https://vvtesh.sarahah.com/

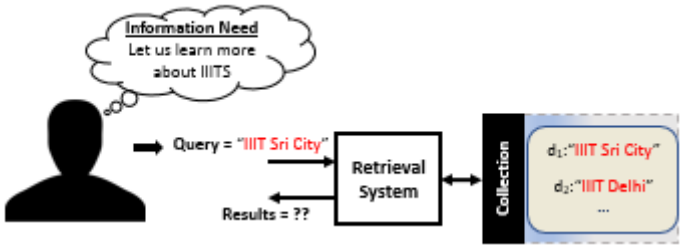
# Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Sep, 2019
Chennai Mathematical Institute

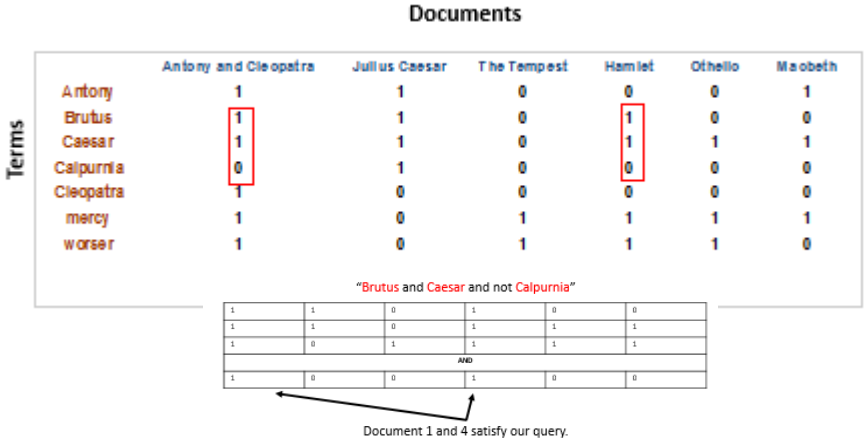
**Mission defines strategy, and strategy defines structure.**
**– Peter Drucker.**

# Review – Structures for Search

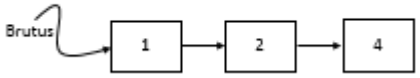



## One (bad) Approach

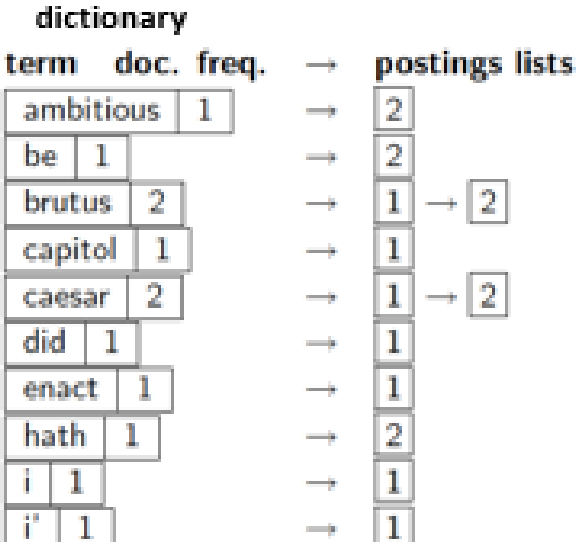
- First match the **term** IIIT.
  - Filter out documents that contain this term.
- Next match the **term** Sri.
  - Filter out documents that contain this term.
- Next match the **term** City.
  - Filter out documents that contain this term.



