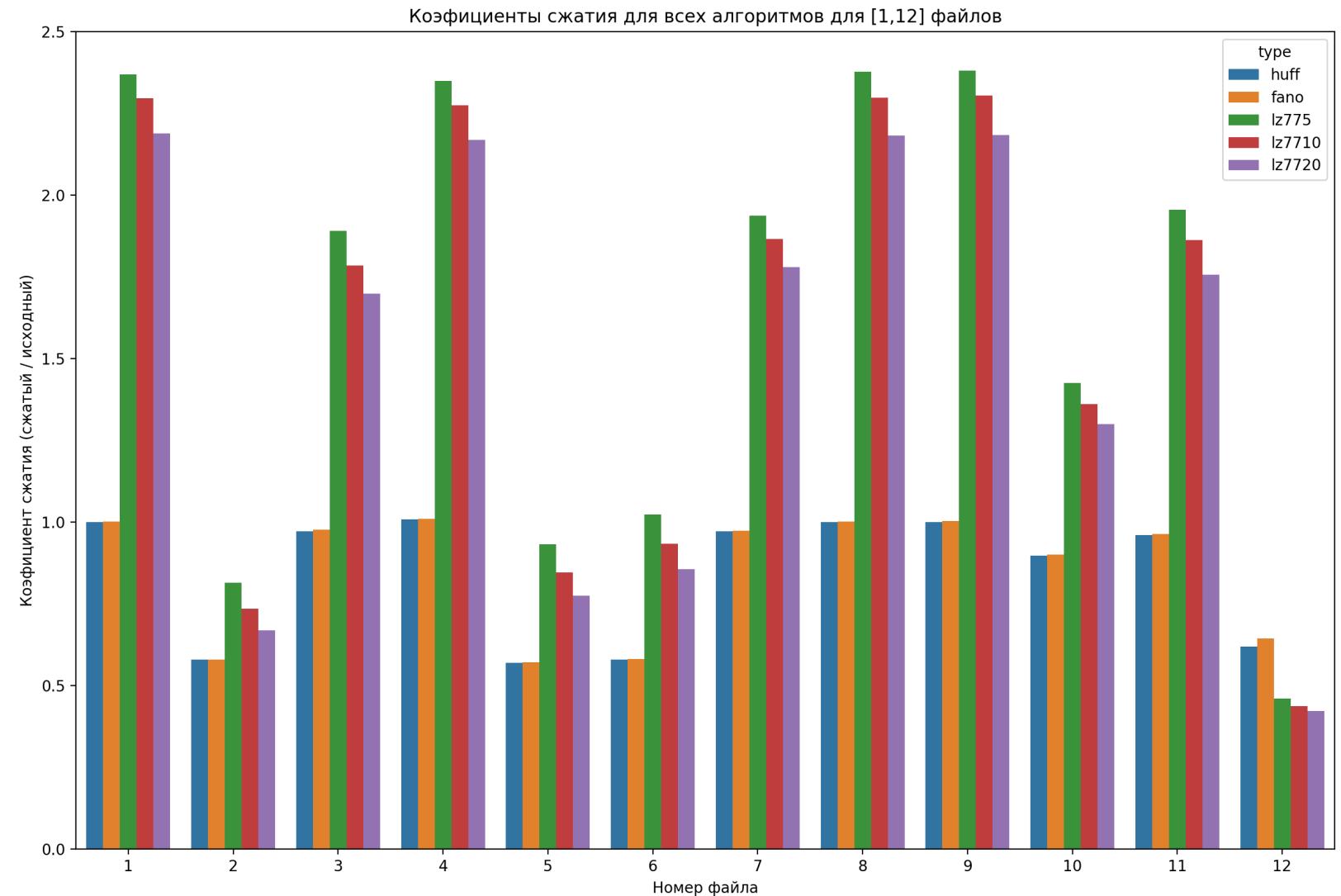


Все примерно схожи  
и энтропия  
колеблется в районе  
0.7 и 0.8.

