ОТЧЕТ
Фортов Егор, БПИ214
КДЗ #2

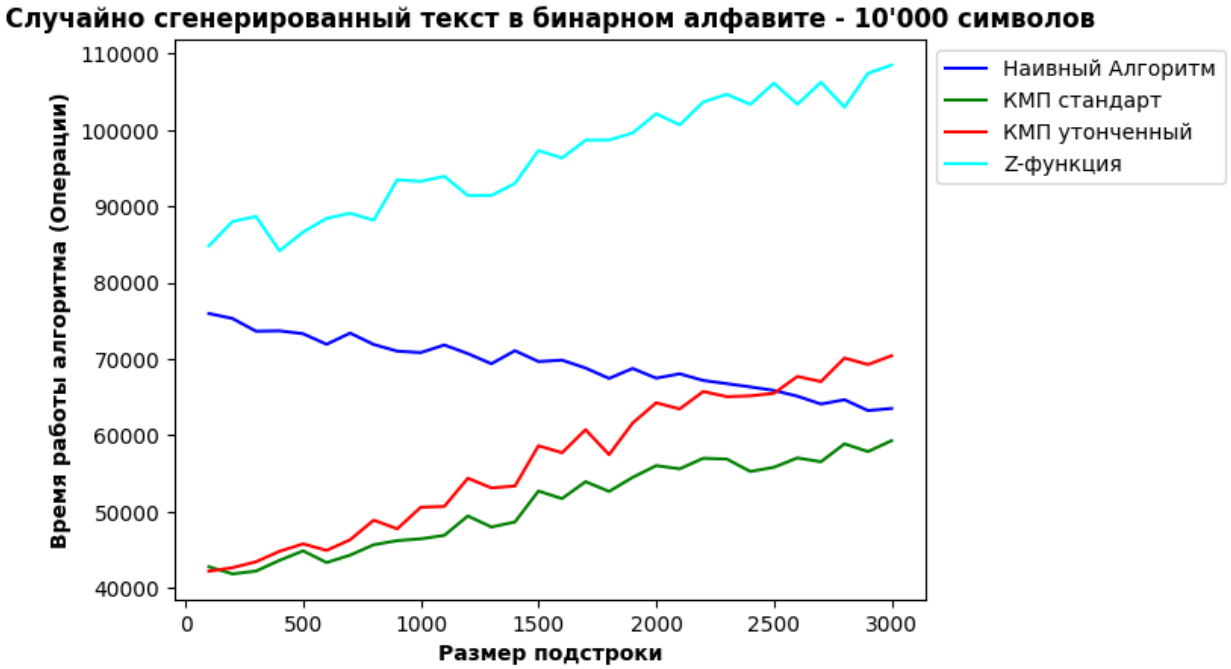
Ключевые интересные наблюдения, сделанные после анализа графиков, построенных в рамках требований задания:

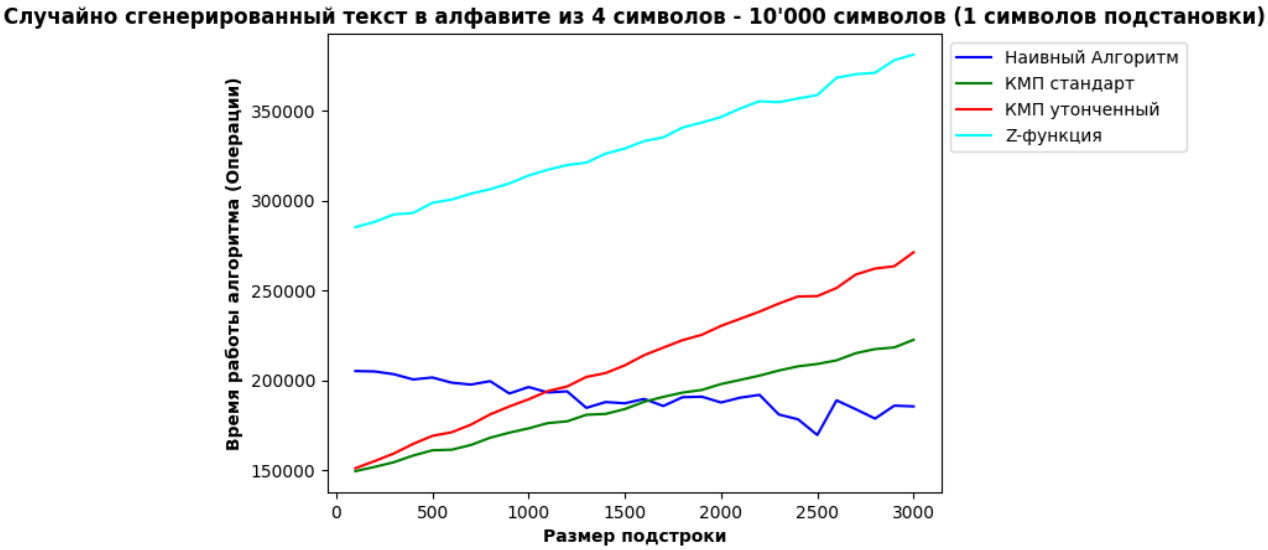
1) Z-функция имеет большую константу, из-за чего работает она дольше других алгоритмов.

2) Наивный алгоритм на первый взгляд ведет себя странно: при бОльших данных он работает чуть быстрее. Объясняется это тем, что, во-первых, при росте размера паттерна поиска не растет размер основной строки. Во-вторых, алфавит бинарный => очень часто при сравнении паттерна поиска будет раньше происходить break, что приводит к меньшему числу дальнейших сравнений. И в-третьих, размеры подстроки и строки сильно отличаются (100-3000 VS 10К/100К) .

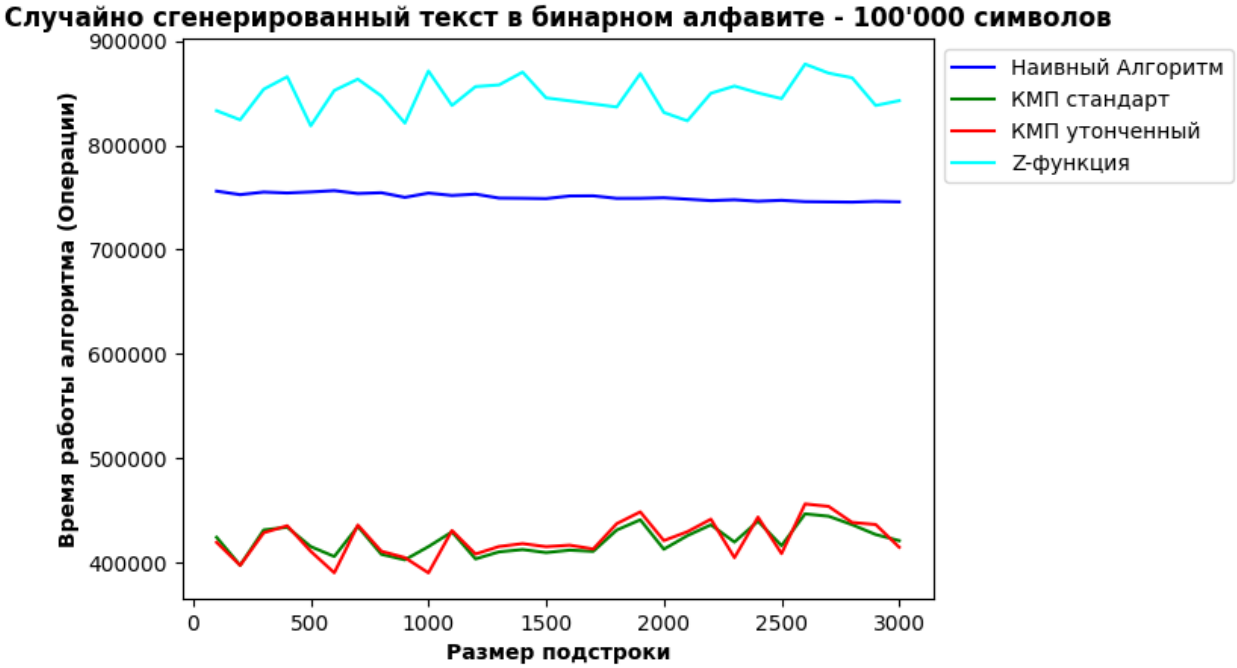
3) Сильные скачки (например, в обоих реализациях алгоритма КМП) объясняются тем, что у нас критически малая мощность входного алфавита (0 и 1).

