https://vvtesh.sarahah.com/

Information Retrieval

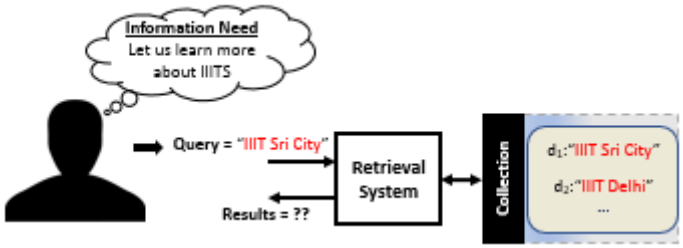
Venkatesh Vinayakarao

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

The study of mathematics, like the Nile, begins in minuteness but ends in magnificence.
Charles Caleb Colton.

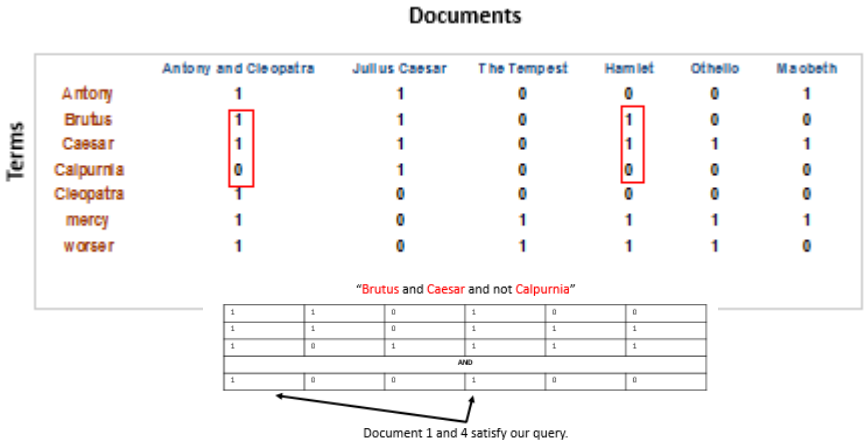
The endless study of information retrieval, like the Taj Mahal, begins in magnificence and stays in magnificence.
Venkatesh Vinayakarao.

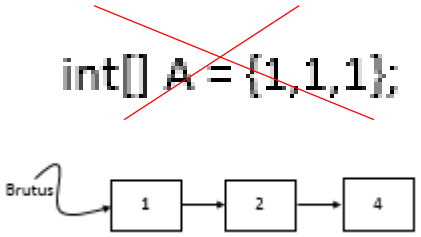
Review



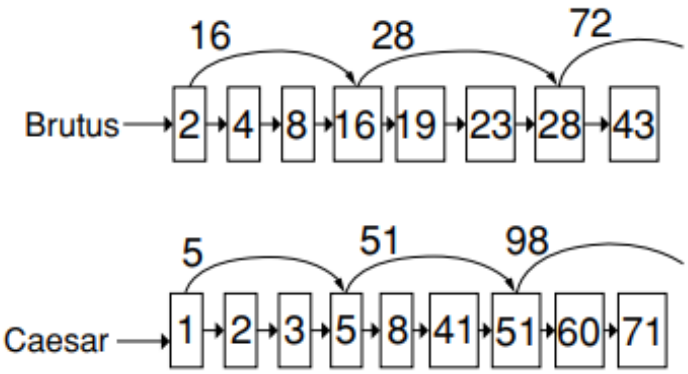
One (bad) Approach

- 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.