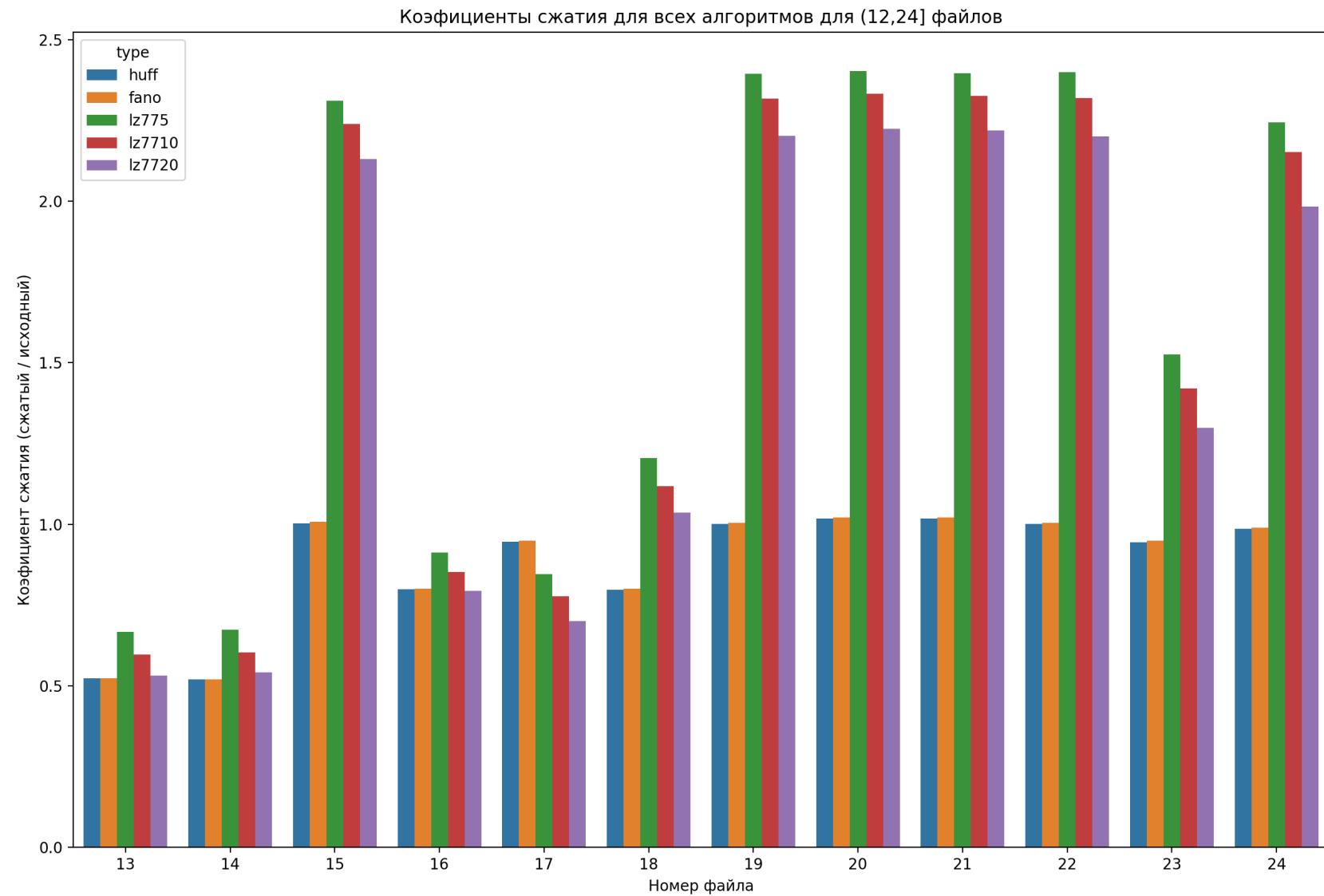


## Коэффициенты сжатия каждого файла для каждого алгоритма

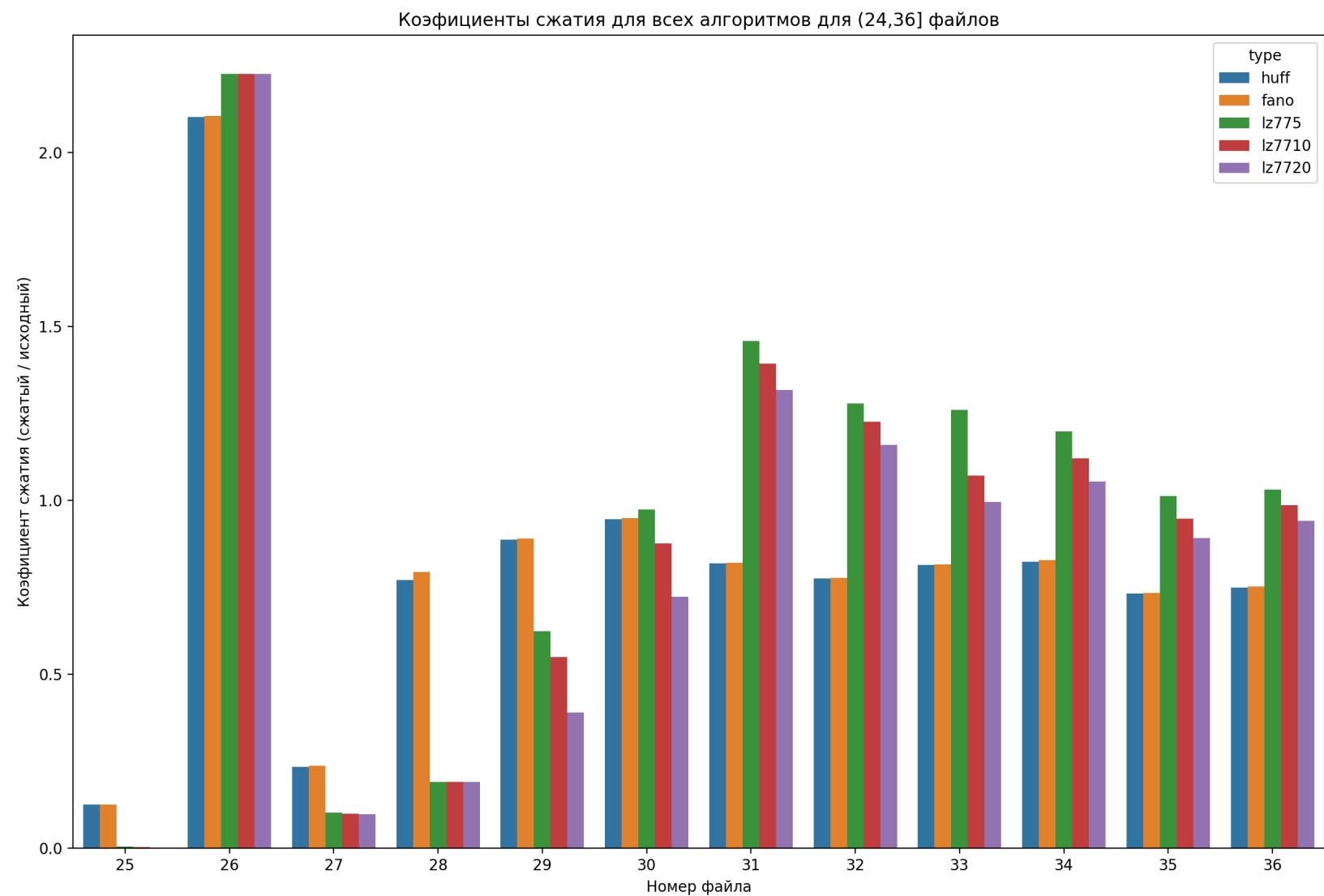
На первых 11ти файлах Хаффман и Фано показывают примерно одинаковые коэффициенты сжатия. Причем они в среднем в 2 раза меньше чем у любого lz77. Однако на 12м файле lz77 сжимает сильнее (ну не удивительно — это же текст про LZ78). И тут же видно четкое отличие разных lz77 — с увеличением окна, улучшается сжатие (можно считать, что зависимость линейная). Причем на всех файлах, кроме, 2,5,6,12, Хаффман и Фано либо не сжимают, либо сжимают чуть-чуть, а LZ77 вообще разжимает в 1.5, 2.5 раза.



Картина в среднем такая же как на первых 12ти.  
А тут LZ77 показал себя лучше только на черно-белой фотографии.