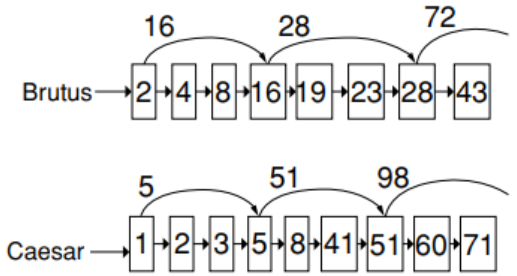

~~int[] A = {1,1,1};~~



# Review - Indexing




**Dictionary and Postings List**



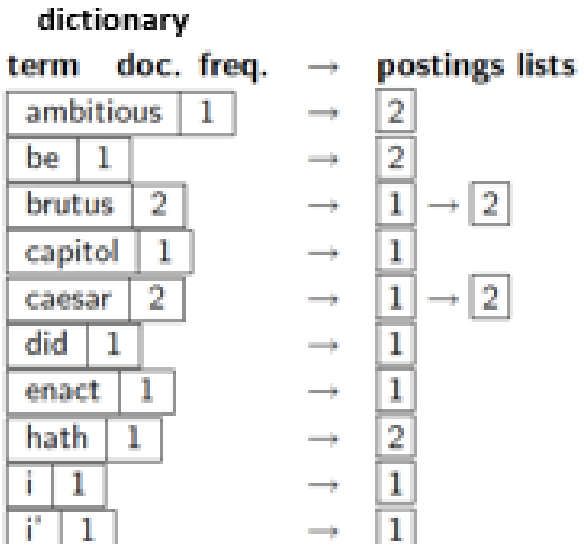

**Skip Pointers**

to, 993427:
⟨ 1, 6: ⟨7, 18, 33, 72, 86, 231⟩;
2, 5: ⟨1, 17, 74, 222, 255⟩;
4, 5: ⟨8, 16, 190, 429, 433⟩;
5, 2: ⟨363, 367⟩;
7, 3: ⟨13, 23, 191⟩; . . . ⟩

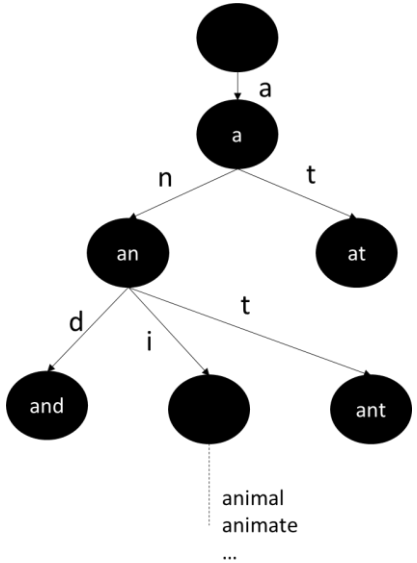
be, 178239:
⟨ 1, 2: ⟨17, 25⟩;
4, 5: ⟨17, 191, 291, 430, 434⟩;
5, 3: ⟨14, 19, 101⟩; . . . ⟩

