https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018

Indian Institute of Information Technology, Sri City

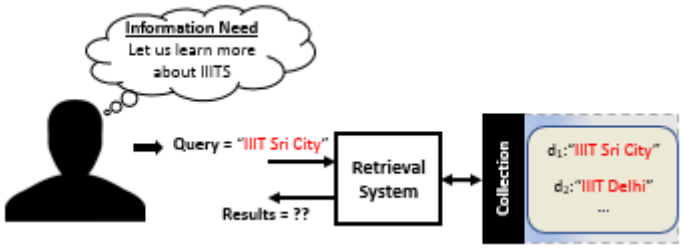
For me, data compression is more than a manipulation of numbers; it is the process of discovering structures that exist in the data**.**

– Khalid Sayood, University of Nebraska.

Review

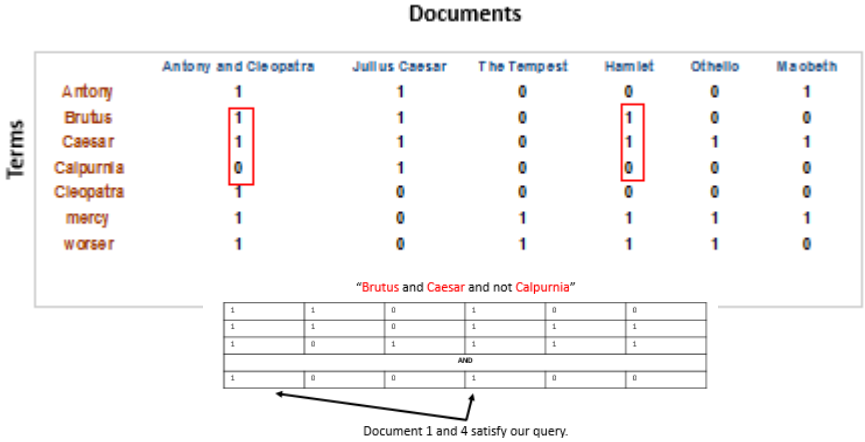
Dictionary Compression

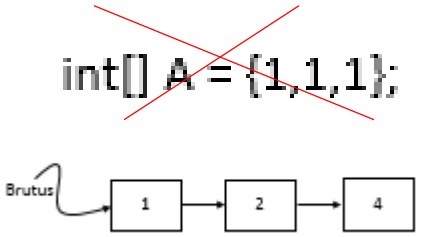
Introduction to Retrieval



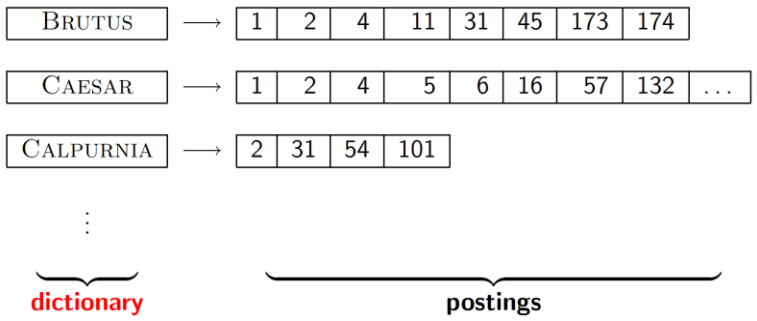
One (bad) Approach

- First match the **term** IIIT.
  - Filter out documents that contain this term.
- Next match the **term** Sri.
  - Filter out documents that contain this term.
- Next match the **term** City.
  - Filter out documents that contain this term.

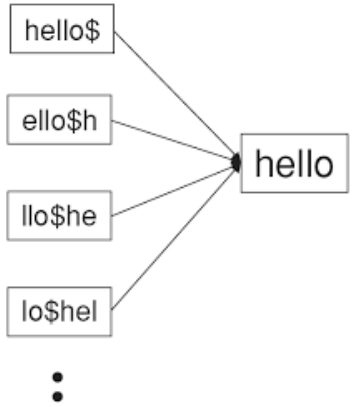




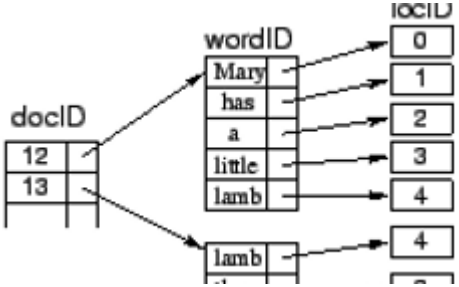
Indexing



Inverted Index



Permuterm Index



Forward Index

Term	Freq	Postings & Positions
and	1	(6,1) (6,6)
big	2	(2,3) (2,8) (3,8)
dark	1	(6,5)
did	1	(4,7)
gown	1	(2, 10)
had	1	(3,6)
house	2	(2,5) (3,2)
in	5	<(1,8)> <(2,1)> <(2,6)> <(3,3)> <(5,7)> <(6,3)> <(6, 8)>
keep	3	(1,7) (3,10) (5,6)

