https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

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

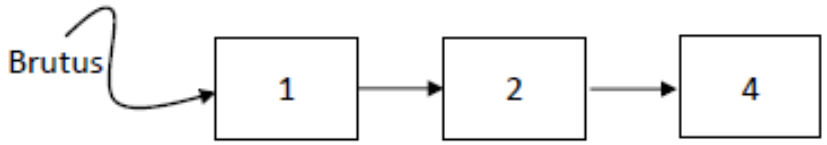
Text is the primary way that human knowledge is stored, and after speech, the primary way it is transmitted**.**

– Bill Frakes and Ricardo Baeza Yates

Term Document Matrix & Postings List

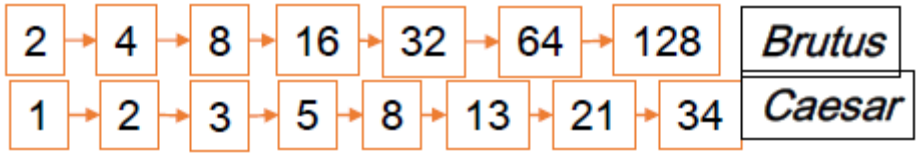
Documents

Terms	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0



Query Processing

- Brutus AND Caeser



- Which document(s) should result?
- How many comparisons did you do?

Query Processing

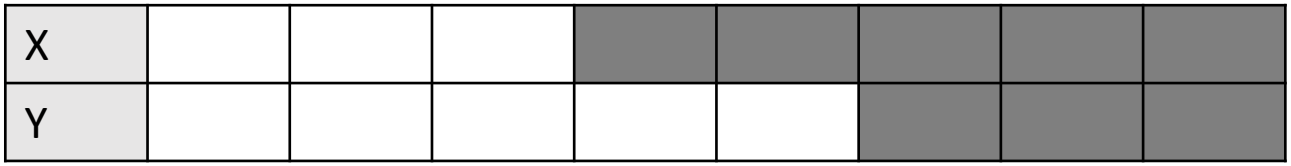
- X AND Y

X	1	2	3	4	5	6	7	8
Y	1	2	3					

- Which document(s) should result?
- How many comparisons did you do?
- Is there any way we could get:
  - |result| > min(|x|, |y|)
    - No!

Query Processing

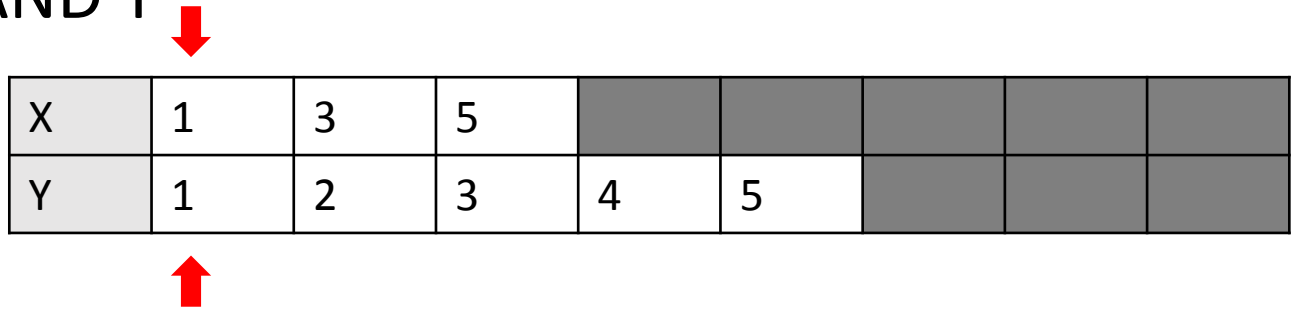
- X AND Y



- If |X| = 3, |Y| = 5, Can you fill the boxes in two different ways such that the number of comparisons are different?

Query Processing

- X AND Y

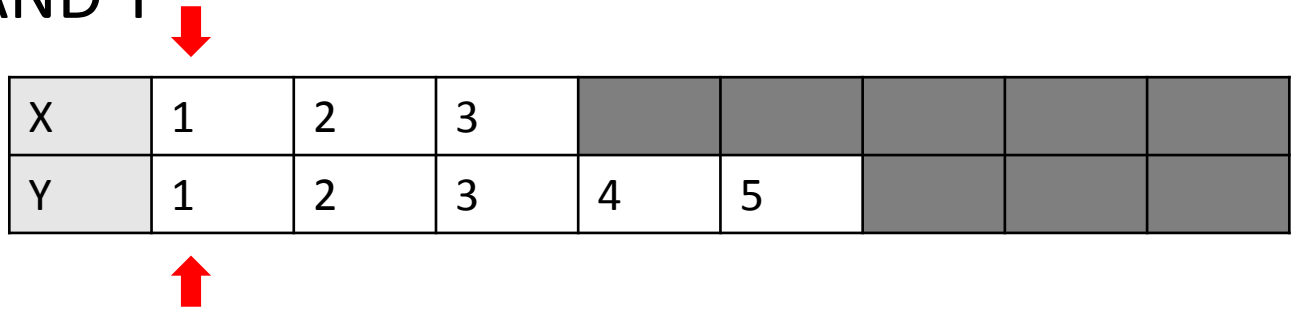


X	1	3	5					
Y	1	2	3	4	5			

- No. of Comparisons = |(1,1), (3,2), (3,3), (5,4), (5,5)| = 5.