Review



Skip Pointers

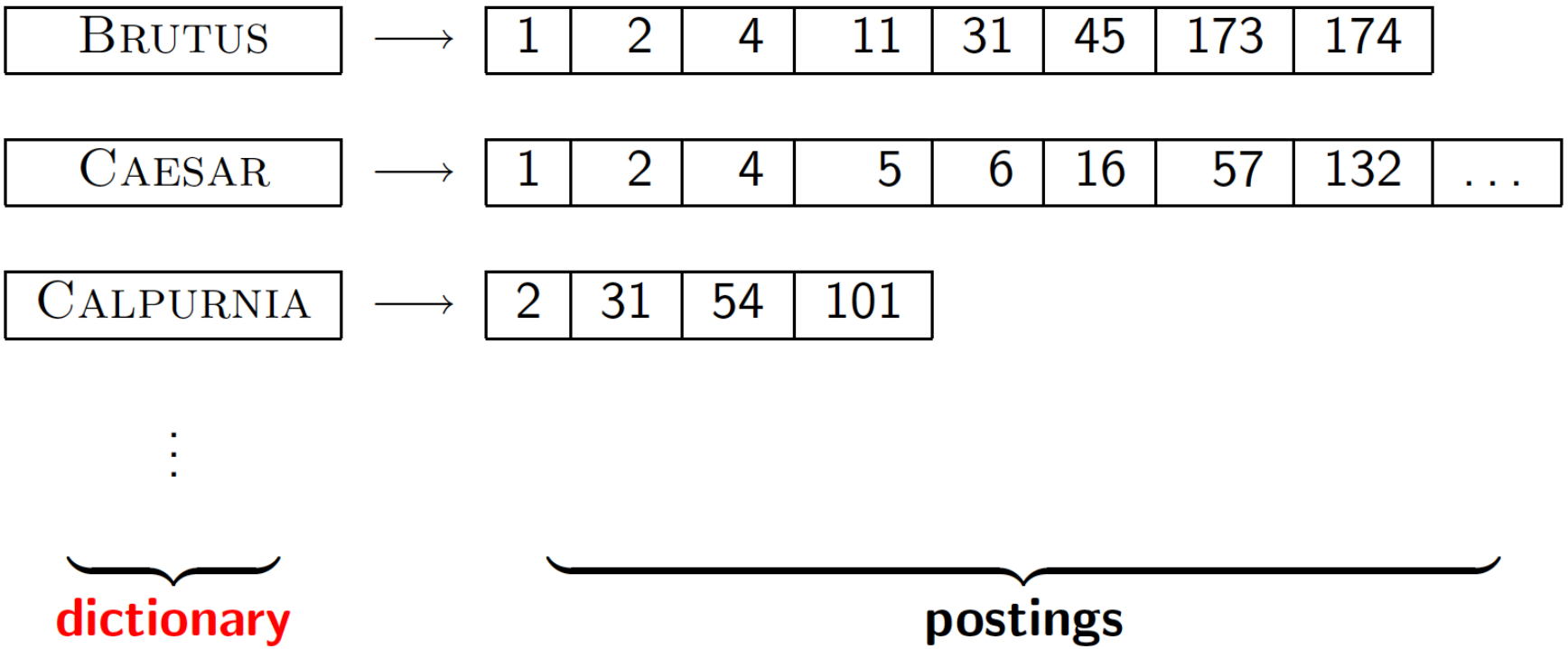
Content Processing

$$\text{Precision } P = tp/(tp + fp)$$

$$\text{Recall } \quad R = tp/(tp + fn)$$

Evaluation

How to Store a Dictionary?



One (bad) Approach

- Store them all in a file.
- Go linearly (one by one) and compare.

Avg. no. of Comparisons $\propto$ No. of Words in Dictionary

Second (still bad) Approach

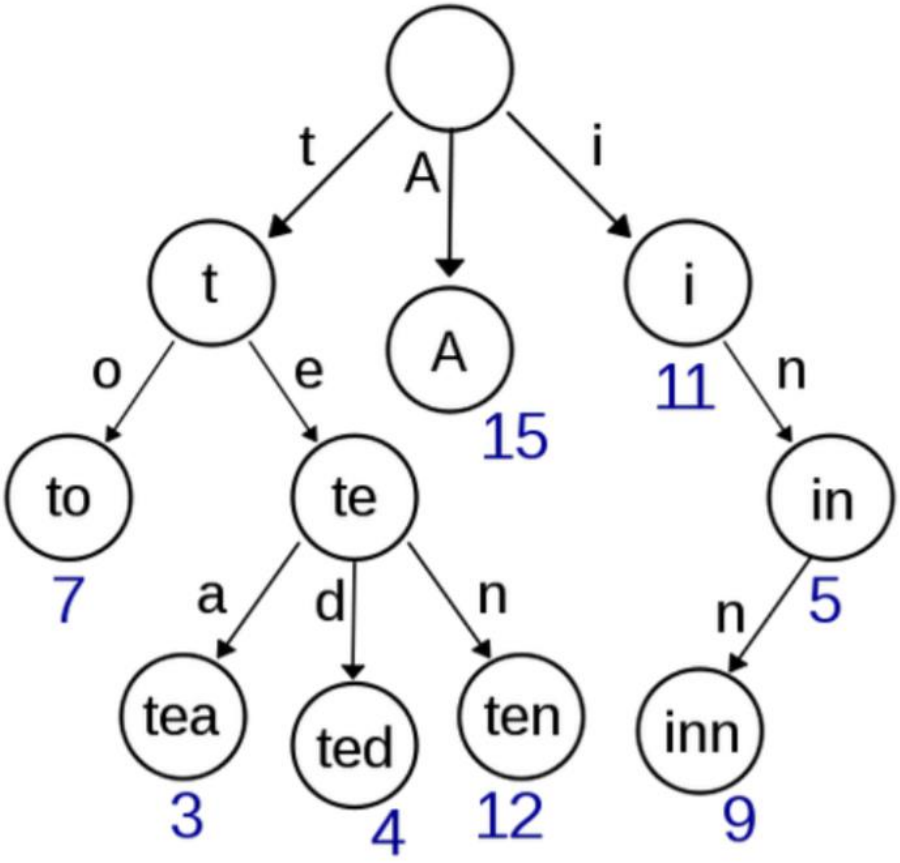
- Sort them.
- Store them in a file.
- Do a binary search.