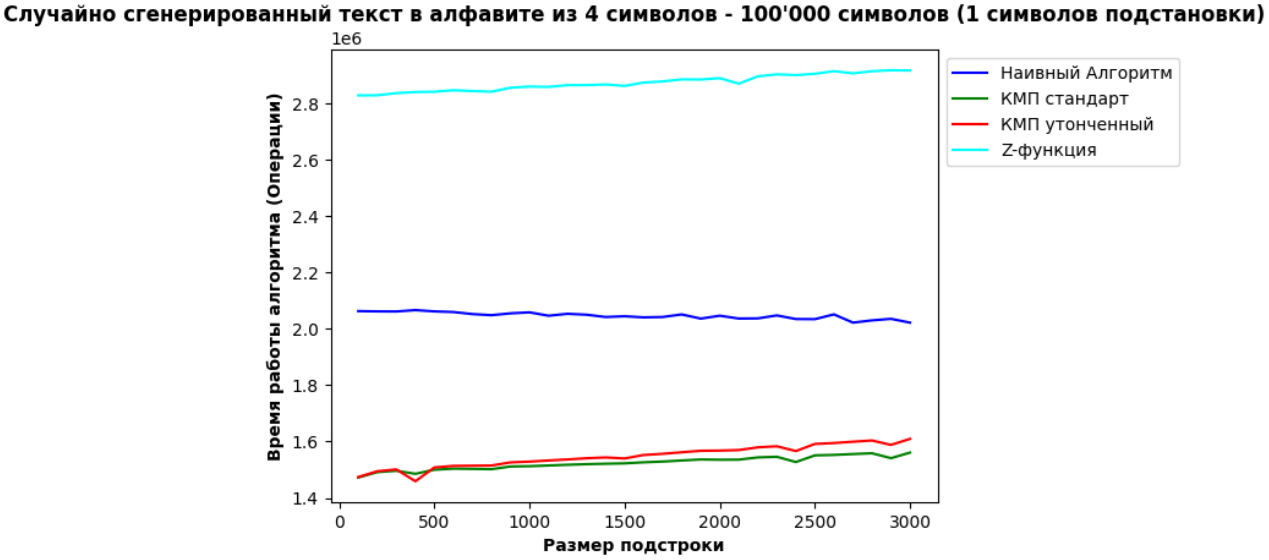
График для наблюдений 1-3:



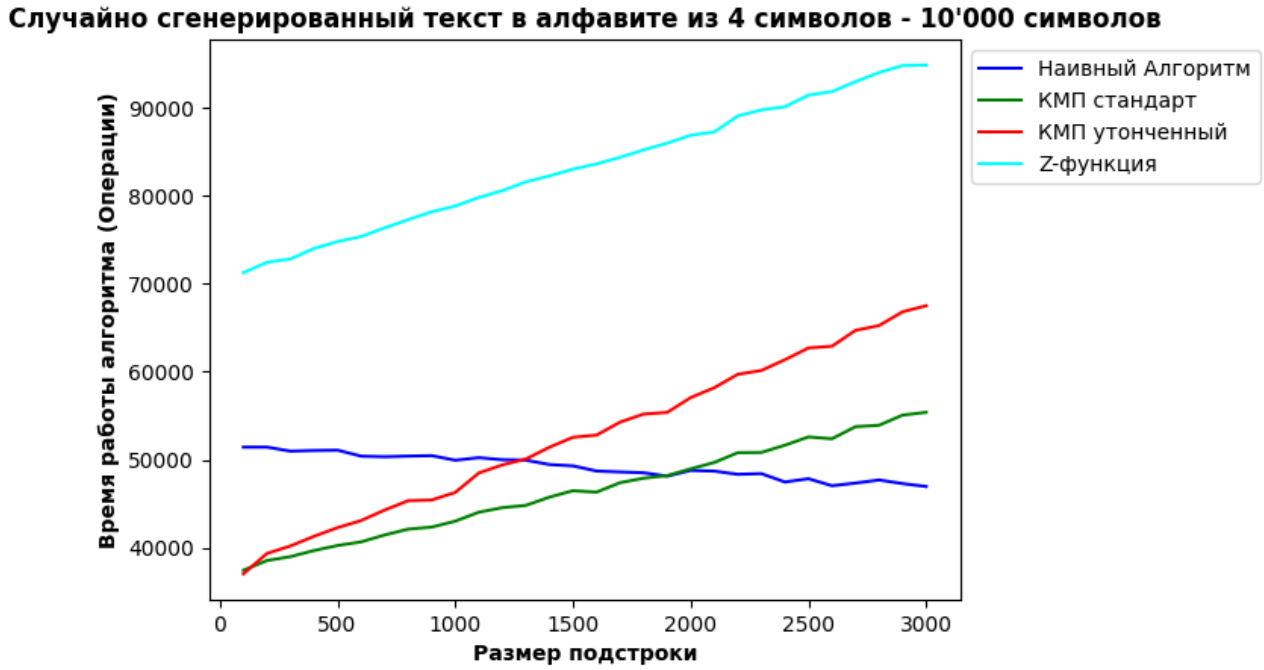


4) График наивного алгоритма вырождается в горизонтальную линию - это объясняется тем, что размер данных существенно вырос (про скачки утверждение то же: чем больше алфавит и основная строка, тем меньше скачки):