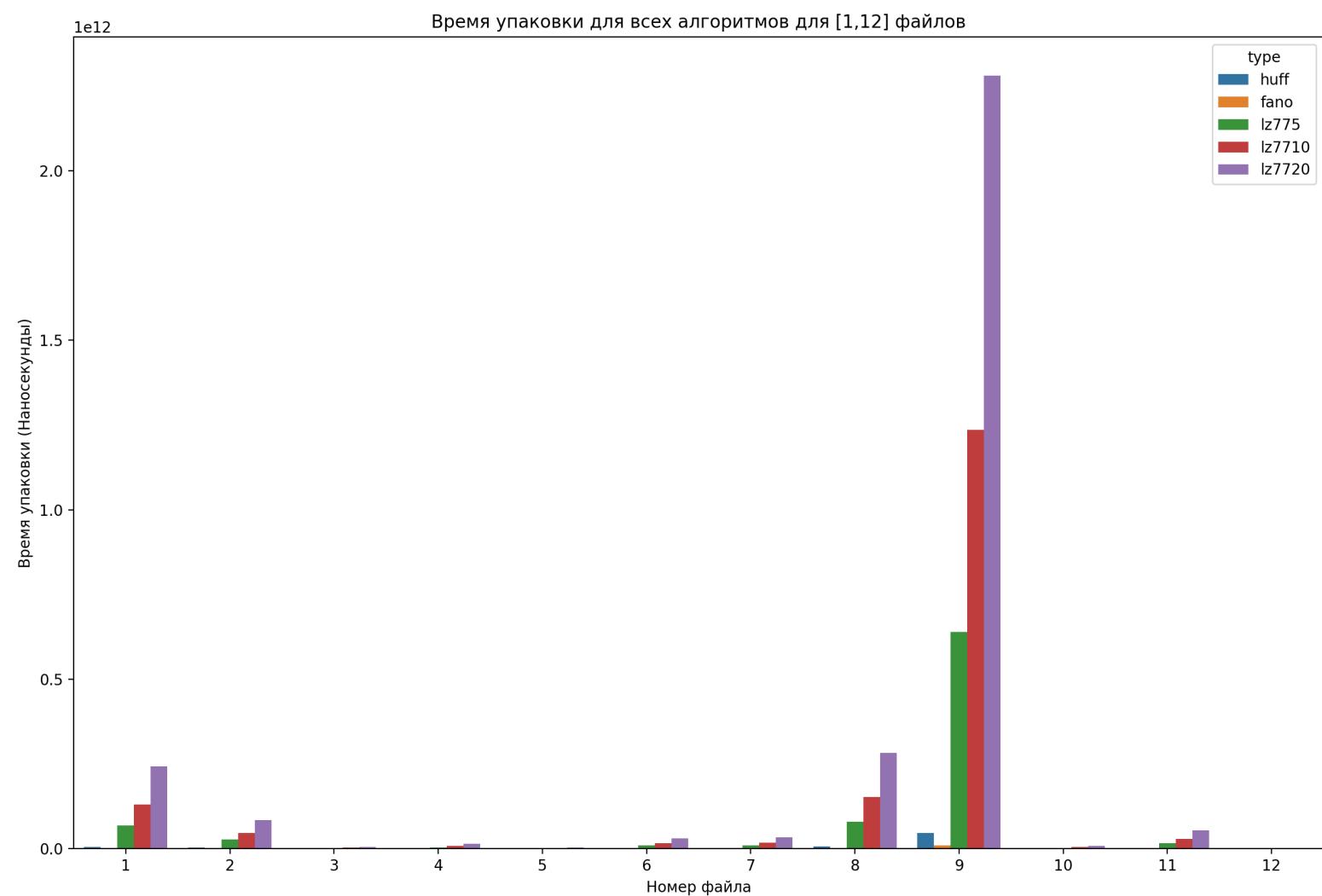


Интересно, что на 25, 27, 28, 29, 30 файлах (картинки в разных форматах), себя лучше показывает LZ77, однако 26й выбивается из общей картины, на нем все алгоритмы показывают себя плохо.