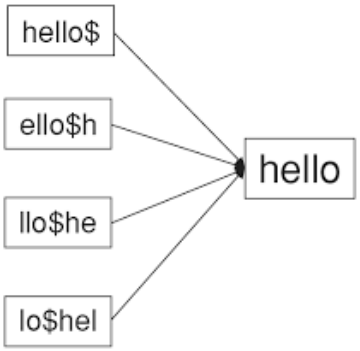
**Postings List for Positional Index**

# Review – Storing Dictionaries




**Dictionary and Postings List**




**Part of :**
**Permuterm Index**

# Review


**Content Processing**

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

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

**Evaluation**

# Agenda

- String Handling Problems
- Spelling Correction

# Some 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.

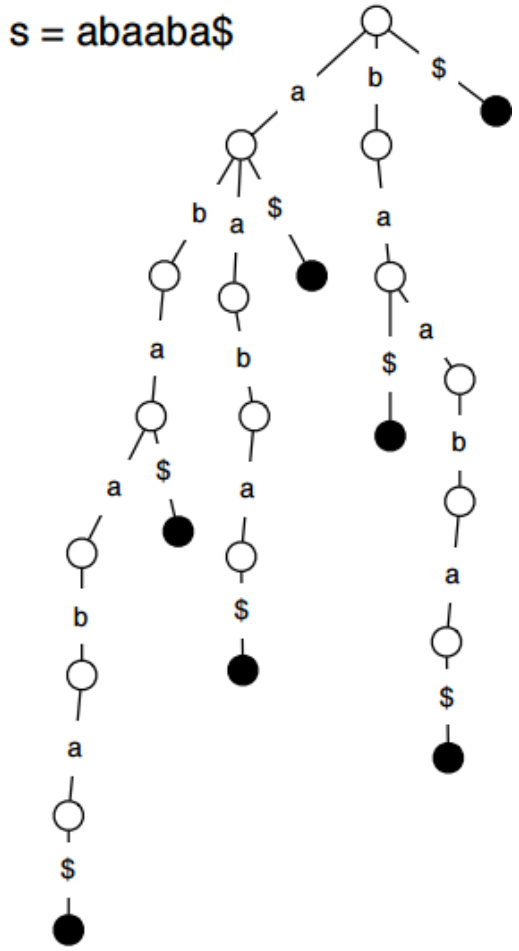
Example

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.

# Quiz: Flex your brain!

- Draw suffix tree for
  - banana
- Highlight and explain how to find the longest repeating substring of banana.

# Spelling Correction

# Spelling Errors in Query

- 10% to 15% of queries carry misspelt words.
- English is not 100% phonetic.
- How many of these phrases contain spelling errors?
  - cigarete lighter
  - fourty dollars
  - going to libary today
  - unforgetable holiday
  - successful businessman

# Notorious Britney

The data below shows some of the misspellings detected by our spelling correction system for the query [ britney spears ], and the count of how many different users spelled her name that way. -- Google.

| | | | | | |
|---|---|---|---|---|---|
| 488941 britney spears | 29 britent spears | 9 brinttany spears | 5 brney spears | 3 britiy spears | 2 brirreny spears |
| 40134 brittany spears | 29 brittnany spears | 9 britanay spears | 5 broitney spears | 3 britmeny spears | 2 brirtany spears |
| 36315 brittney spears | 29 britttany spears | 9 britinany spears | 5 brotny spears | 3 britneeey spears | 2 brirttany spears |
| 24342 britany spears | 29 btiney spears | 9 britn spears | 5 bruteny spears | 3 britnehy spears | 2 brirttney spears |
| 7331 britny spears | 26 birttney spears | 9 britnew spears | 5 btiyney spears | 3 britnely spears | 2 britain spears |
| 6633 briteny spears | 26 breitney spears | 9 britneyn spears | 5 btrittney spears | 3 britnesy spears | 2 britane spears |
| 2696 britteny spears | 26 brinity spears | 9 britrney spears | 5 gritney spears | 3 britnetty spears | 2 britaneny spears |
| 1807 briney spears | 26 britenay spears | 9 brtiny spears | 5 spritney spears | 3 britnex spears | 2 britania spears |
| 1635 brittny spears | 26 britneyt spears | 9 brtittney spears | 4 bittny spears | 3 britneyxxx spears | 2 britann spears |
| 1479 brintey spears | 26 brittan spears | 9 brtny spears | 4 bnritney spears | 3 britnity spears | 2 britanna spears |
| 1479 britanny spears | 26 brittne spears | 9 brytny spears | 4 brandy spears | 3 britntey spears | 2 britannie spears |
| 1338 britiny spears | 26 btittany spears | 9 rbitney spears | 4 brbritney spears | 3 britnyey spears | 2 britannt spears |
| 1211 britnet spears | 24 beitney spears | 8 birtiny spears | 4 breatiny spears | 3 britterny spears | 2 britannu spears |
| 1096 britiney spears | 24 birteny spears | 8 bithney spears | 4 breetney spears | 3 brittneey spears | 2 britanyl spears |
| 991 britaney spears | 24 brightney spears | 8 brattany spears | 4 bretiney spears | 3 brittnney spears | 2 britanyt spears |
| 991 britnay spears | 24 brintiny spears | 8 breitny spears | 4 brfitney spears | 3 brittnyey spears | 2 briteeny spears |
| 811 brithney spears | 24 britanty spears | 8 breteny spears | 4 briattany spears | 3 brityen spears | 2 britenany spears |
| 811 brtiney spears | 24 britenny spears | 8 brightny spears | 4 brieteny spears | 3 briytney spears | 2 britenet spears |
| 664 birtney spears | 24 britini spears | 8 brintay spears | 4 briety spears | 3 brltney spears | 2 briteniy spears |
| 664 brintney spears | 24 britnwy spears | 8 brinttey spears | 4 briitny spears | 3 broteny spears | 2 britenys spears |
| 664 briteney spears | 24 brittni spears | 8 briotney spears | 4 briittany spears | 3 brtaney spears | 2 britianey spears |
| 601 bitney spears | 24 brittnie spears | 8 britanys spears | 4 brinie spears | 3 brtiiany spears | 2 britin spears |
| 601 brinty spears | 21 biritney spears | 8 britley spears | 4 brinteney spears | 3 brtinay spears | 2 britinary spears |
| 544 brittaney spears | 21 birtany spears | 8 britneyb spears | 4 brintne spears | 3 brtinney spears | 2 britmy spears |
| 544 brittnay spears | 21 biteny spears | 8 britnrey spears | 4 britaby spears | 3 brtitany spears | 2 britnaney spears |

