

Ссылка на GitHub: <https://github.com/LiDa06/Algorithms/tree/main/SET5/A3>

### Цель работы

Целью работы является реализация стандартного алгоритма HyperLogLog для оценки количества различных элементов (кардинальности,  $F_0$ ) в потоке данных, а также экспериментальная проверка его точности и соответствия теоретическим оценкам ошибки.

### Описание эксперимента

Был сгенерирован поток данных, состоящий из попарно различных элементов, с постепенно увеличивающимся размером от 50 000 до 1 000 000 элементов.

Для каждого размера потока:

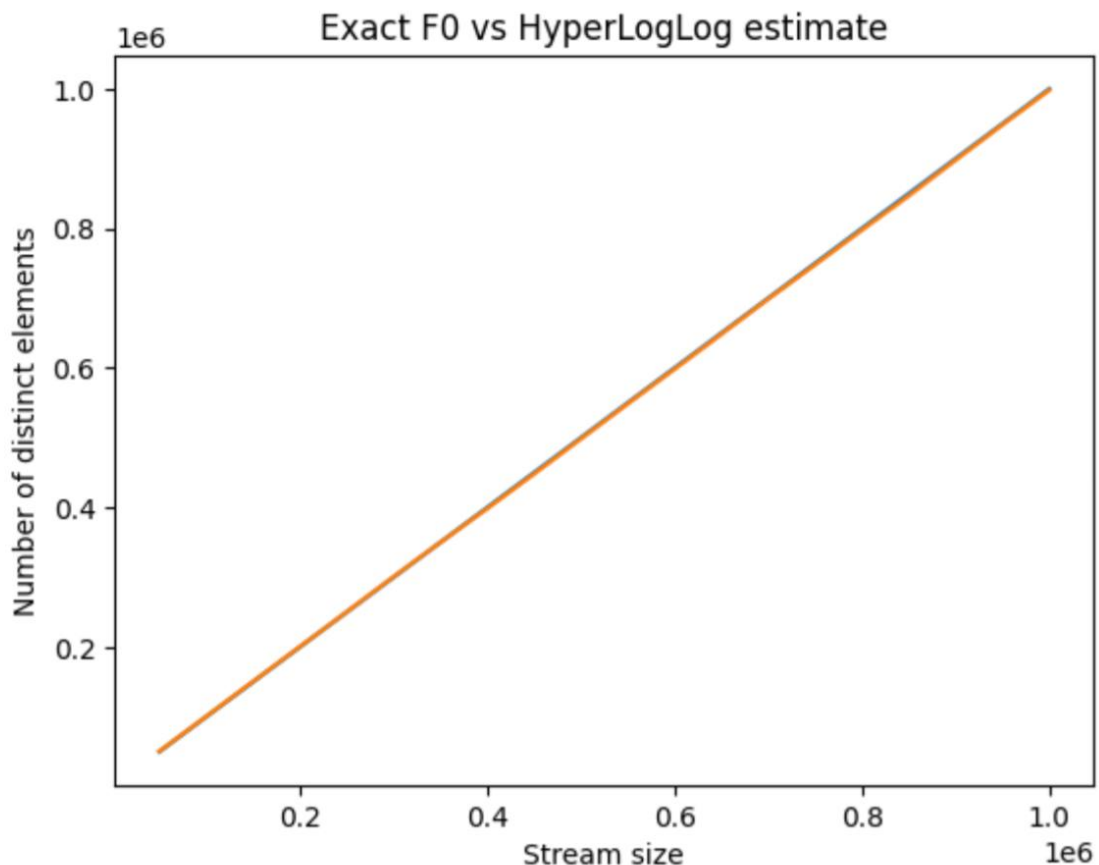
1. вычислялось точное значение  $F_0$
2. выполнялась оценка кардинальности алгоритмом HyperLogLog;
3. оценка усреднялась по нескольким независимым запускам;
4. вычислялись стандартное отклонение и относительная ошибка.

Число регистров HyperLogLog:  $m=16384$

Теоретическая относительная ошибка алгоритма:  $\varepsilon \approx \frac{1.04}{\sqrt{m}}$

### Результаты эксперимента

**График 1. Сравнение точного значения  $F_0$  и оценки HyperLogLog**



На первом графике показано сравнение:

1. точного количества различных элементов  $F_0$ ;
2. средней оценки, полученной алгоритмом HyperLogLog.

