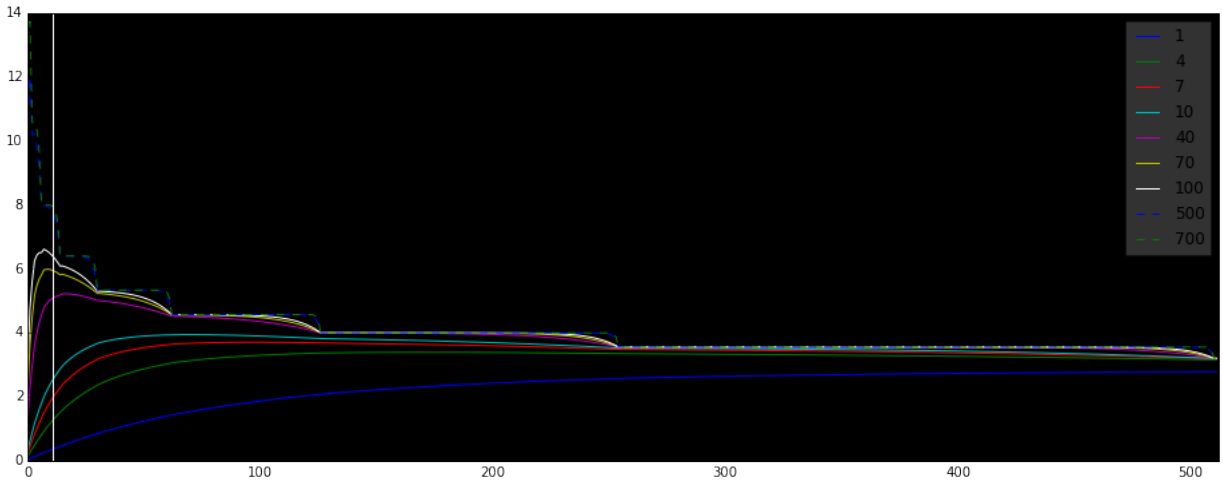
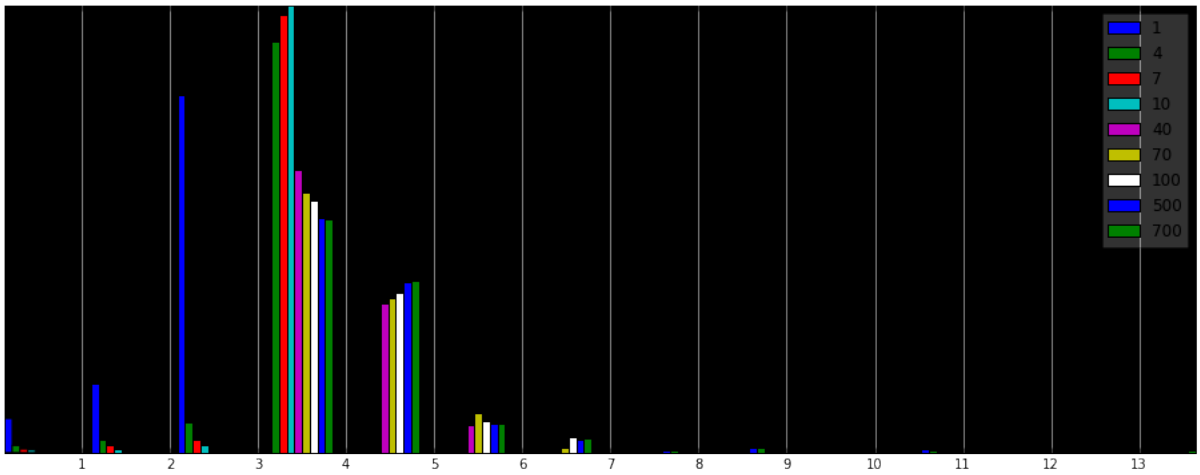
Арзуманян Виталий, CS, 2 курс

Распределение степеней сжатия

Для параметра параметра m в интервале от 1 до 512:



Гистограмма распределения степеней сжатия:



Оптимальный параметр - для которого средняя степень сжатия максимальна (т.к. начальный размер одинаков) - $m_opt = 12$. Степени сжатия: 4.36, степени сжатия листов:

$$[0.36494337, 1.27794159, 1.98655463, 2.55080547, 5.09261153,$$

$$5.94558047, 6.38075691, 7.75722598, 7.94622788]$$

Оценка параметра распределения

Т.к. длина постинг-листов одинаковая, считаем самым длинным меньше всего сжатый с оптимальным параметром. Подберем параметры распределения по максимуму правдоподобия (здесь геометрическое разделение в numpy немного отличается от геометрического распределения в R - начинается с 1, а в R, судя по сгенерированным данным, с нуля):

$$L = \Pi(1-p)^{X_i}p$$

$$l = N\log p + \sum(X_i - 1)\log(1-p)$$

$$l' = 0 = \frac{N}{p} - \frac{\sum X_i - N}{1-p}$$

$$N(1-p) = p\sum X_i - NP$$

Получаем оценку максимального правдоподобия:

$$\widehat{p} = \frac{N}{\sum X_i}$$

Полученное значение $p = 0.0010005$, что соответствует параметру $p = 0.001$ генерации распределения.

Оптимальный параметр кода

Для этого кода оптимальный параметр из перебираемых 512 - максимальный. Степень сжатия 2.78.

Теоретическая оценка

Теоретическая оценка дает m=1.

Очевидно, что для некоторых p ($p+p^2 < 1$) будет оценка $m = 1$ и так же что при этом кодировании длина кода будет длиной унарной записи числа + 1, что будет длинее начальной записи числа. Следовательно, для таких p теоретическая оценка некорректна - даже простая унарная запись будет более оптимальным кодом.

□