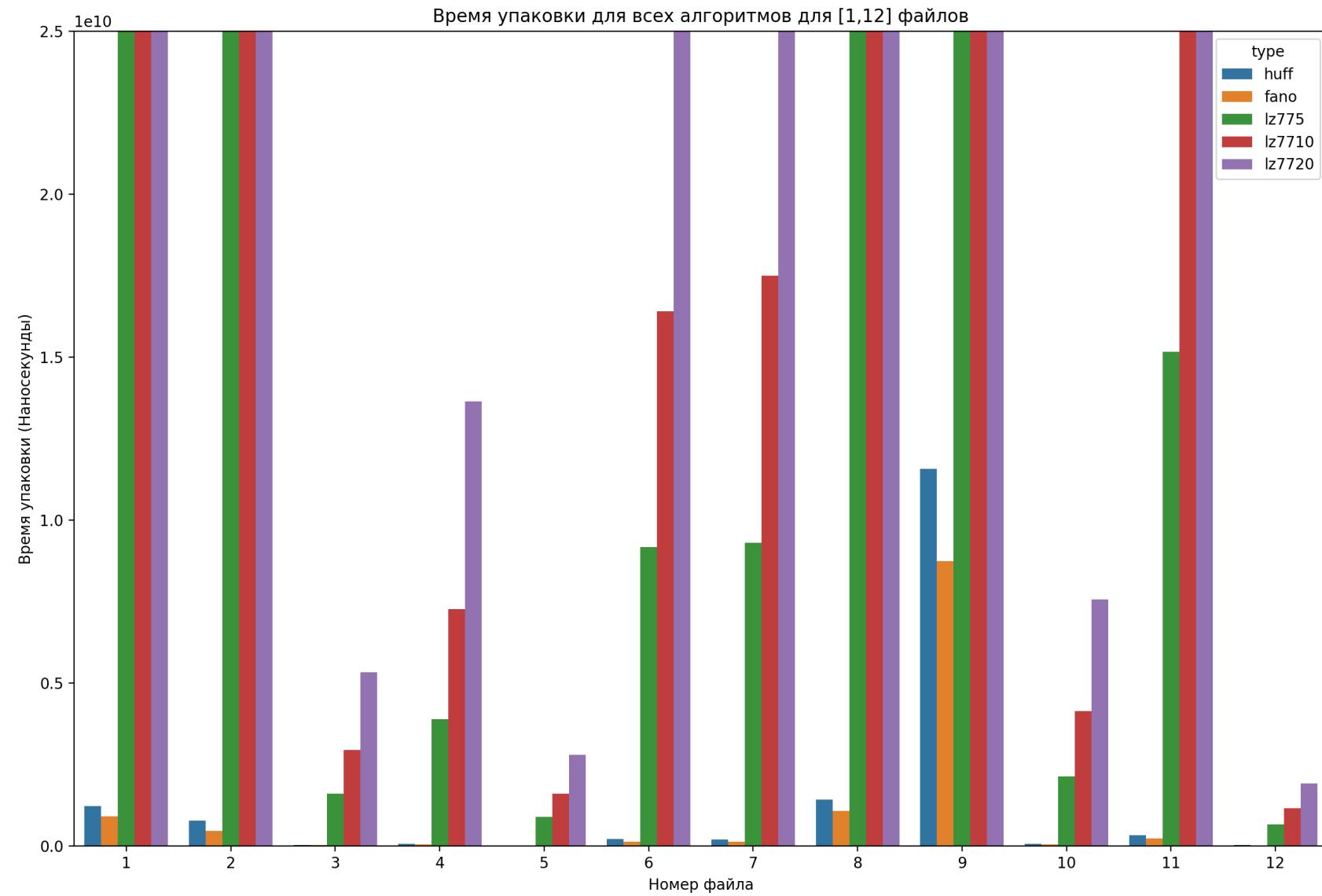


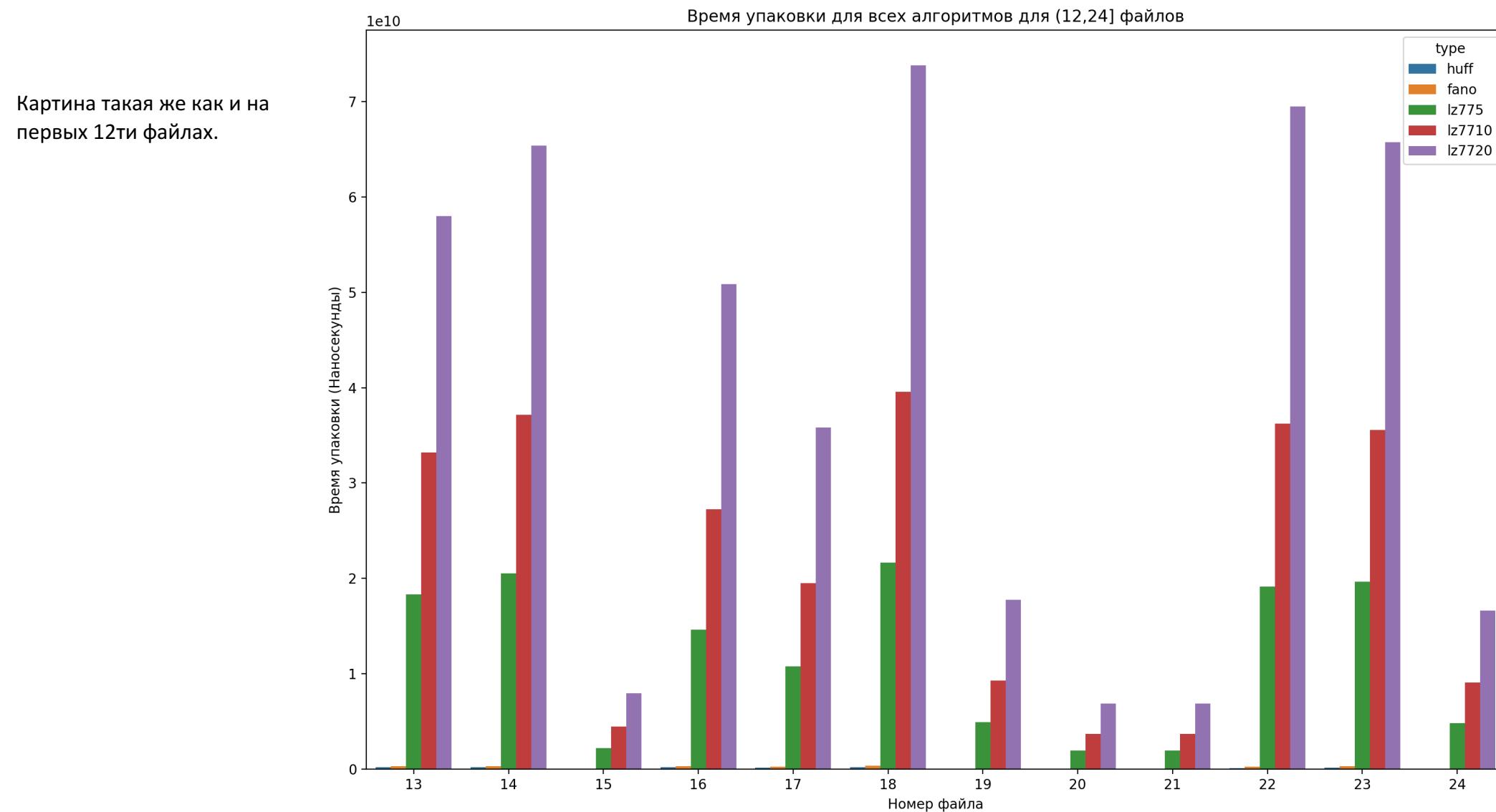
## Время сжатия для каждого файла для каждого алгоритма

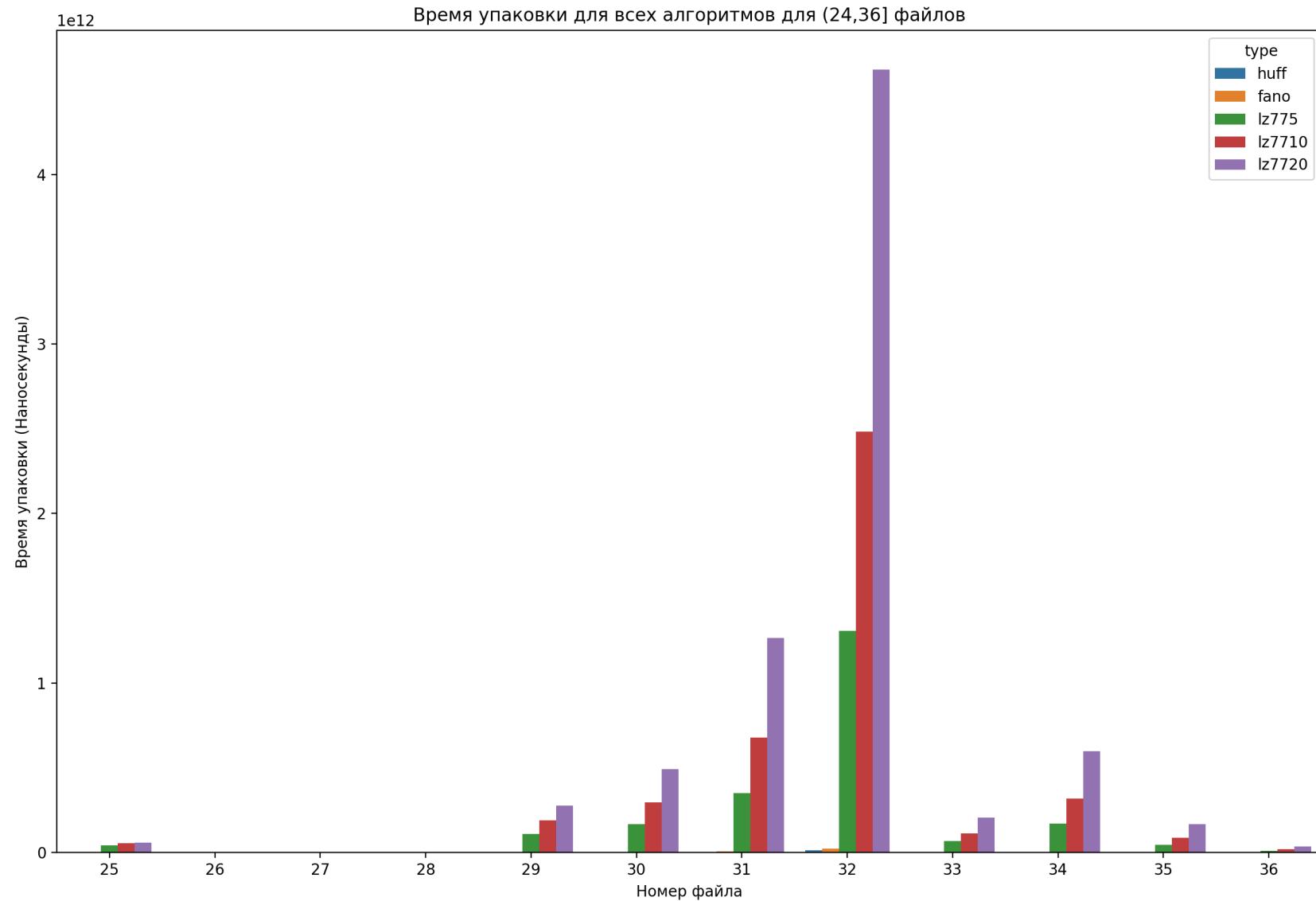
Интересный факт, по такой картине можно судить, только о размере файлов и производительности LZ77, потому что Хаффман и Фано настолько быстрее это делают, что их тут даже не видно (и это подтверждает, асимптотическую сложность алгоритмов —  $n \log n$  у Хаффмана и Фано, против  $n^3$  у LZ). Итак, 9й файл сжимался много больше других, потому что у него максимальный размер на текущем промежутке. В среднем LZ7720 работает в 2 – 2.5 раза медленнее 2 двух других.