Source: https://archive.google.com/jobs/britney.html

# Two Major Approaches

- Two major approaches exist for spelling correction:
  - finding "**nearest**" dictionary term.
  - finding "**most commonly used**" dictionary term when there are multiple "nearest terms".
- Two major kinds:
  - Isolated-Term Correction
    - Correct one word at a time.
  - Context-Sensitive Correction
    - "flew ***form*** New York" – Note that form is a dictionary term. Yet, this requires to be corrected to "flew from New York".

# Spelling Correction: Edit distance

- Given two strings $S_1$ and $S_2$, the minimum number of operations to convert one to the other
- Operations are typically character-level
  - Insert, Delete, Replace, (and perhaps Transposition*)
- E.g., the edit distance from ***dof*** to ***dog*** is 1
  - From ***cat*** to ***act*** is 2 (Just 1 with transpose.)
  - from ***cat*** to ***dog*** is 3.

*In this course, we do not consider transposition.

# Quiz

What is the edit distance between Sunday and Saturday?

*You are allowed to perform only Insert, Delete, and Replace operations.

# Answer

- Saturday = Sunday = S*day
- Problem is same as
  - What is the edit distance between atur and un?
  - Answer
    - Delete a,t. Replace r with n.
    - 3.

# Levenshtein Example

| | | S | a | t | u | r | d | a | y |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| **s** | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| **u** | 2 | 1 | 1 | 2 | 2 | 3 | 4 | 5 | 6 |
| **n** | 3 | 2 | 2 | 2 | 3 | 3 | 4 | 5 | 6 |
| **d** | 4 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 5 |
| **a** | 5 | 4 | 3 | 4 | 4 | 4 | 4 | 3 | 4 |
| **y** | 6 | 5 | 4 | 4 | 5 | 5 | 5 | 4 | 3 |

**Sunday**

**Keep s. Insert a, t.**

**Keep u.**

**Replace r.**

**Keep day.**

**Saturday**

# Levenshtein Algorithm

```
EDITDISTANCE(s1, s2)
 1  int m[i, j] = 0
 2  for i ← 1 to |s1|
 3  do m[i, 0] = i
 4  for j ← 1 to |s2|
 5  do m[0, j] = j
 6  for i ← 1 to |s1|
 7  do for j ← 1 to |s2|
 8       do m[i, j] = min{m[i − 1, j − 1] + if (s1[i] = s2[j]) then 0 else 1fi,
 9                        m[i − 1, j] + 1,
10                        m[i, j − 1] + 1}
11  return m[|s1|, |s2|]
```

# Efficiency Improvement

- How to limit the set of terms for which we compute edit distances to the query term?