Positional Inverted Index

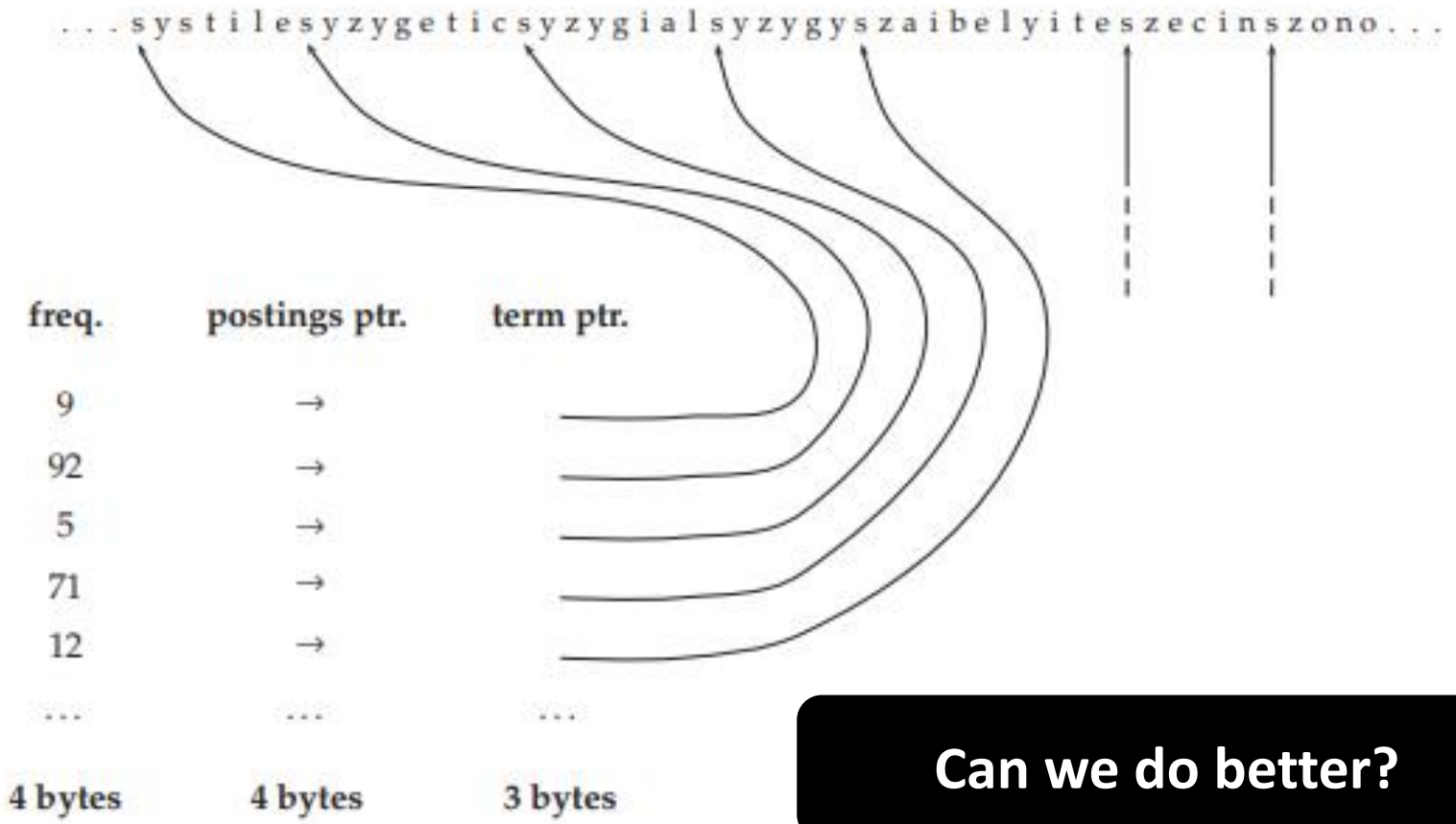
Dictionary as a Sorted Array

term	document frequency	pointer to postings list
a	656,265	⟶
aachen	65	⟶
…	…	…
zulu	221	⟶
20 bytes	4 bytes	4 bytes

Apply binary search to search the term array!