Все для тех же файлов, но  
чуть-чуть обрезанный  
график, чтобы увидеть  
Хаффмана и Фано.  
И тут становится заметно,  
что Фано работает  
немного быстрее, чем  
Хаффман. (Хаффману еще  
приходится строить  
дерево)

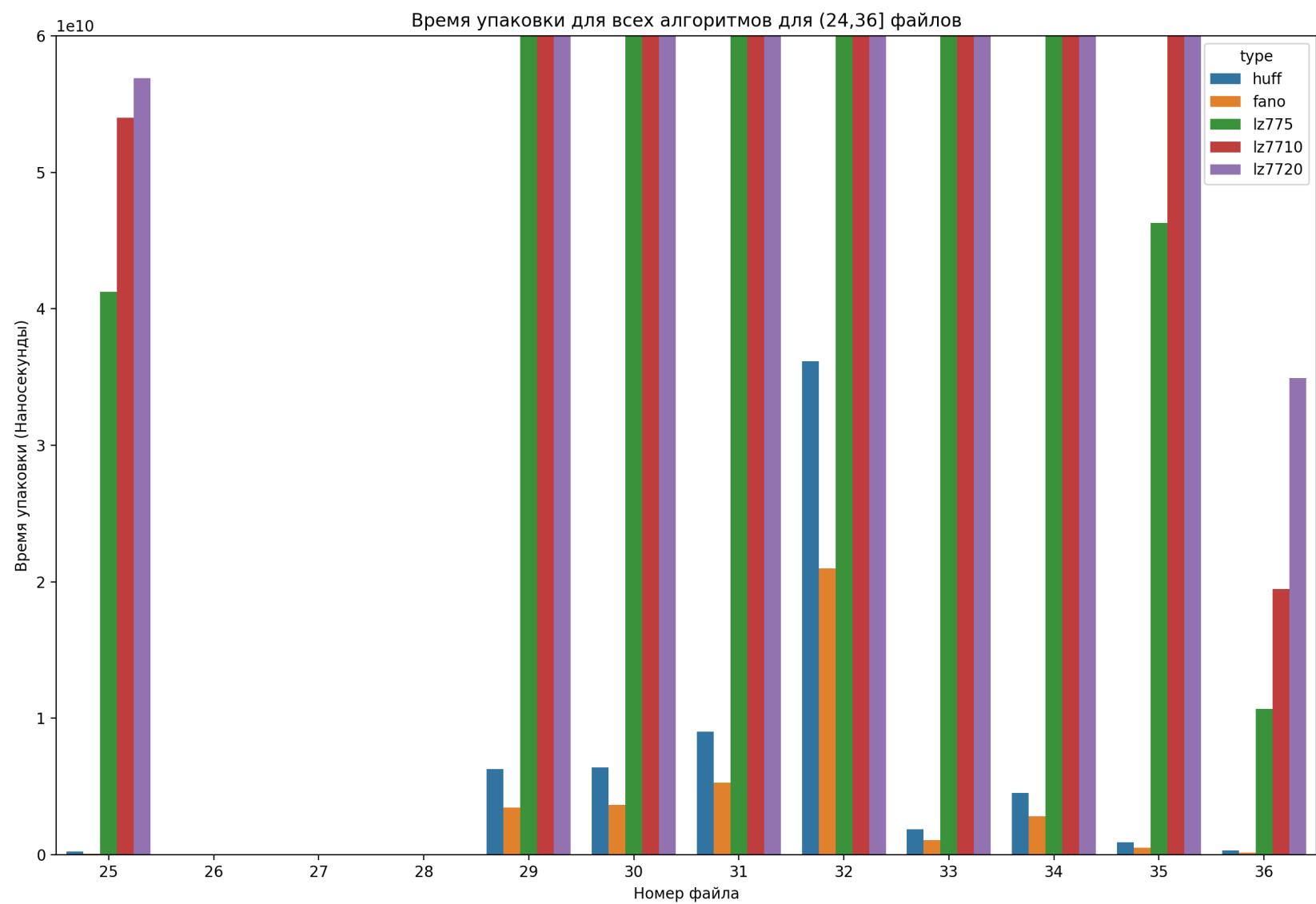




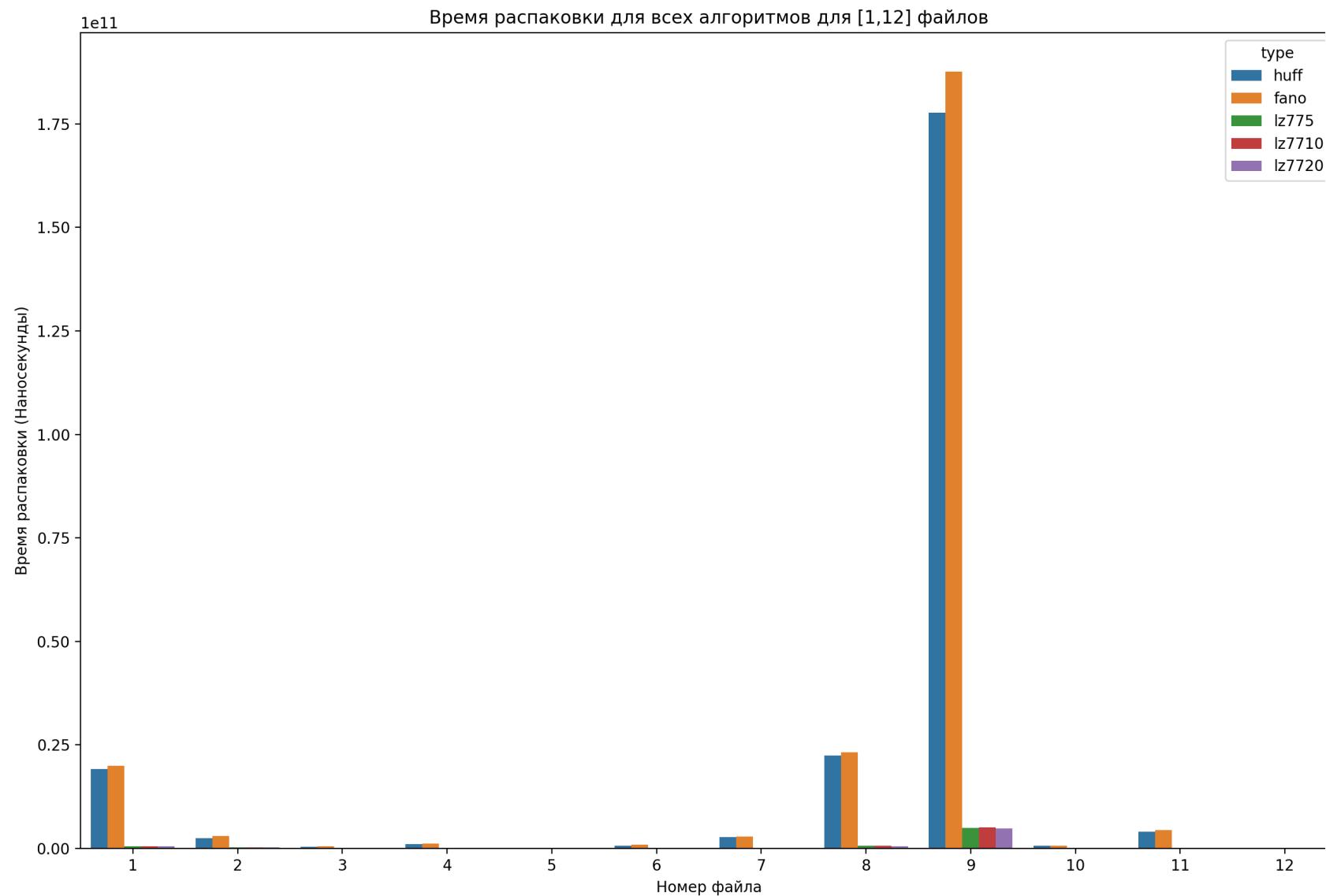


Обрезанный по Y  
график.