# k-gram Idea for Spelling Correction

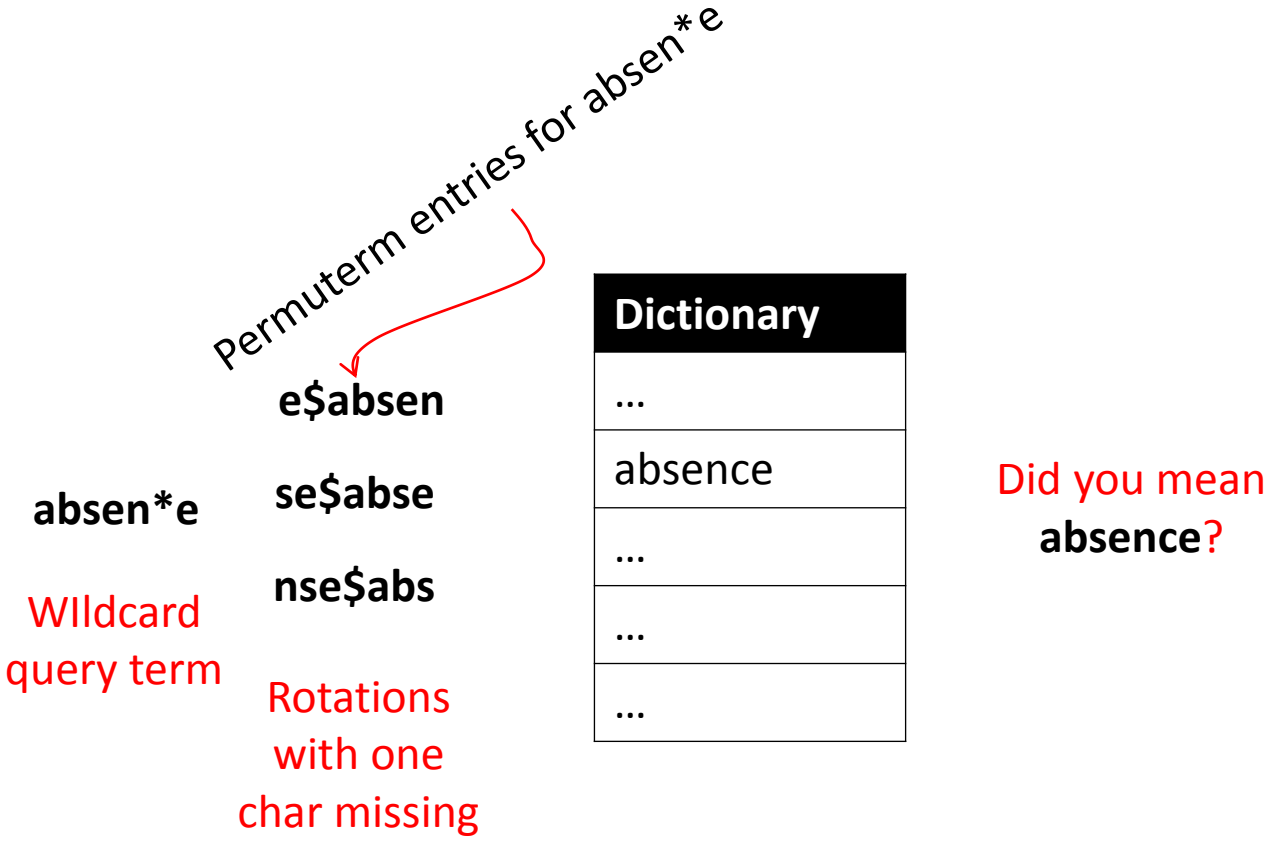
- Many heuristics lead to poor matches.
- For example, “bored” misspelt as “bord” may match “boardroom” if the heuristic is
  - Match any two bigrams
    - and we matched “bo” and “rd”
- Potential Solution
  - Compute Jaccard Similarity between k-grams of matched term and that of the query term.

# Jaccard Coefficient

- Jaccard Coefficient of two sets A and B
  = |A⋂B|/|AUB|
- Example: JS on bigrams of (“bord”, “boardroom”)
  = |{$b, bo, rd}|/|{$b, bo, or, rd, d$, oa, ar, dr, ro, oo, om, m$}|
  = 3/12.

*If you do not use end markings, we get 2/9.

# Permuterm Index




Compute edit distance for each query term with each of its permuterm based matches. Very expensive!
Assume first letter will be correct. Apply such heuristics.

# Context-Sensitive Spelling Correction

- Our heuristics may lead to
  - “flew form Delhi” → “flew fore Delhi”, “flew from Delhi”
- Surrounding words may determine the correction
- Potential Solution
  - Simply use query log frequency or collection frequency of these phrases to choose the best.