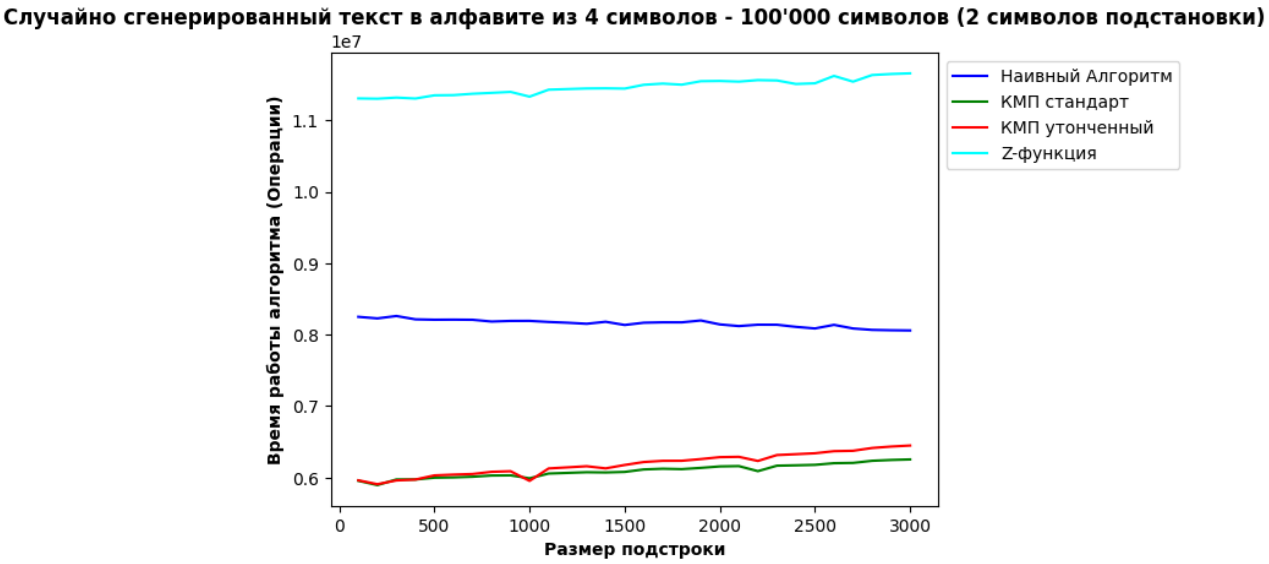


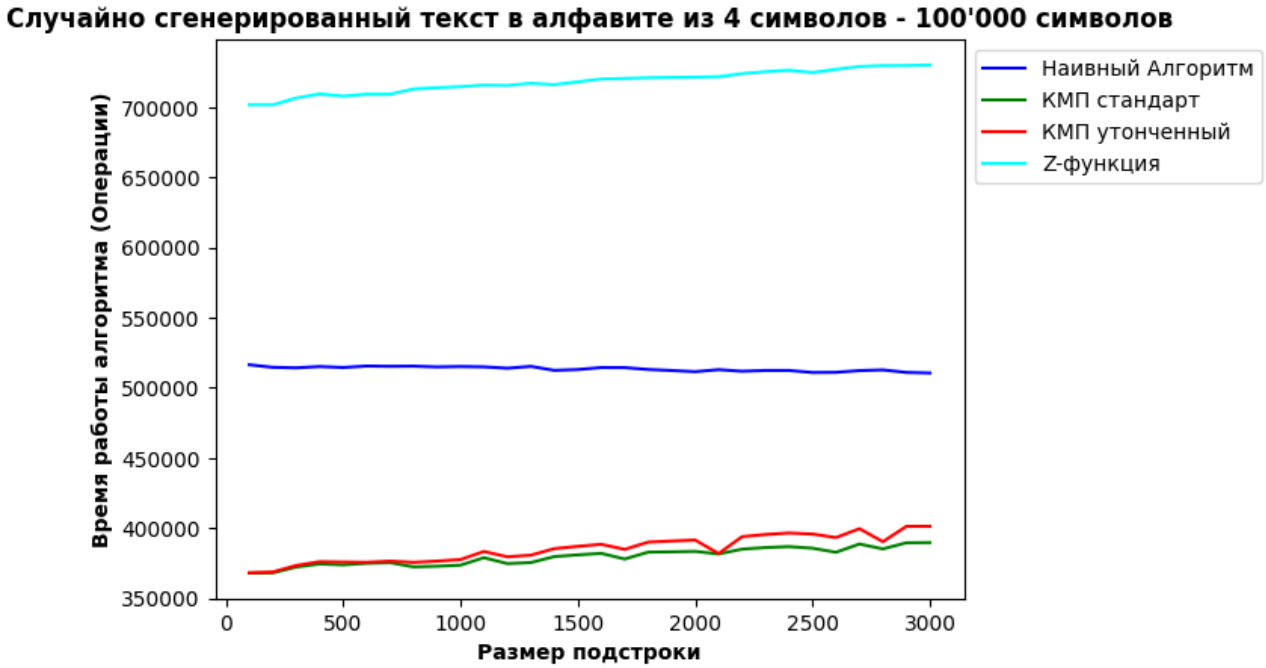


5) КМП (обе реализации) вырождаются в линию, что подтверждает их линейную сложность работы:

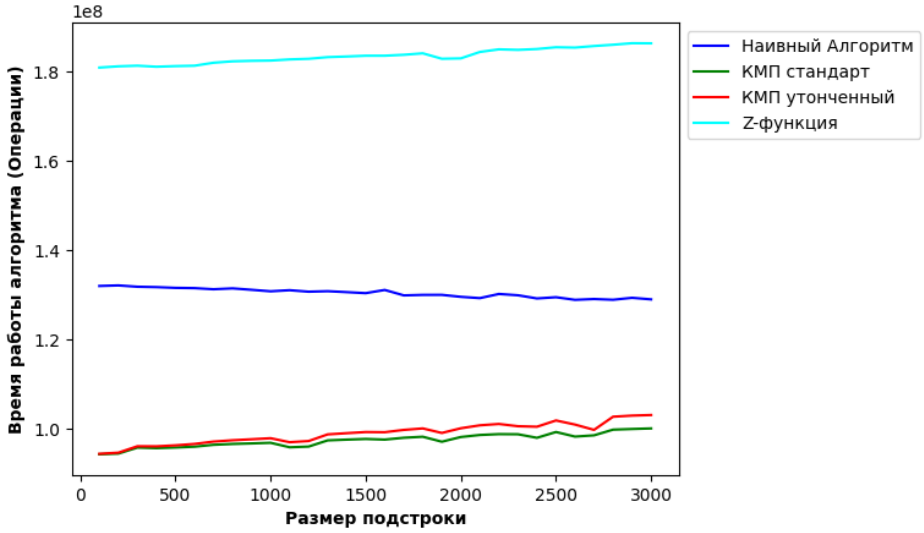


6) Скачки минимальны. линия. То, что подтверждает линейная сложность.



