INFORMATION RETRIEVAL

Week 7 – Ranked Retrieval

11.04.2025

Severin Mills

Today

1

Exercise Recap

- Discussion
- Questions

2

Theory

- Recap: Heap’s / Zipf’s Law
- Parametric Search
- Shared inverted index
- Scores / Weights

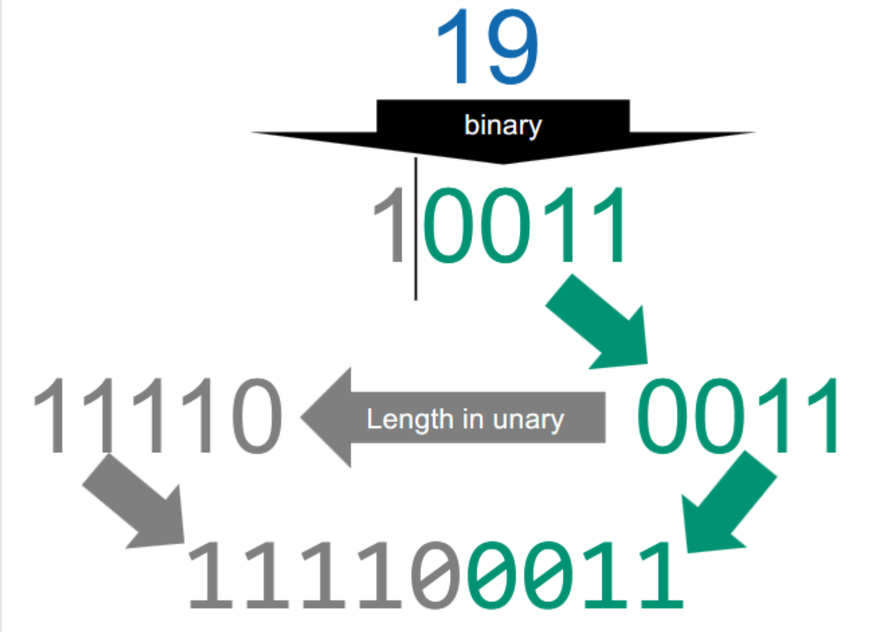
3

Kahoot / Exam questions

- Exercise 4: Heap’s and Zipf’s Law

Gamma code mapping

Recall:



Gamma code mapping

Example: 7

1. To binary: 111
2. Trim leading 1: 11
3. Unary length: 110
4. Combine: 11011

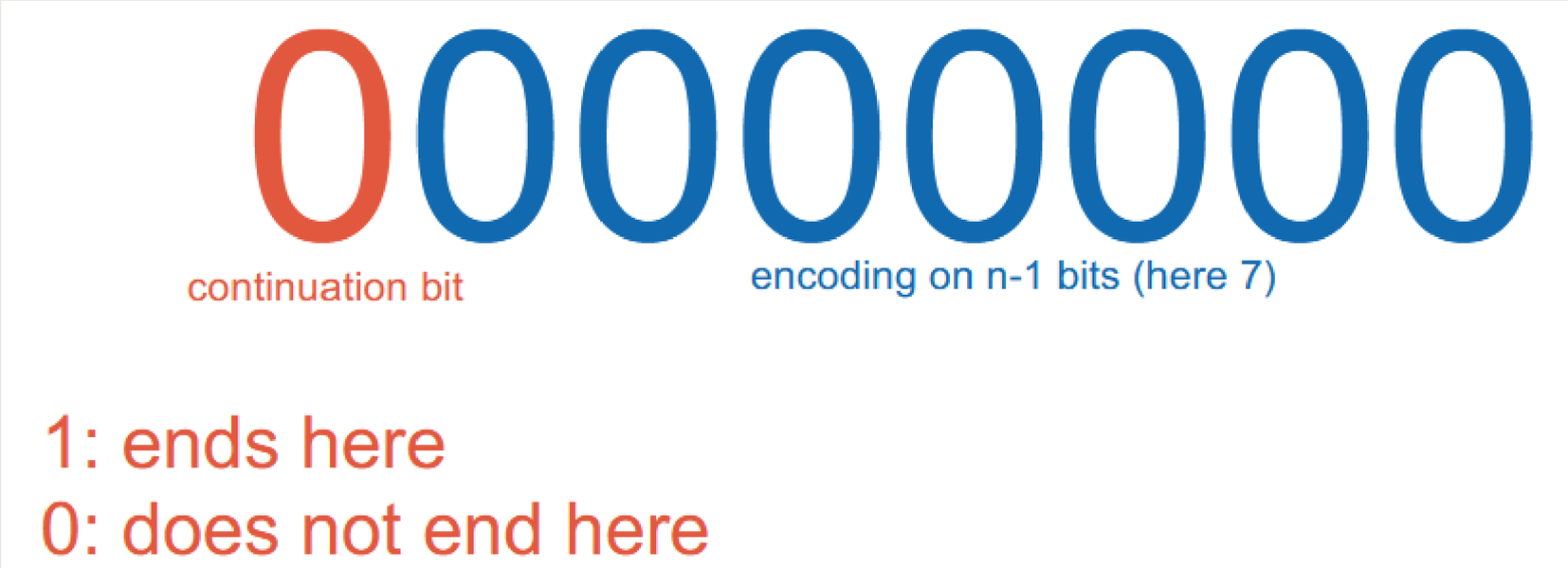
Gamma code mapping

Example: 12

1. 1100
2. 100
3. 1110
4. 1110100

Variable byte encoding

Recall:



Variable byte encoding

Example: 356, 8-bit packets

1. To binary: 101100100
2. Split up into 7-bit segments: 10 1100100
3. Add continuation / termination bits: 00000010 1100100

Variable byte encoding

Example: 356, 8-bit packets

1. To binary: 101100100
2. Split up into 7-bit segments: 10 1100100
3. Add continuation / termination bits: 00000010 1100100

Variable byte encoding

Example: 46

1. To binary: 101110
2. Split up into 7-bit segments: Not needed
3. Add continuation / termination bits: 10101110

Variable byte encoding

Example: 767 (Binary: 1011111111)

1. To binary: 1011111111
2. Split up into 7-bit segments: 101 1111111
3. Add continuation / termination bits: 00000101 11111111

Ordering encoding methods

Gamma encoding: Always starts with a string of 1s, middle bit is 0, always odd size

Unary encoding: The value of the number + 1

Fix-length encoding: Number in binary rounded up to the next multiple of the packet size

Variable length: If n := Number in binary and p := packet size, then the size is ceil(| n | / (p – 1)) * p

Largest code points

Variable length encoding: Check packet size, remove first bit of every segment

Fix-length: Size of maximum encoding

Unary: size – 1

Gamma: If n is the size, the largest number that fits in ceil((n – 1) / 2)

Discussion

Lot of exam questions! Practice is key here.

Terminology

N: number of documents
T: number of tokens (positional postings)
M: number of terms