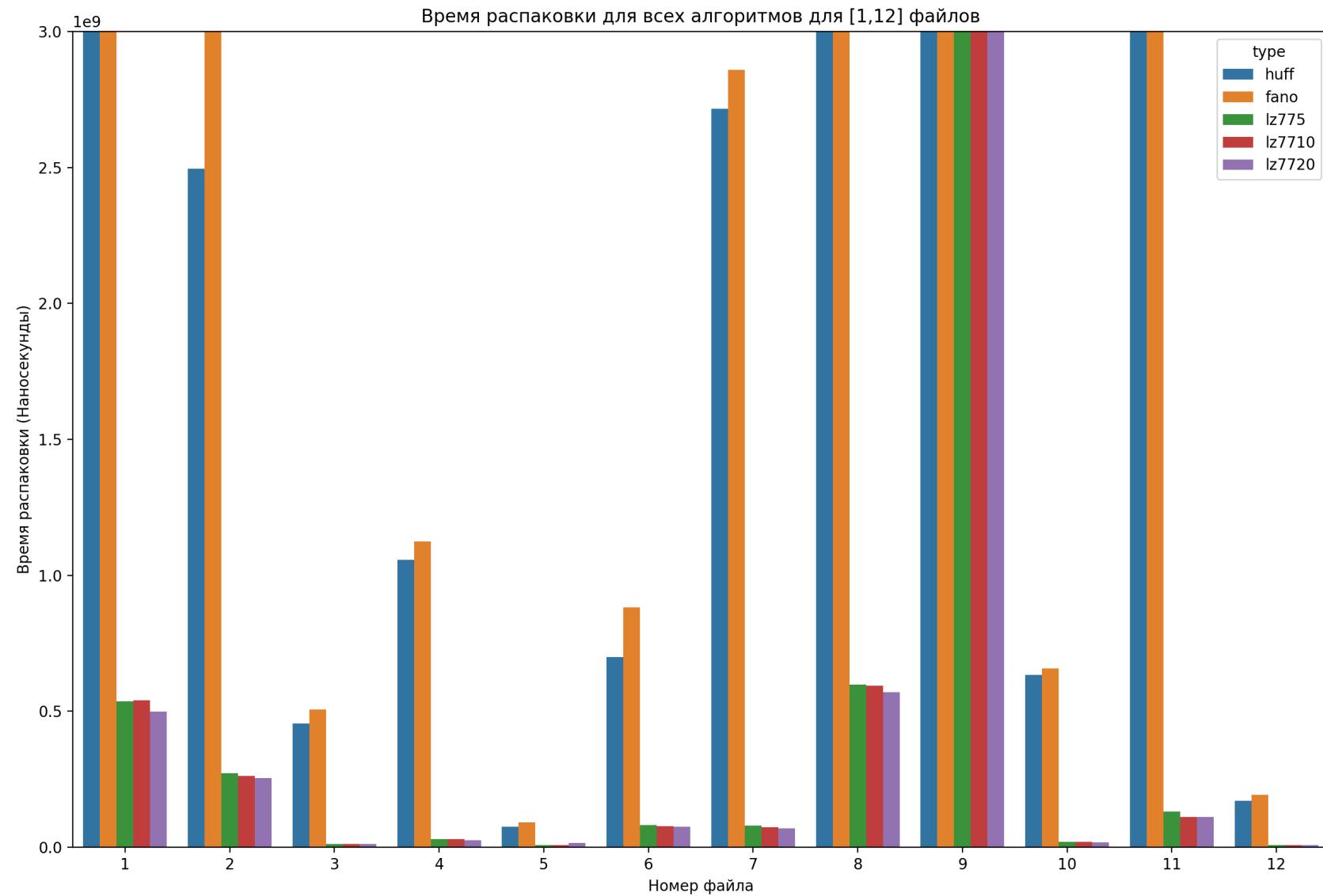
Тут опять же можно  
увидеть, что Хаффман  
сжимает немного  
меньше чем Фано.