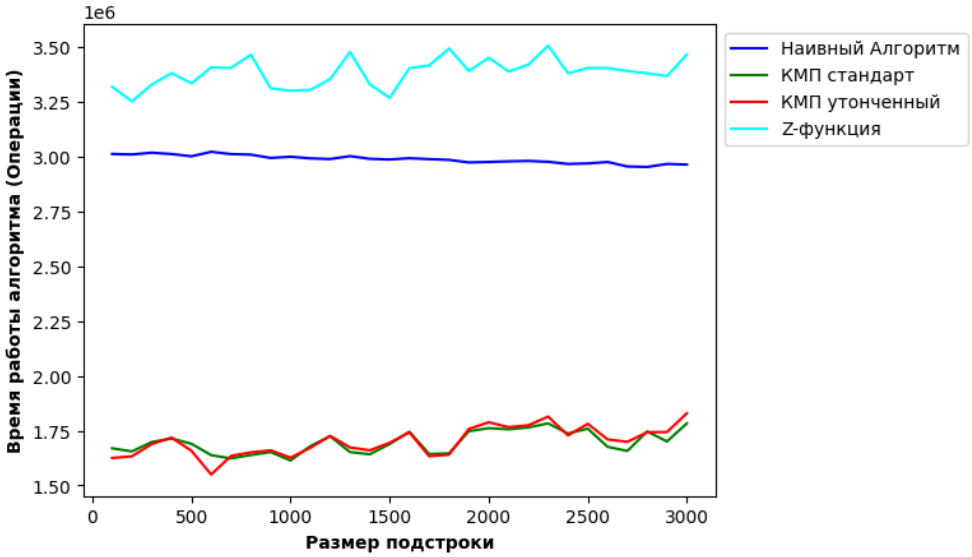


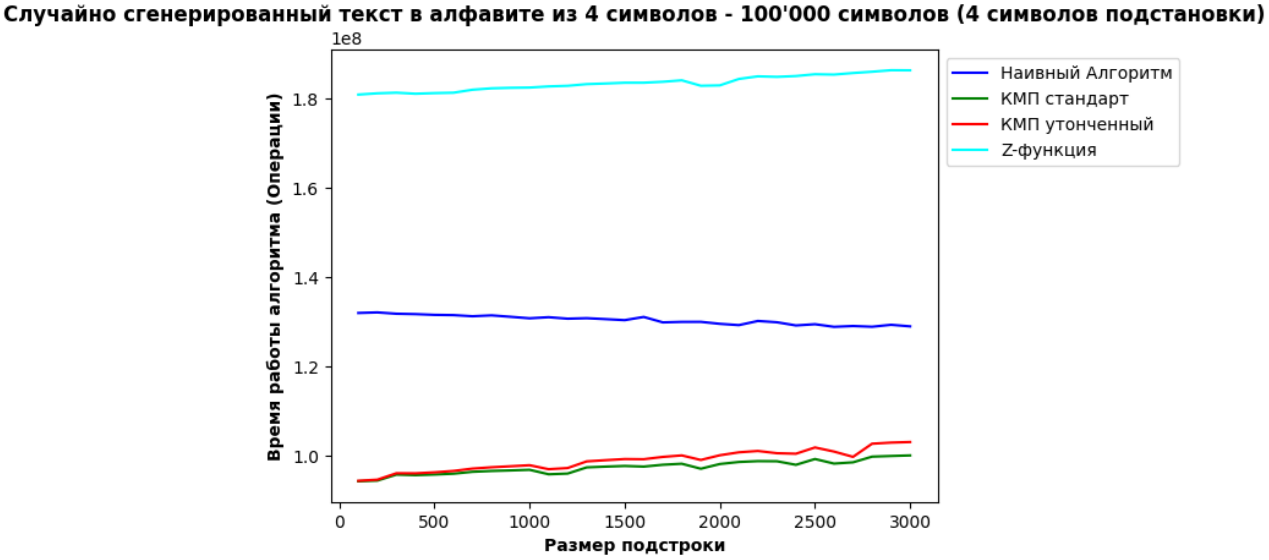
Случайно сгенерированный текст в алфавите из 4 символов - 100'000 символов (4 символов подстановки)



7) Строка больше => скачки меньше:

Случайно сгенерированный текст в бинарном алфавите - 100'000 символов (1 символов подстановки)





8) Удачные входные данные заставляют работать алгоритм намного быстрее, из-за чего на графиках появляются впадины:

9) КМП с утонченными гранями при наших входных данных работает хуже, чем КМП со стандартными гранями (это видно из всех предыдущих графиков). Это объясняется тем, что данные недостаточно большие и вычисление массива brs в случае утонченных граней занимает немало времени.

10) Из графиков видно, что графики алгоритмов КМП с утонченными гранями и КМП со стандартными гранями сильно похожи (практически все изгибы повторяются). Это объясняется тем, что по сути эти алгоритмы работают одинкаково, единственное их различие — в том, что КМП с утонченными гранями считает дополнительный массив brs. Именно эта деталь дает графику оптимизированного КМП чуть большую константу, что видно из вышеприведенных графиков.

ОБЩИЕ ВЫВОДЫ:

- Несмотря на преимущество по асимптотической сложности алгоритмов КМП и Z-функции перед нативным алгоритмом, последний ведет себя отнюдь неплохо. Стоит учесть, что при наших входным данных этот алгоритм работает тем быстрее, чем больше входные данные (объяснение см. в наблюдении №2), поэтому в подобных случаях (когда строка фикс. Длины, а паттерн все время увеличивается + алфавит очень маленький), стоит использовать именно его.
- Несмотря на одинаковую асимптотику двух реализаций КМП и Z-функции, последняя имеет большую константу, что отражается на ее «не самых лучших» результатах.
- Несмотря на оптимизацию алгоритма КМП (КМП с утонченными гранями), его следует использовать на много бОльших данных, когда длина строки >1млн и длина подстроки >500000. Использование оптимизации на наших входных данных нецелесообразно: массив brs считается долго, именно поэтому этот алгоритм ведет себя хуже, чем его стандартный, неоптимизированный вариант.