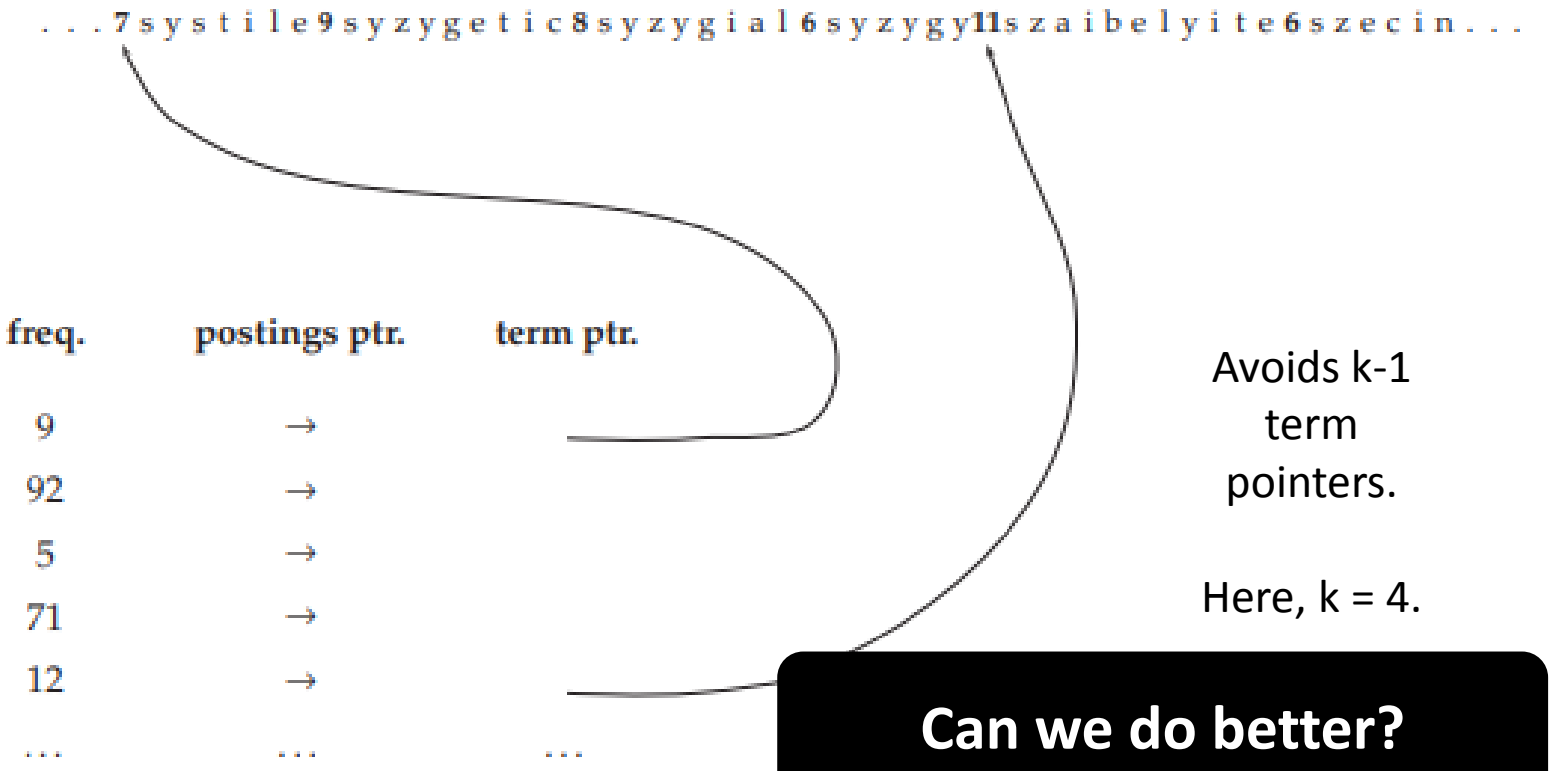
Can we do better?

Dictionary as a String



Can we do better?

Blocked Storage



Avoids k-1 term pointers.

Here, k = 4.

Can we do better?

Front Coding

One block in blocked compression ($k = 4$) . . .
8automata8automate9automatic10automation

. . . further compressed with front coding.
8automat∗a1⋄e2 ⋄ ic3⋄ion

Can you front-code at k = 3, "interspecies","interstellar","interstate"?

12inter*species7◊stellar5◊state

or

12inters*pecies6◊tellar4◊tate

Right Answer

- **12inters*pecies6◊tellar4◊tate**

Thank You.