Avg. no. of Comparisons $\propto$ Log(No. of Words in Dictionary)

Can we do better?

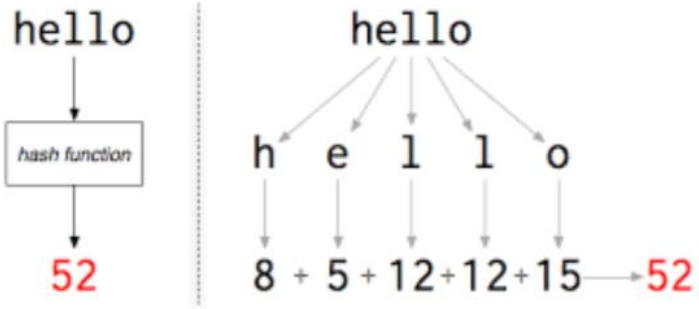
How can we store all dictionary words for a fast look up?

B Trees and Tries

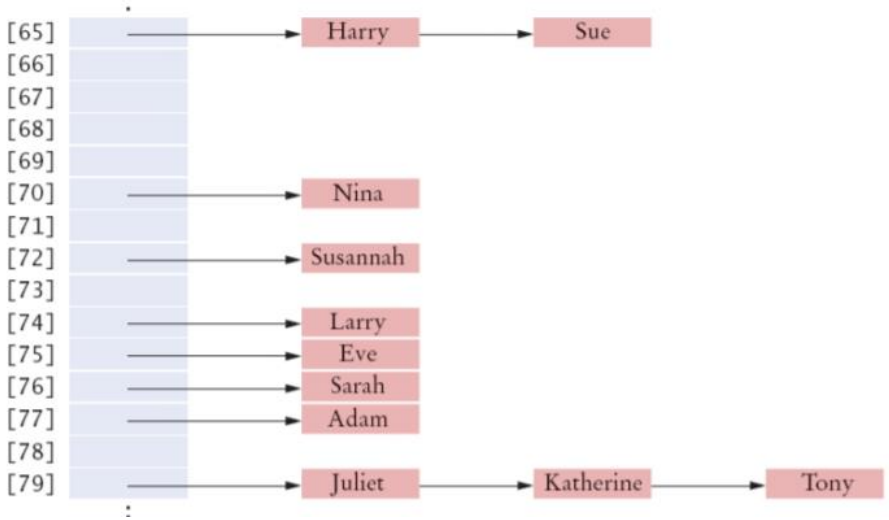


Hashing

Simple hash function



Simple hash table



More String Handling Problems

How can you find if a given string S is a substring of another string T?

Example
S = “ego”, T = “category”

How can you find the number of times S occurs in T?

Example
“a” appears thrice in Banana

Is S a suffix of T?

Example
“dia” is a suffix of India

Find the longest repeating substring of T

Example
“geeks” is the longest repeating substring in T = “geeksforgeeks”

Given two strings X and Y, find the longest common substring of X and Y.