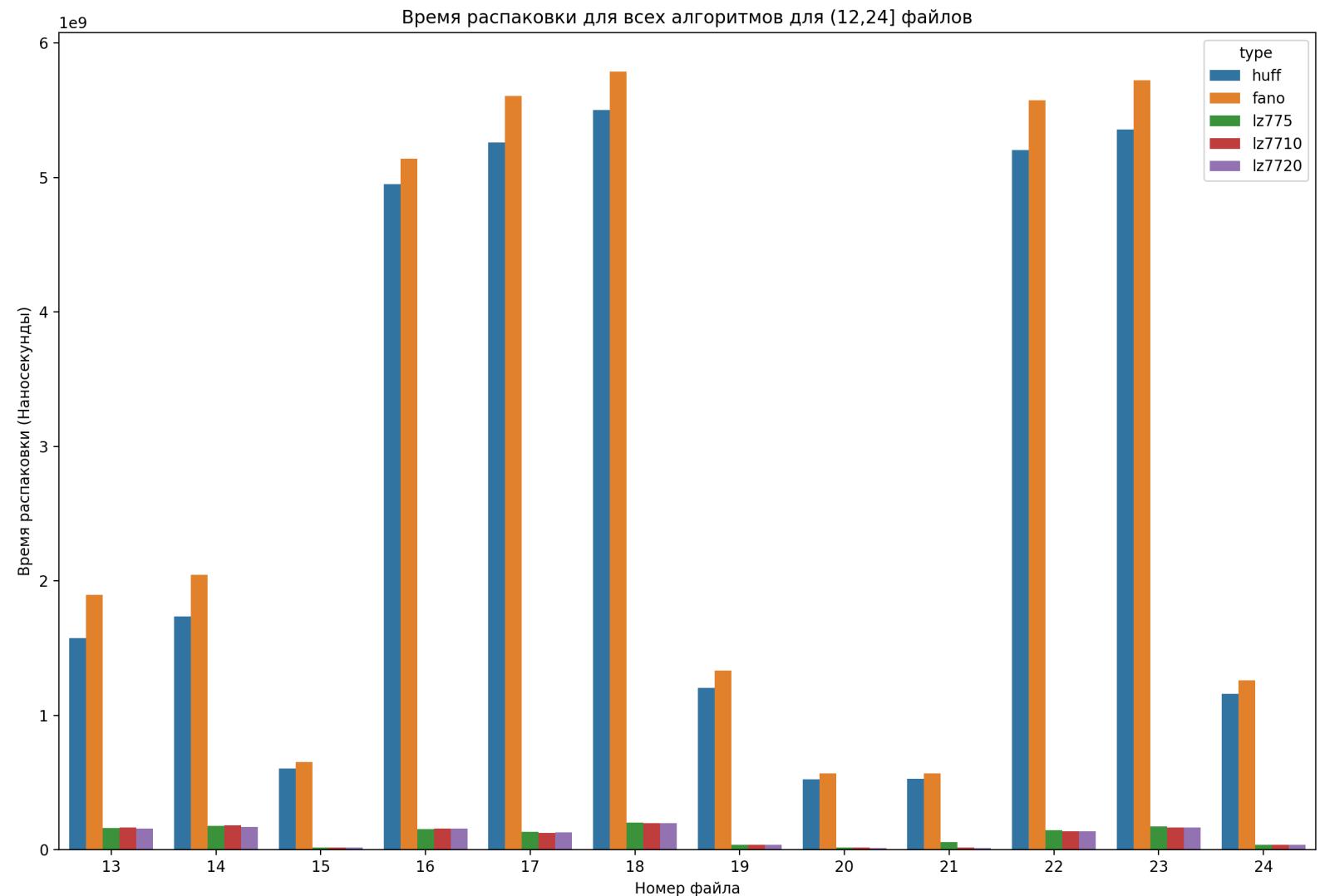


## Время разжатия для каждого файла для каждого алгоритма



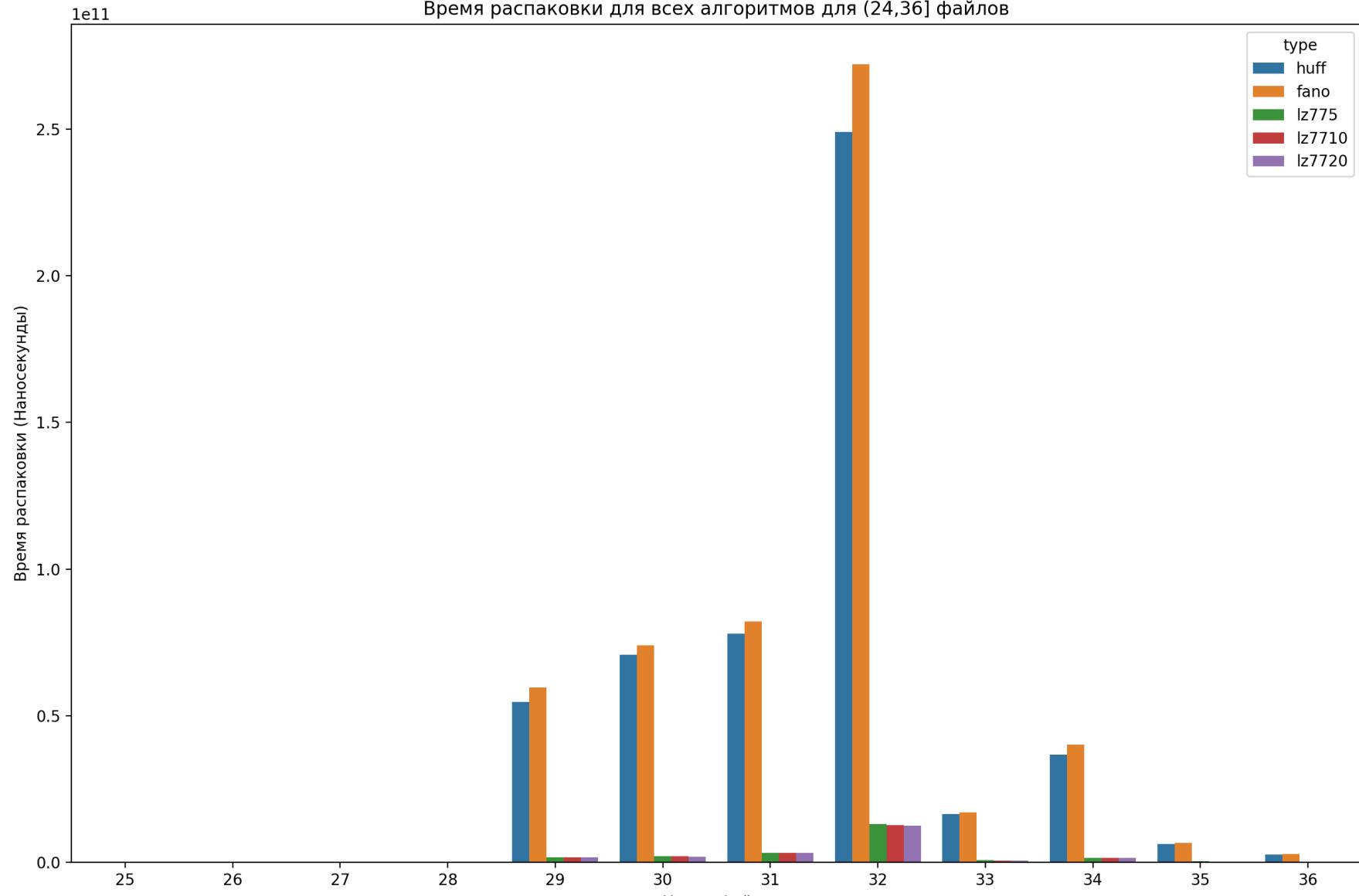
При разжатии картина сильно меняется в сравнении со сжатием. Теперь LZ работает быстро, а Хаффман и Фано проигрывают в 10ки раз. Это объяснимо — им приходится перебирать коды, для поиска первого совпадения. Однако, Хаффман работает все-таки немнога быстрее, чем Фано, потому что он создает оптимальные коды.