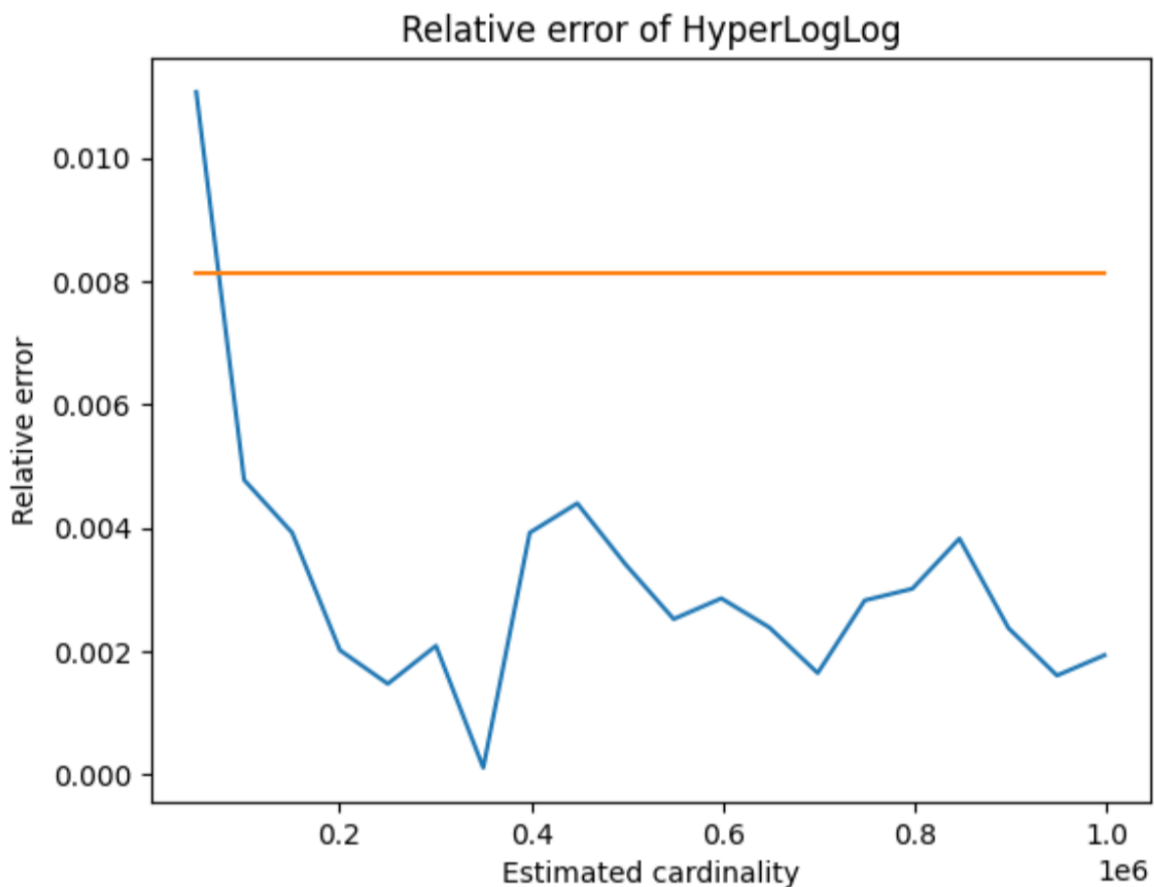
**Наблюдения:**

1. Графики практически совпадают на всём диапазоне значений.
2. Оценка HyperLogLog является несмещённой.
3. С ростом размера потока расхождение между оценкой и точным значением не увеличивается.

**Вывод:**

Алгоритм HyperLogLog корректно аппроксимирует значение  $F_0$  даже для больших потоков данных.

**График 2. Относительная ошибка оценки**



На втором графике показана относительная ошибка:  $\frac{|F'_0 - F_0|}{F_0}$

в зависимости от оценки кардинальности.

Также на графике отображена теоретическая граница ошибки  $\epsilon$

**Наблюдения:**

1. Экспериментальная относительная ошибка практически всегда ниже теоретической оценки.
2. Максимальная ошибка наблюдается на малых значениях потока и не превышает  $\sim 1.1\%$ .
3. При увеличении кардинальности ошибка стабилизируется в диапазоне  $0.1\text{--}0.4\%$ .

**Вывод:**

Экспериментальные результаты полностью согласуются с теоретическими оценками точности HyperLogLog.