https://vvtesh.sarahah.com/

Information Retrieval

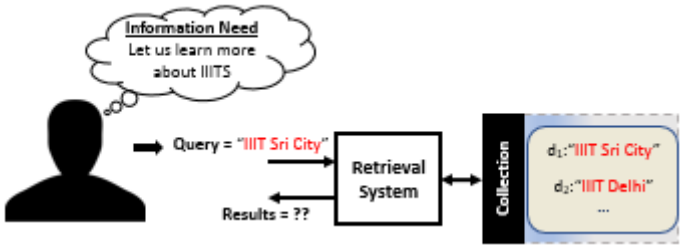
Venkatesh Vinayakarao

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

அட பாடல் போல தேடல் கூட ஒரு சுகமே
Ada Padal Pola Thedal Kooda Oru Sugame
Search, like a song, is also a joy.

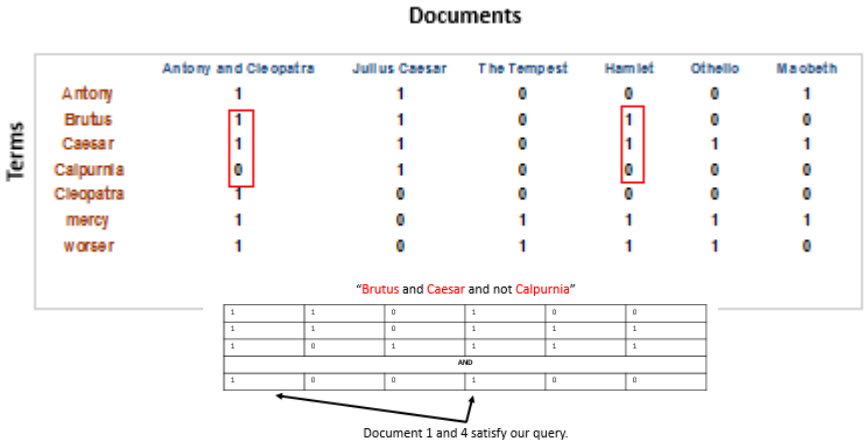
- From the movie, Thulladha Manamum Thullum. Lyrics by Vaali.

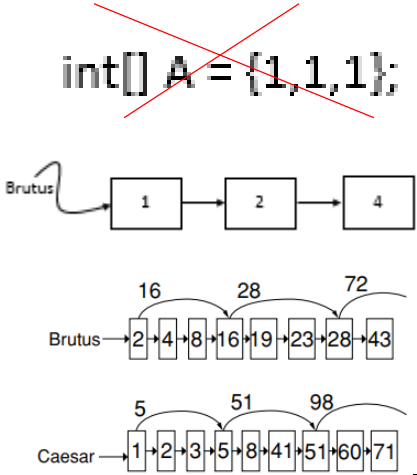
Review



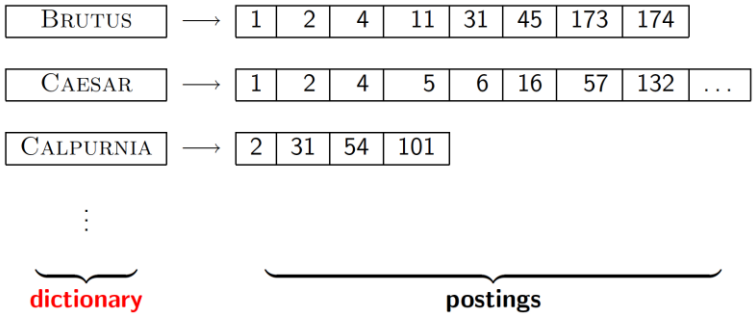
One (bad) Approach

- 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.

