https://vvtesh.sarahah.com/

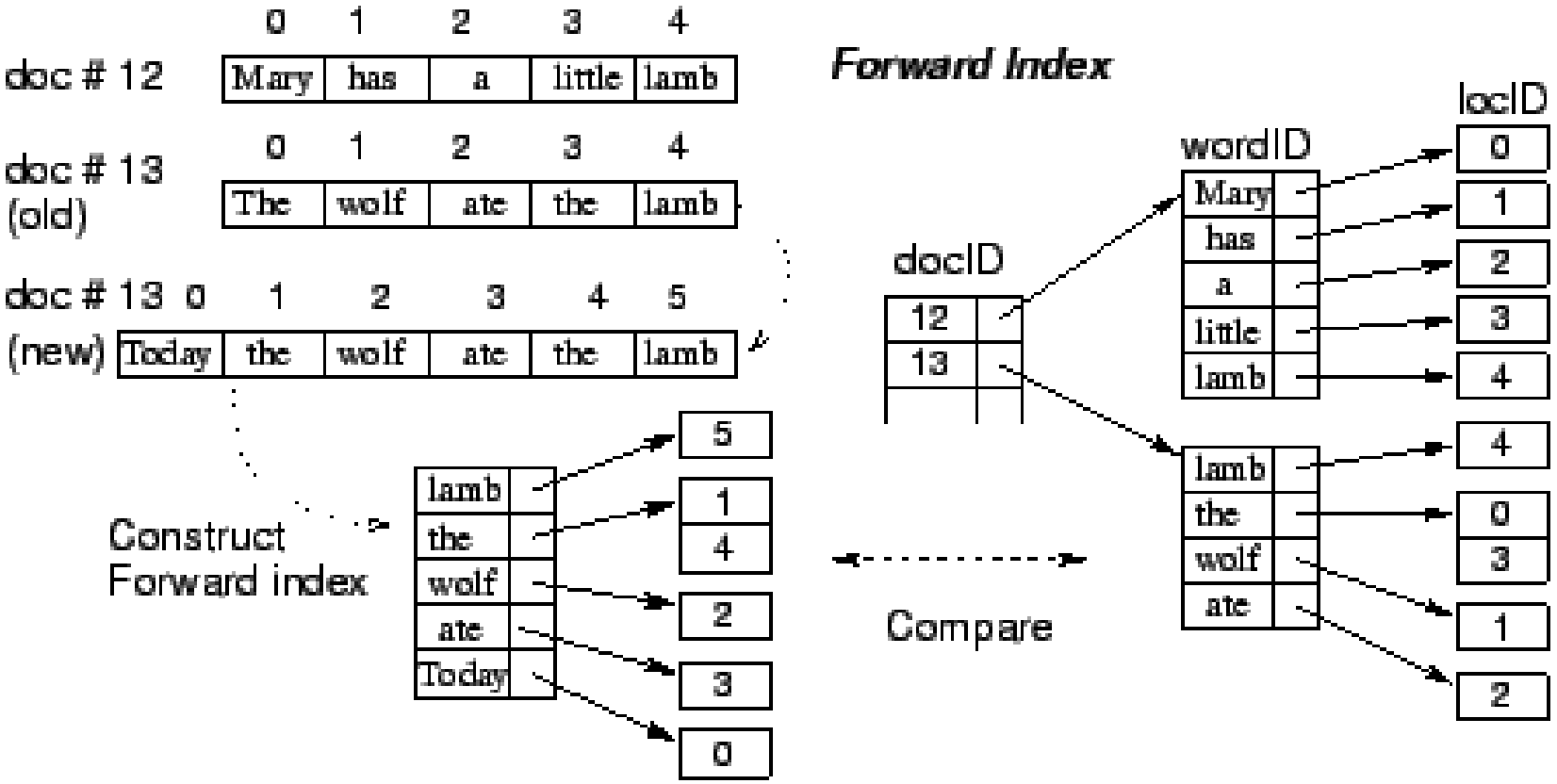
Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018
Indian Institute of Information Technology, Sri City

The human-computer interface is less well understood than other aspects of IR, in part because humans are more complex than computer systems and their motivations and behaviors are more difficult to measure and characterize.
– Marti A. Hearst.

Forward Index



Inverted Index

Doc 1
I did enact Julius Caesar: I was killed
i' the Capitol; Brutus killed me.

