Information Retrieval
Topic- Index Compression (Term statistics)
Lecture-20

Prepared By

Dr. Rasmita Rautray & Dr. Rasmita Dash
Associate Professor
Dept. of CSE

Content

- Term statistics
- Heap’s law
- Zipf’s law

Why compression? (in general)

- Use less disk space (saves money)
- Keep more stuff in memory (increases speed)
- Increase speed of transferring data from disk to memory (again, increases speed)
  - [read compressed data and decompress in memory] is faster than [read uncompressed data]
- Decompression algorithms are fast.

Why compression in information retrieval?

- First, we will consider space for dictionary
  - Main motivation for dictionary compression: make it small enough to keep in main memory
- Then for the postings file
  - Motivation: reduce disk space needed, decrease time needed to read from disk

Lossy vs. lossless compression

- Lossy compression: Discard some information.
- Lossless compression: All information is preserved.

Term Statistics

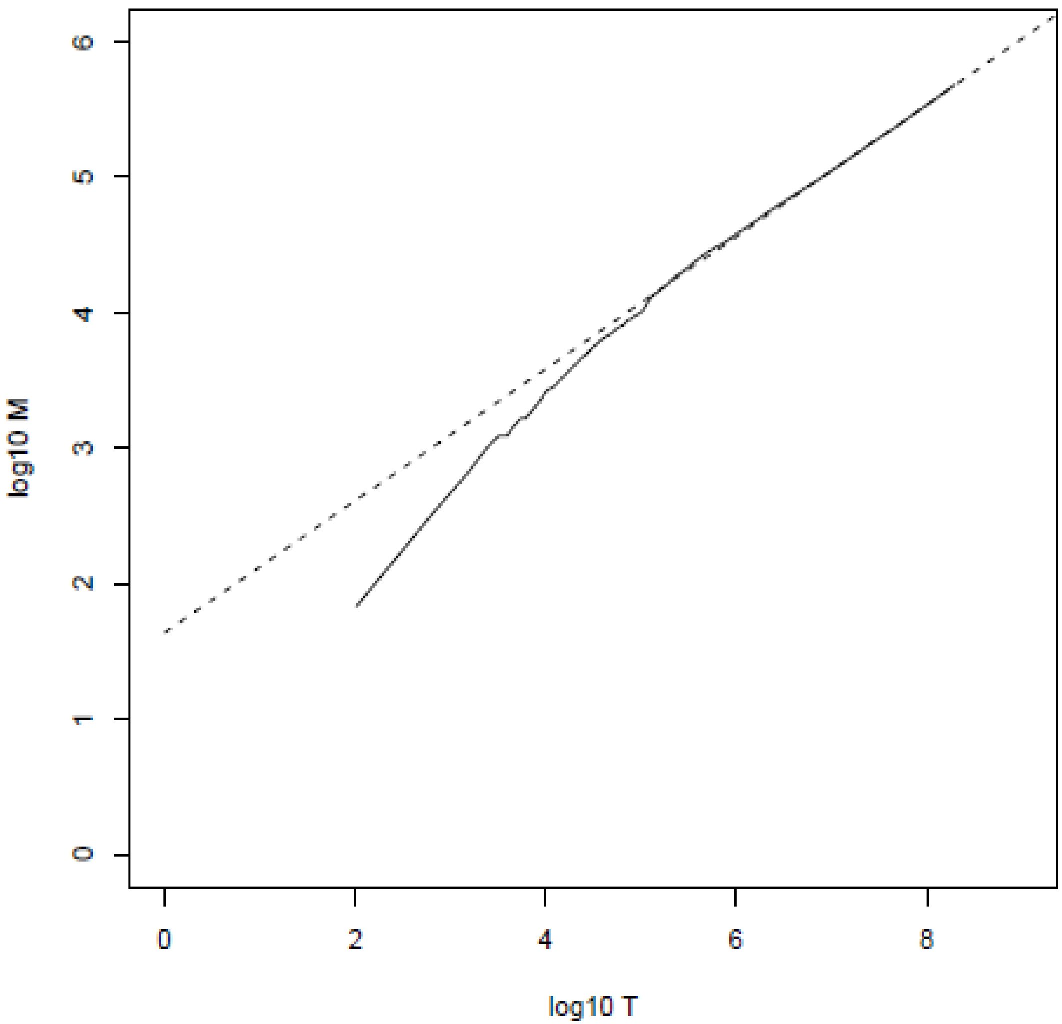
How big is the term vocabulary?

- That is, how many distinct words are there?
- Can we assume there is an upper bound?
- Not really: At least 7020 ≈ 1037 different words of length 20.
- The vocabulary will keep growing with collection size.