Heap's law

$M = k\sqrt{T}$

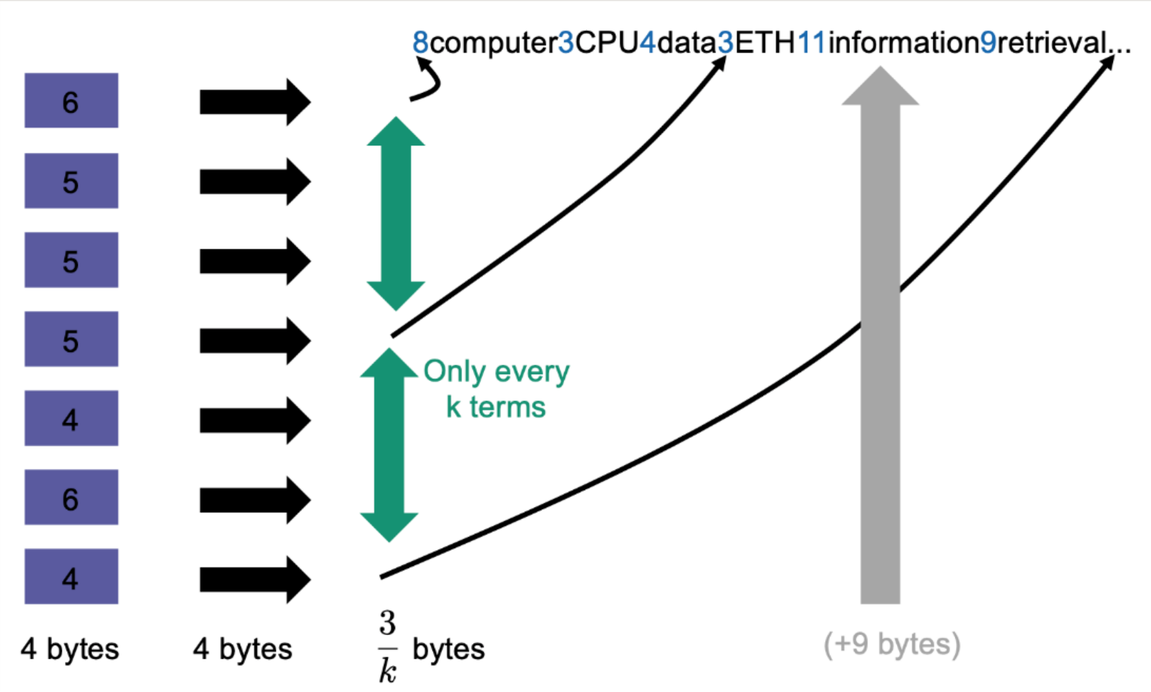
Empirically, $30 \leq k \leq 100$

Zipf's law

$$\text{Frequency} = \frac{k}{Rank}$$

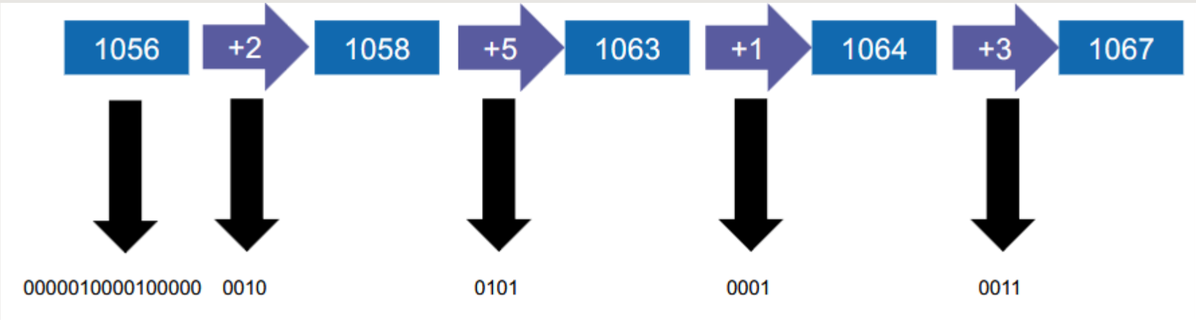
Compression of the dictionary

- Store pointers
