Dynamic Indexing

For n = 2 and $1 \leq T \leq 30$, perform a step-by-step simulation of the algorithm on page 38 of the lecture "index construction". Create a table that shows, for each point in time at which $T = 2 * k$ tokens have been processed ($1 \leq k \leq 15$), which of the three indexes $I_0, ..., I_3$ are in use.

Map Reduce

- Wie kann das Map-Reduce Programmier Paradigma verwendet werden, um für jeden Term in einem Korpus die Korpus-Frequenz *cf* zu berechnen.
- Spezifizieren Sie den Input und Output der beiden Phasen.

Heaps' law

- Looking at a collection of web pages, you find that there are 3000 different terms in the first 10,000 tokens and 30,000 different terms in the first 1,000,000 tokens.
- Assume a search engine indexes a total of 20,000,000,000 (2×10^{10}) pages, containing 200 tokens on average
- What is the size of the vocabulary of the indexed collection as predicted by Heaps' law?
- Heaps' law: $M = k * T^b$
  with M: Vocabulary size
  T: Number of Tokens
  k, b: Parameters