Heaps' law

- Heaps' law: $M = kT^b$
- M is the size of the vocabulary, T is the number of tokens in the collection.
- Typical values for the parameters k and b are: $30 \leq k \leq 100$ and $b \approx 0.5$.
- Heaps' law is linear in log-log space.
  - It is the simplest possible relationship between collection size and vocabulary size in log-log space.
  - Empirical law

Heaps' law for Reuters



Vocabulary size M as a function of collection size T (number of tokens) for Reuters-RCV1. For these data, the dashed line $\log_{10} M = 0.49 * \log_{10} T + 1.64$ is the best least squares fit. Thus, $M = 10^{1.64} T^{0.49}$ and $k = 10^{1.64} \approx 44$ and $b = 0.49$.

Empirical fit for Reuters

- Example: for the first 1,000,020 tokens Heaps' law predicts 38,323 terms:

$$44 \times 1000020^{0.49} \approx 38{,}323$$

- The actual number is 38,365 terms, very close to the prediction.
- Empirical observation: fit is good in general.

Zipf's law

Zipf's law: The i^{th} most frequent term has frequency proportional to $1/i$.

$\mathrm{cf}_i \propto \frac{1}{i}$

cf is collection frequency: the number of occurrences of the term in the collection.

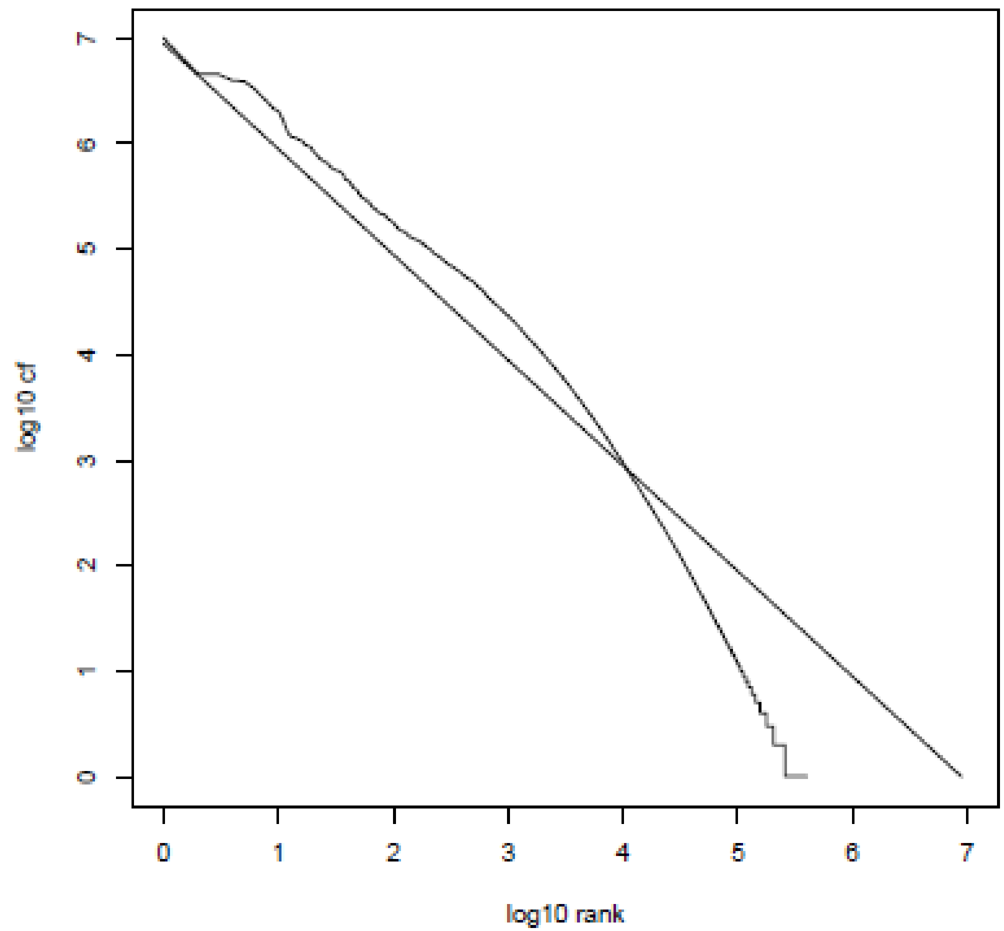
So if the most frequent term (*the*) occurs cf_1 times, then the second most frequent term (*of*) has half as many occurrences $\mathrm{cf}_2 = \frac{1}{2}\mathrm{cf}_1 \quad \ldots$

…and the third most frequent term (*and*) has a third as many occurrences $\mathrm{cf}_3 = \frac{1}{3}\mathrm{cf}_1$ etc.

Equivalent: $\mathrm{cf}_i = ci^k$ and $\log \mathrm{cf}_i = \log c + k \log i$ (for $k = -1$)

Example of a power law

Zipf's law for Reuters



Fit is not great. What is important is the key insight: Few frequent terms, many rare terms.

Thank You