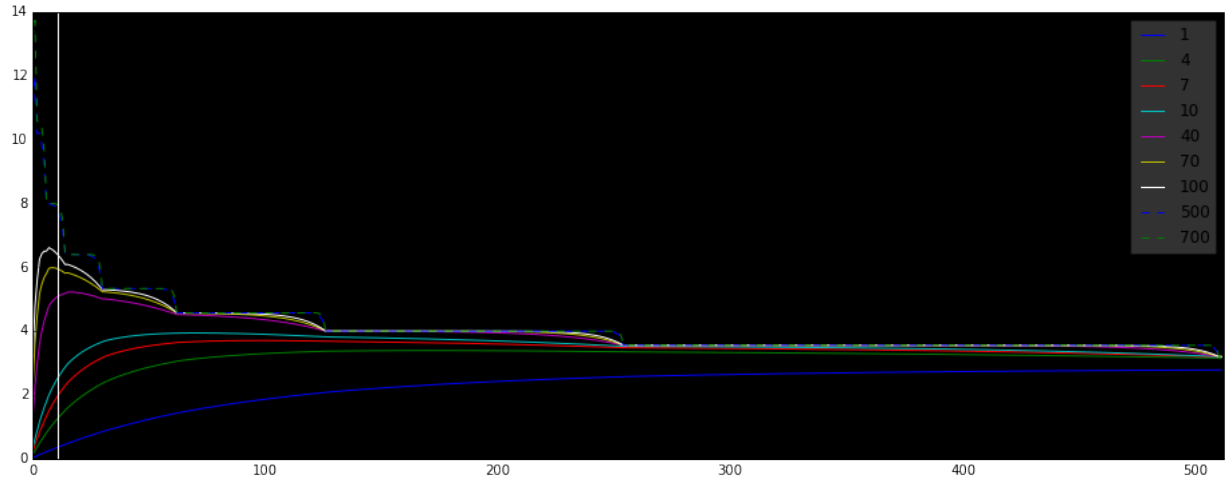
Арзуманян Виталий, CS, 2 курс

Работа на смоделированных данный

Для параметра параметра m в интервале от 1 до 512 на сгенерированных выборках:



Оптимальное значение $m = 12$, отмечено белым.

Работа на большом индексе

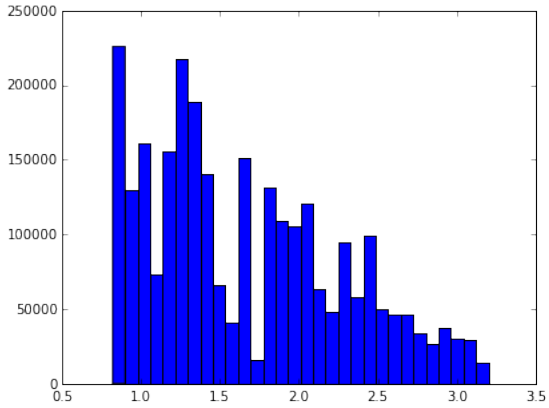
Предобработка

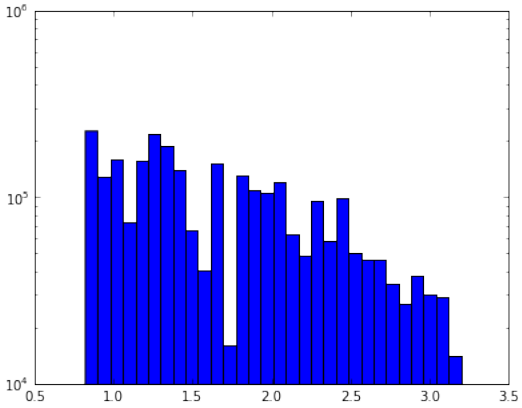
Индекс состоит из записей вида: номер слова, номер документа:количество. Мы проходим по индексу и получаем список массивов, в каждом из которых номер документа повторяется столько раз, сколько в нем встречается заданное слово. Далее список документов упорядочиваетя по возрастанию. На вход для кодирования подаются эти списки. Оптимальный параметр для ускорения подбираем сначала две итерации по степеням двойки в предполагаемом диапазоне, затем с некоторым шагом на отрезке предполагаемого минимума. Минимум рассматриваем для суммарного размера сжатого индекса (не рассчитывая степени сжатия).

Оптимальный параметр - для которого минимальный размер. $m = 988$

Гистограмма распределения степеней сжатия для этого параметра (слева - линейная шкала, справа - логарифмическая. Отношения полно-

го размера к размеру сжатого индекса от 0.82 до 3.2, большая часть в диапазоне до 1.5, дальше линейно убывают):





Оценка параметра распределения

Самый длинный постинг-лист номер 209362 (начиная с 0), состоит из $N = 435272$ записей. Подберем параметры распределения по максимуму правдоподобия:

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

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

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

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

Получаем оценку максимального правдоподобия:

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

Полученное значение $p = 0.939304$.

Оптимальный параметр кода

Для этого кода оптимальный параметр из перебираемых (до $2^{16} - 1$) $m = 1$. Степень сжатия 30.

Теоретическая оценка

Теоретическая оценка дает $m = 11$ для такого p.

Причина расхождения в специфическом виде последовательности - она состоит из большого количества нулей и больших чисел, т.е. слово встречается в каком-то документе несколько раз, потом некоторое время не встречается, затем снова несколько раз и т.д. Оптимальный параметр 1 понятен - выигрыш за счет экономии кодирования (нет остатка - код длины 0) на большом количестве нулей (434619 из 435272) перебивает потери за счет унарного кодирования нескольких чисел.

Эта последовательность маловероятна при любом p, а теоретическая оценка дается для типичных последовательностей.

□