- Words start with number indicating length



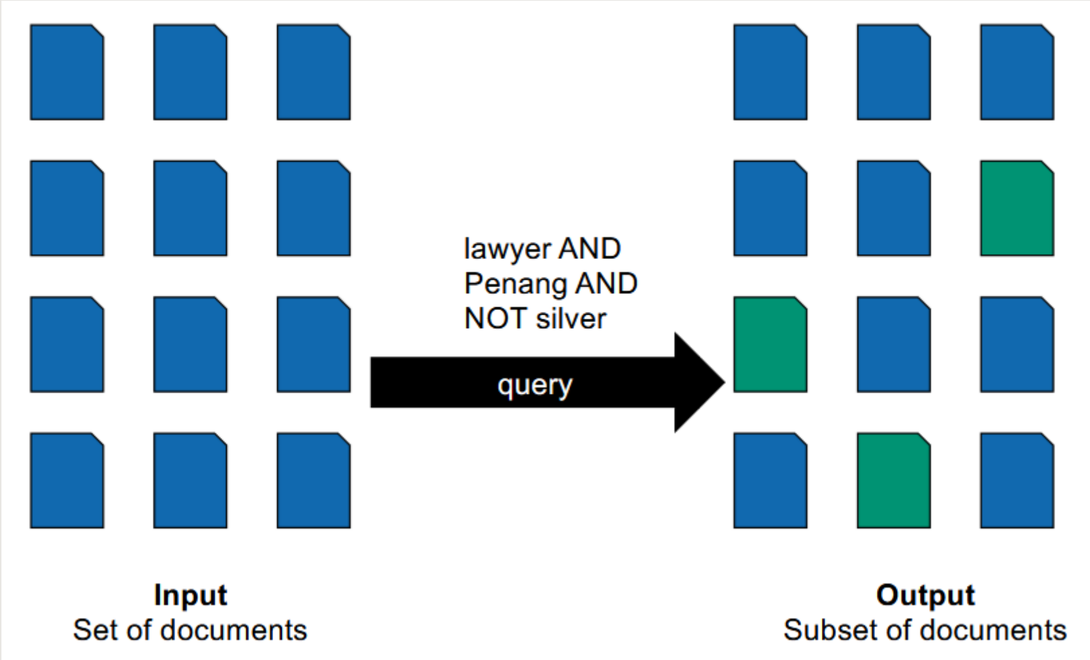
Compression of the postings list

- Encoding the gaps => What issue arises?



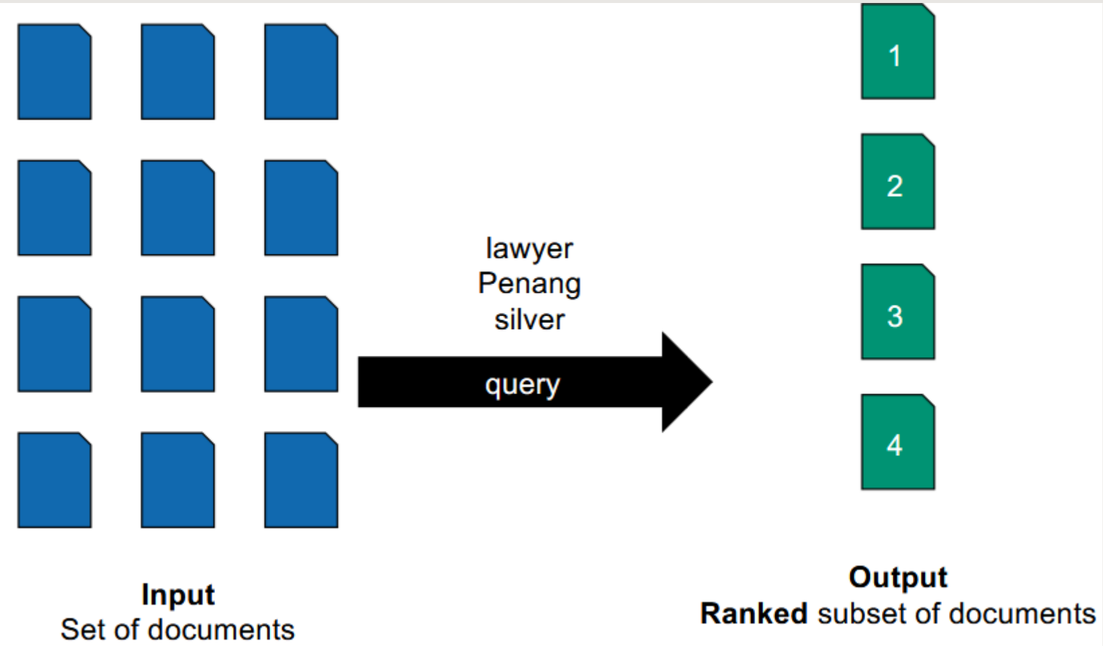
Boolean retrieval

- Until now:



Boolean retrieval

- What we want:



Parametric search

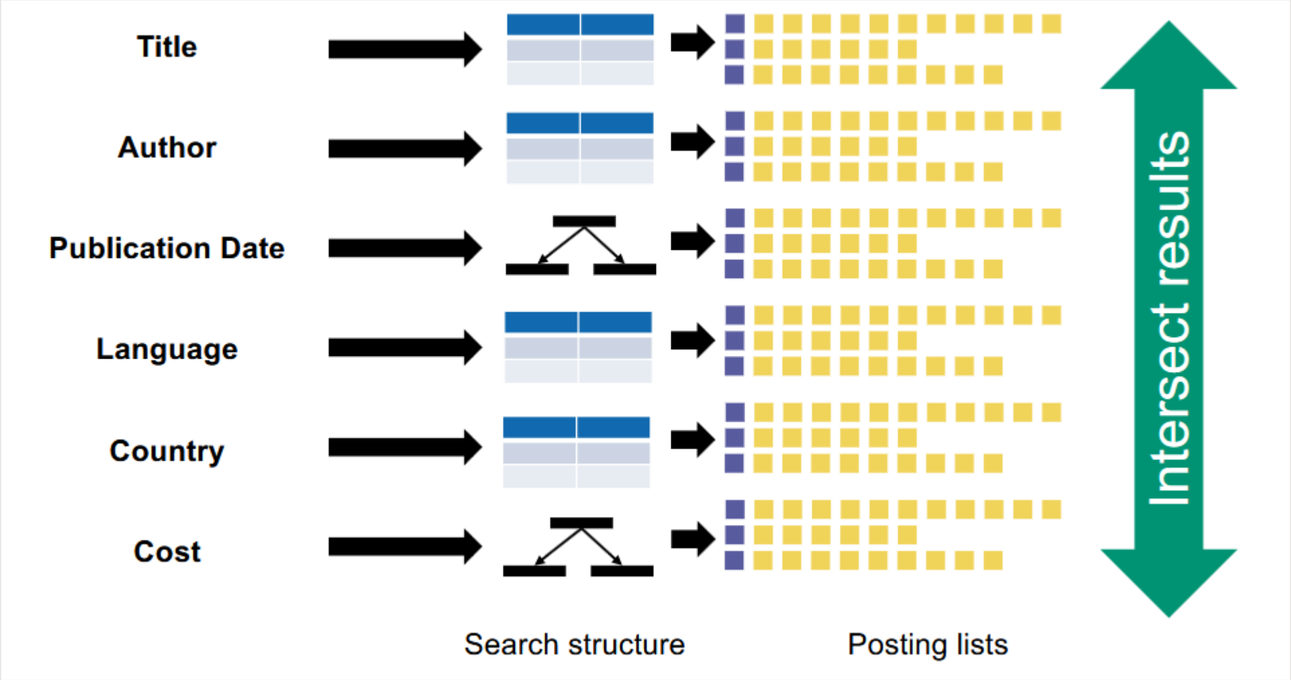
- Split index up between metadata

Title: Introduction to Information Retrieval
Authors: Christopher D. Manning and Prabhakar Raghavan and Hinrich Schütze
Hardcover: 506 pages
Publisher: Cambridge University Press; 1 edition (July 7, 2008)
Language: English
ISBN-10: 0521865719
ISBN-13: 978-0521865715
Product Dimensions: 7 x 1.2 x 10 inches
Shipping Weight: 2.2 pounds (View shipping rates and policies)
Average Customer Review: ★★★★½ 27 customer reviews

Parametric search

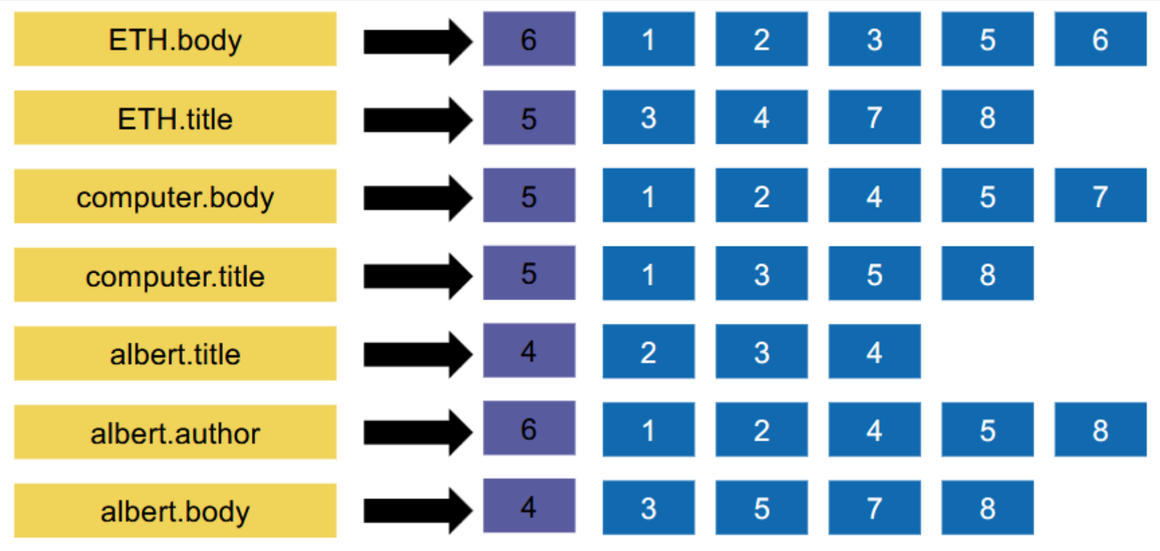
- One index for each metadata entry (i.e. author, title, etc.)
- Intersect results at the end