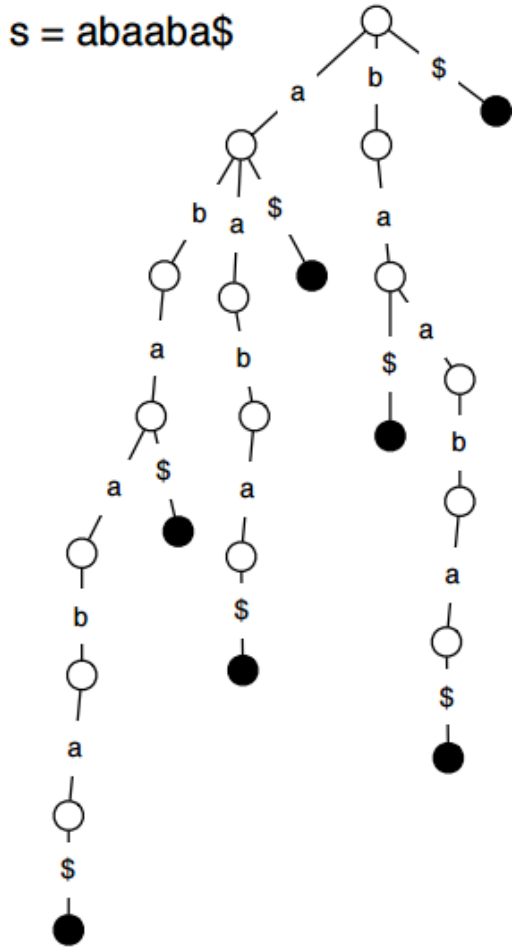
Example

X = “geeksforgeeks”, Y = “geeksquiz”. Longest common substring is “geeks”

Flex your brain!

- How can you find if a given string S is a substring of another string T?
  - S = “ego”, T = “category”
- How can you find the number of times S occurs in T?
  - S = “a”, T = “Banana”
- Is S a suffix of T?
  - S = “dia”, T = “India”
- Find the longest repeating substring of T.
  - T = “geeksforgeeks”
- Given two strings X and Y, find the longest common substring of X and Y.
  - X = “geeksforgeeks”, Y = “geeksquiz”

Suffix Tree



Flex your brain!

- How can you find if a given string S is a substring of another string T?
- How can you find the number of times S occurs in T?
- Is S a suffix of T?
- Find the longest repeating substring of T.
- Given two strings X and Y, find the longest common substring of X and Y.

Flex your brain!

- Draw suffix trees for
  - banana
  - ssnsace
  - elephant

Wild-card queries: *

- ***mon*:*** retrieve all words in range: ***mon ≤ w < moo***
- ****mon:*** find words ending in "mon": harder
  - Maintain an additional B-tree for terms *backwards.*

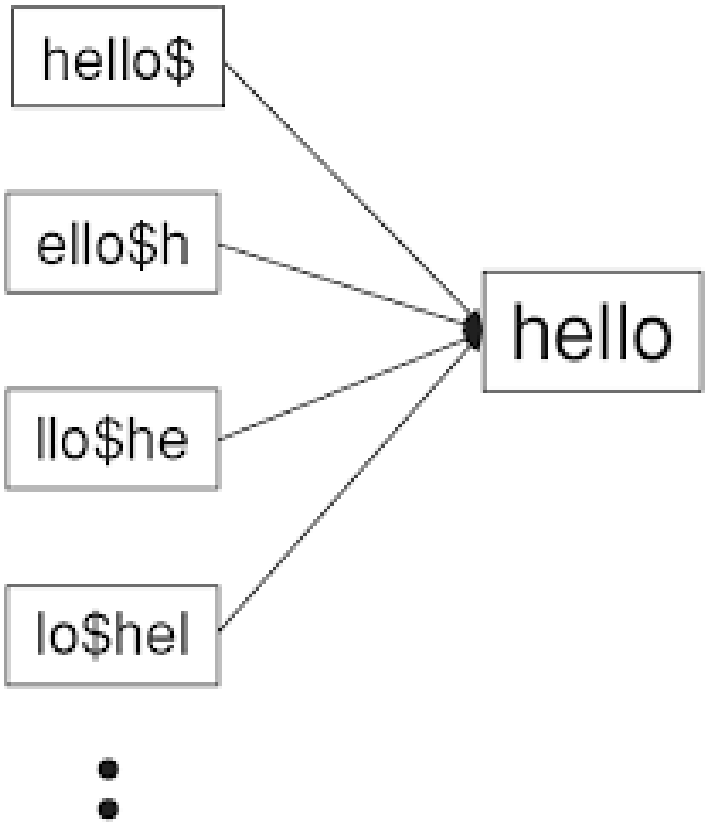
  Can retrieve all words in range: ***nom ≤ w < non****.*

B-trees handle *'s at the end of a query term

- How can we handle *'s in the middle of query term?
  - ***co*tion***
- We could look up ***co**** AND ****tion*** in a B-tree and intersect the two term sets
  - Expensive
- The solution: transform wild-card queries so that the *'s occur at the end
- This gives rise to the **Permuterm** Index.

How to Match he*lo?

- Rotate he*lo → he*lo$ → $he*lo → o$he*l → lo$he*
- Till * is at the end.
- Exercise: How will you match h*l*o?



Questions?