Query Processing

- X AND Y

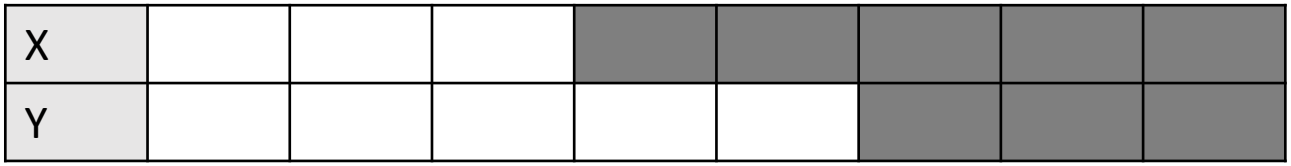


X	1	2	3					
Y	1	2	3	4	5			

- No. of Comparisons = |(1,1), (2,2), (3,3)| = 3.

Query Processing

- X AND Y



- If |X| = 3, |Y| = 5, Can you fill the boxes in two different ways such that the number of comparisons are ~~different~~ maximum?

Query Processing

- X AND Y

X	1	2	8					
Y	4	5	6	7	8			

- No. of Comparisons = |(1,4),(2,4),(8,4),(8,5),(8,6),(8,7),(8,8)| = 7

- Can you do any better?

Query Processing

- X AND Y
- Min. No. of Comparisons =3

X	1	2	3					
Y	1	2	3	4	5			

- Max. No. of Comparisons = 7

X	1	2	8					
Y	4	5	6	7	8			

- Is there a better answer?

Query Processing

- Query: Brutus AND Caeser AND Calpurnia
- Assumption:
  - Brutus appears in 10 documents.
  - Caeser appears in 5 documents.
  - Calpurnia appears in 6 documents.
- How many comparisons?
  - Option 1: Merge Brutus AND Caeser first. Merge the result with Calpurnia.
  - Option 2: Merge Caeser AND Calpurnia first. Merge the result with Brutus.

Query Processing Order

- Option 1: Merge Brutus AND Caeser first. Merge the result with Calpurnia.
  - Brutus AND Caeser: In worst case, requires (10 + 6) = 16 comparisons.
  - Result AND Calpurnia: In worst case, requires (6 + 5) = 11 comparisons.
  - Therefore, requires 16 + 11 = 27 comparisons.
- Option 2: Merge Caeser AND Calpurnia first. Merge the result with Brutus.
  - Caeser AND Calpurnia: In worst case, requires 6 + 5 = 11 comparisons.
  - Result AND Brutus: In worst case, requires 5 + 10 = 15 comparisons.
  - Therefore, 15 + 11 = 26 comparisons.

*We approximate worst case comparisons to x+y for convenience.

Process in increasing order by frequency.

Do you now see why we store frequency with our Dictionary terms?

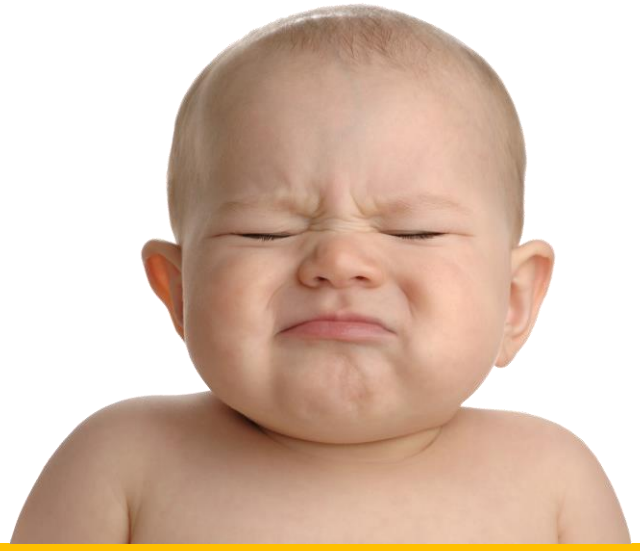
Quiz

What is the best order of processing "eyes and skies and trees"?

Term	Postings size
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

What about "(eyes or skies) and (trees or tangerine) and (marmalade or kaleidoscope)"?

Find the query processing order for
(A OR B) AND (C OR D) AND (E OR F)



(A OR B) AND (C OR D) AND (E OR F)

- Let x = Freq(A) + Freq(B)
- Let y = Freq(C) + Freq(D)
- Let z = Freq(E) + Freq(F)

- A OR B leaves us with A union B items. In worst case, we have freq(A) + freq(B) items.

- We know how to solve (x AND y AND z)

Quiz

Term	Postings size
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

What about (tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes)**?**

Answer: ((kaleidoscope OR eyes) AND (tangerine OR trees)) AND (marmalade or skies)

Thank You

*It is not **how much** we know, it is **how well** we know that matters.*