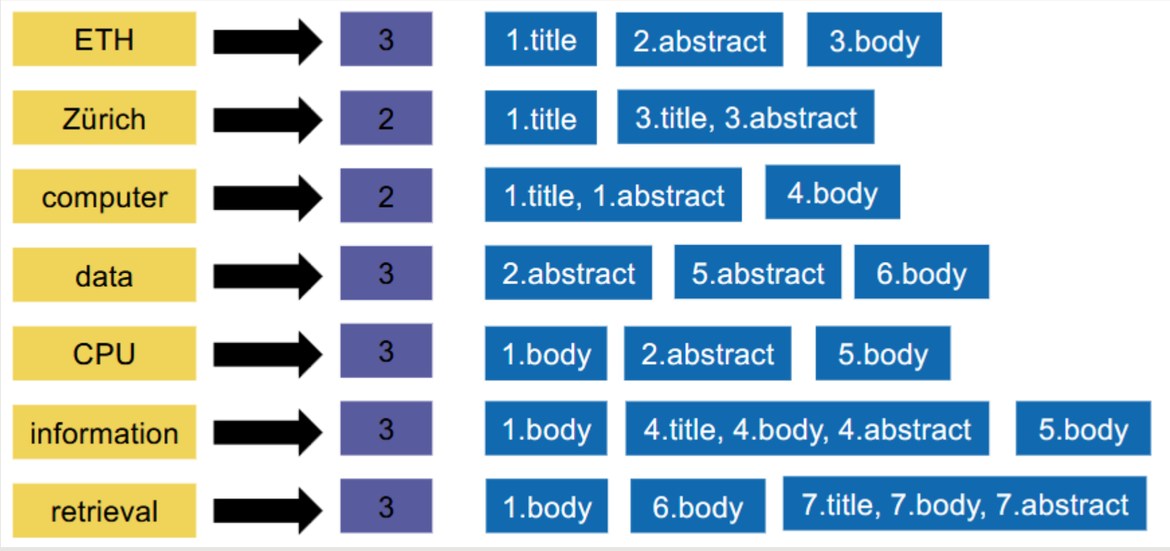
→ Zone search



Shared inverted index

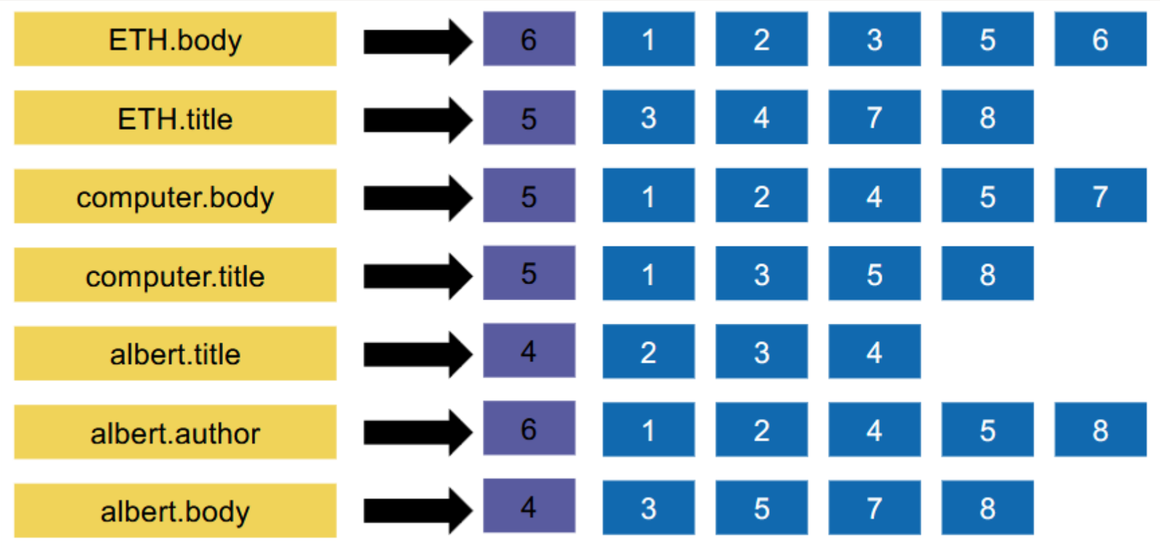
- Zones in terms vs. Zones in postings

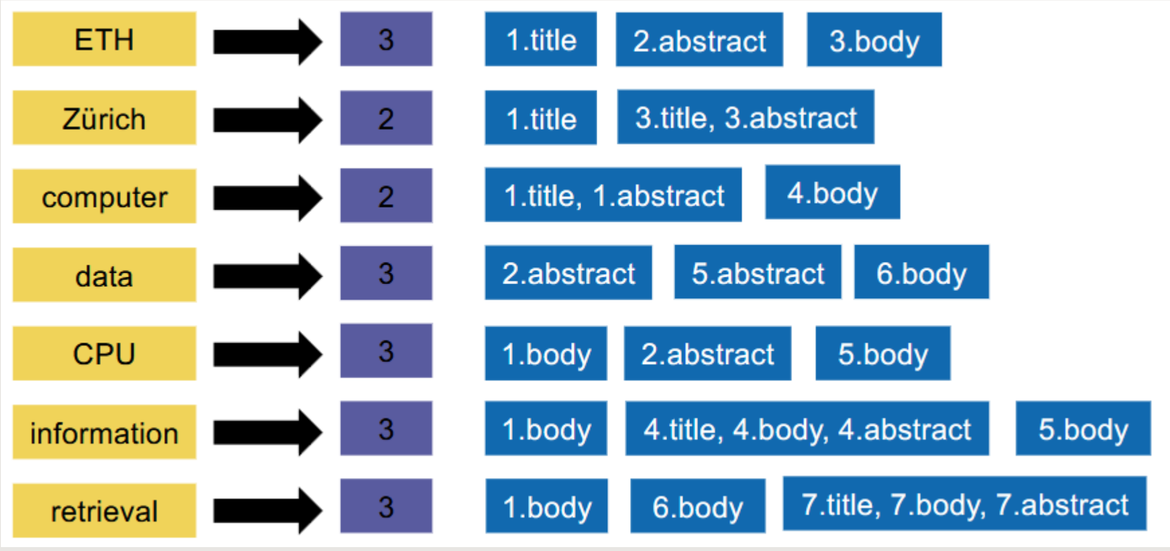




Shared inverted index

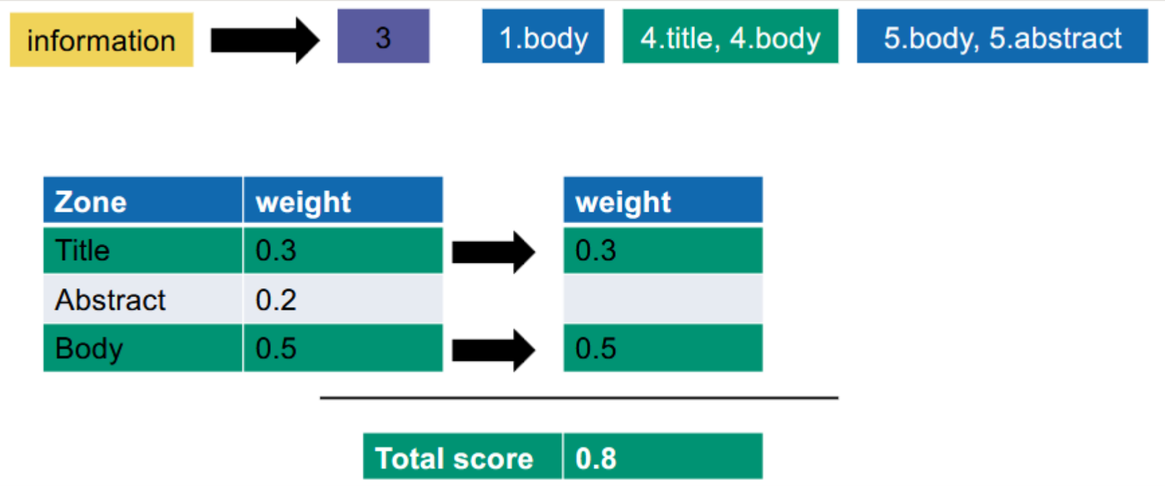
- Zones in terms vs. Zones in postings





Single-term query

- g_i: *zone weight*
- s_i: *zone i contains term or does not contain term*



Score of a document:

$$\sum_{i=1}^{i=l} g_i s_i$$