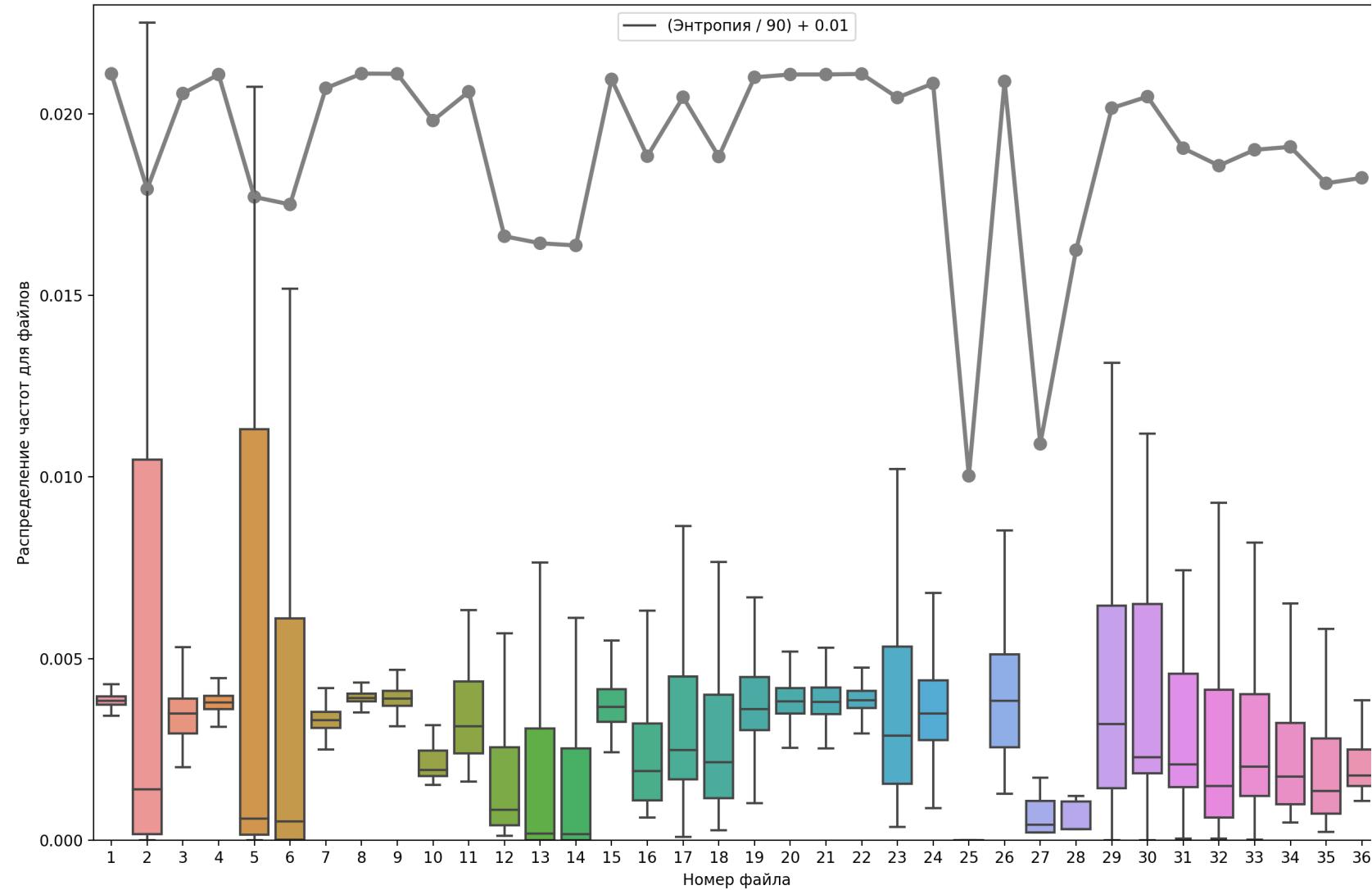
25й, 26й, 27й и 28й  
очень маленькие,  
поэтому их в  
масштабах других  
файлов даже не  
видно.  
А так картина как и  
у предыдущих  
файлов.

Время распаковки для всех алгоритмов для (24,36] файлов



## Сравнение распределений частот файлов и энтропии

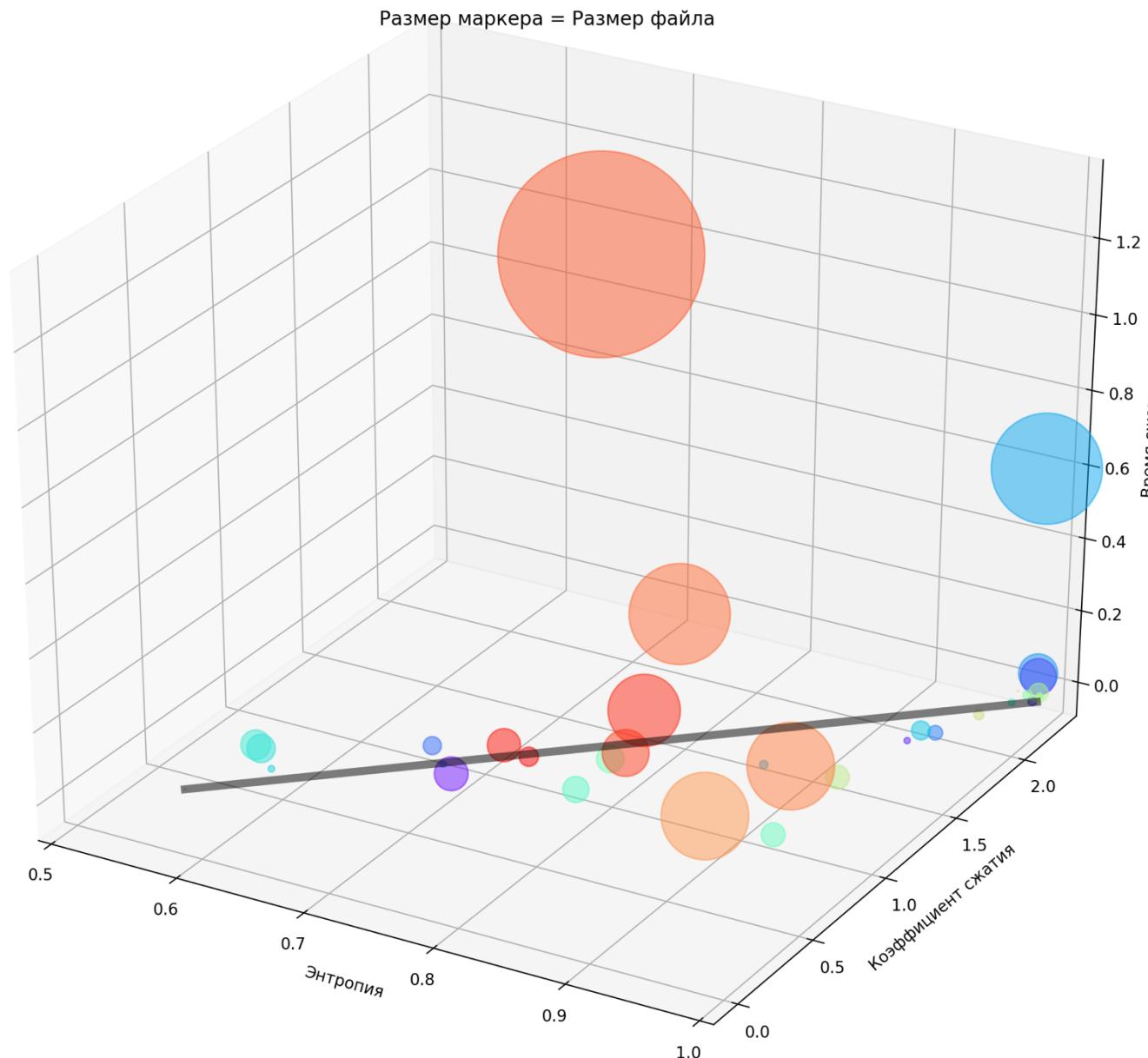
Такой график, во-первых позволяет показать все 36 распределений на одном графике, а не на 36ти, во-вторых можно смотреть на взаимосвязь характеристик распределения и энтропии(нормированной). Например, первое, что бросается в глаза — зависимость медианы и энтропии (чем ближе медиана к первой квартили относительно интерквартильного размаха — тем меньше энтропия).



## Зависимости энтропии, коэффициента сжатия, времени сжатия и размера файла

Такой 4D график содержит в себе сразу несколько направлений информации:

- 1) Можно заметить корреляцию коэффициента сжатия от энтропии, и при этом на эту зависимость никак не влияют размеры файлов.
- 2) Этот график также дает понять, что время сжатия не коррелирует, ни с коэффициентом сжатия, ни с энтропией — большее влияние оказывает именно размер файла.



## Зависимости коэффициента сжатия от энтропии для каждого файла для каждого алгоритма

График позволяет посмотреть на то, как коэффициент сжатия каждого алгоритма зависит от энтропии файла.