Doc 2
So let it be with Caesar. The noble B
hath told you Caesar was ambitious:

term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1

term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1

Dictionary

term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2

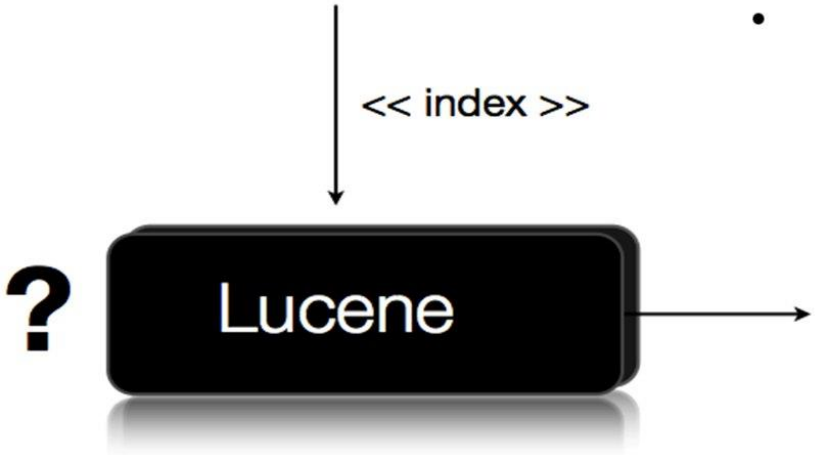
Positional Index

- Query: *"to$_1$ be$_2$ or$_3$ not$_4$ to$_5$ be$_6$"*
  - TO, 993427:
    - ‹ 1: ‹7, 18, 33, 72, 86, 231›;
    - 2: ‹1, 17, 74, 222, 255›;
    - 4: ‹8, 16, 190, 429, 433›;
    - 5: ‹363, 367›;
    - 7: ‹13, 23, 191›; . . . ›
  - BE, 178239:
    - ‹ 1: ‹17, 25›;
    - 4: ‹17, 191, 291, 430, 434›;
    - 5: ‹14, 19, 101›; . . . ›
- Document 4 is a match!

Inverted Index

1	The old night keeper keeps the keep in the town
2	In the big old house in the big old gown.
3	The house in the town had the big old keep
4	Where the old night keeper never did sleep.
5	The night keeper keeps the keep in the night
6	And keeps in the dark and sleeps in the light.

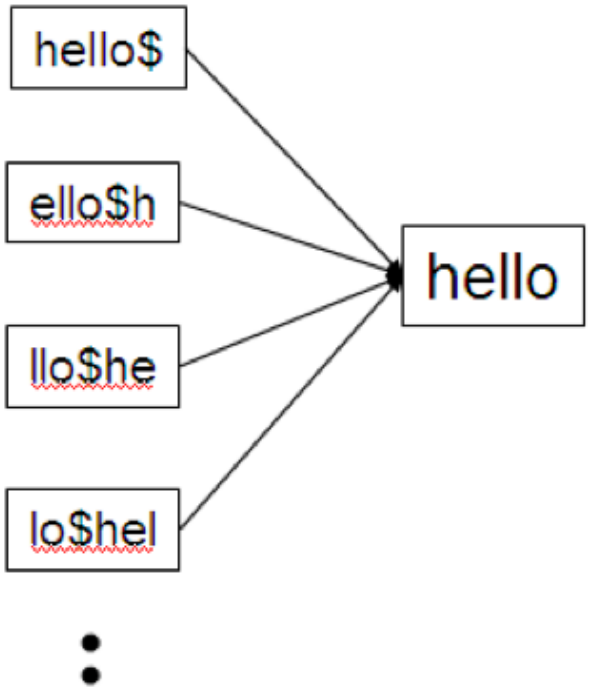
Table with 6 documents



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)
keeper	3	(1, 4) (4, 5) (5, 3)
keeps	3	(1, 5) (5, 4) (6, 2)
light	1	(6, 10)
never	1	(4, 6)
night	3	(1, 3) (4, 4) (5, 2) (5, 9)
old	4	(1, 2) (2, 4) (2,9) (3, 9) (4, 3)
sleep	1	(4, 8)
sleeps	1	(6, 7)
the	6	<(1, 1)> <(1,6)> <(2, 2)> <(2,7)> <(3, 1)> <(3,4)> <(3,7)> <(4, 2)> <(5, 1)> <(5, 5)> <(5, 8)> <(6, 4)> <(6, 9)>
town	2	(1, 10) (3, 5)
where	1	(4,1)

Permuterm index

- For HELLO, we've stored: *hello$, ello$h, llohe, lohel, and o$hell*



6

K-gram Index

- Enumerate all character *k*-grams (sequence of *k* characters) occurring in a term
- 2-grams are called bigrams.
- Example: from April is the cruelest month we get the bigrams: $a ap pr ri il l$ $i is s$ $t th he e$ $c cr ru ue el le es st t$ $m mo on nt h$
- $ is a special word boundary symbol, as before.

A partial snapshot of a sample 3-gram index

Quiz

- How does the bi-gram index look like?
  - Assume:
    - Collection: d1: INDIA, d2: ASIA

Thank You