- Sort documents based on score, return top-k results

Term frequency

- Number of times a term occurs in a document

t	$tf_{t,d}$
foo	3
bar	2
foobar	2

Collection frequency

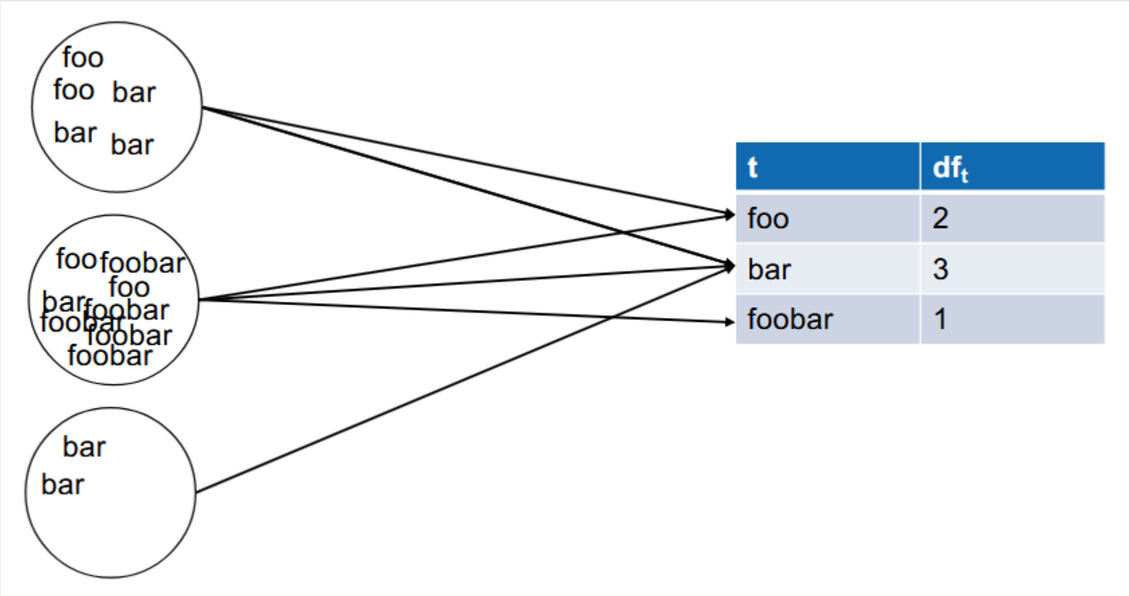
- Number of times a term occurs in the whole collection

t	cf_t
foo	4
bar	5
foobar	5

But foobar is rare, in that it appears only in one document!

Document frequency

- Number of documents where a term occurs
- Inverse document frequency: $idf_t = \ln \frac{N}{m}$, where $N := largest\ df_t, m := df_t$



t	df_t
foo	2
bar	3
foobar	1

t	df_t	idf_t
foo	2	0.41
bar	3	0
foobar	1	1.10

log 3/.

$tf - idf$

- $tf\text{-}idf = tf \times idf$

tf	A	B
foo	5	1
bar	0	4
foobar	2	1

tf-idf	A	B
foo	25	5
bar	0	40
foobar	6	3

idf	
foo	5
bar	10
foobar	3

Questions

- Moodle: Multiple choice
- Coding: Plot Heap’s and Zipf’s law

Kahoot

https://create.kahoot.it/details/information-retrieval-ex-06-repetitions/7ad1b588-2650-41bf-a05e-ee7ac16f8b77