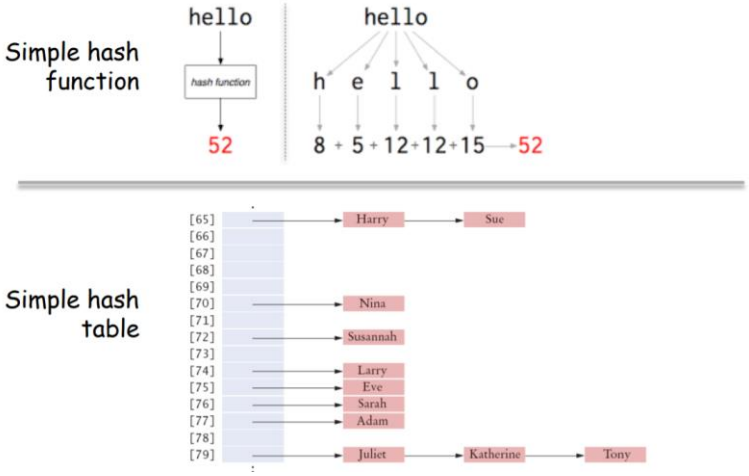




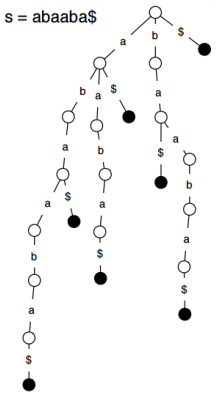
How to store a dictionary?



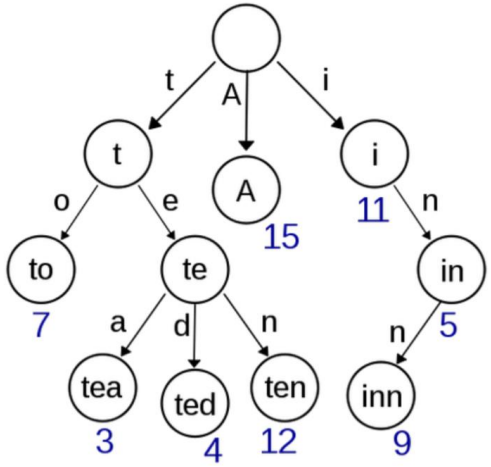
Hashing



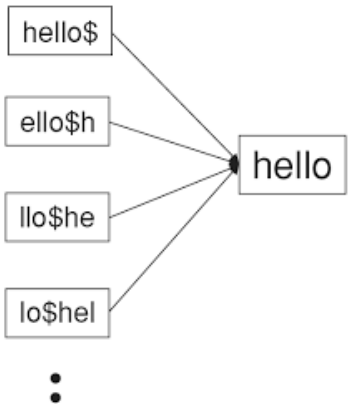
Suffix Trees



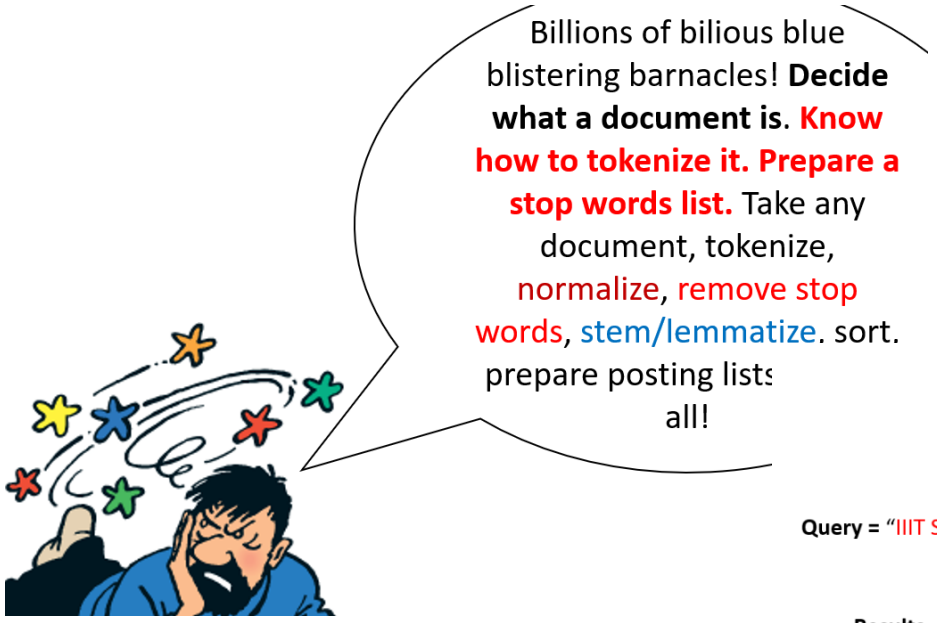
Trie

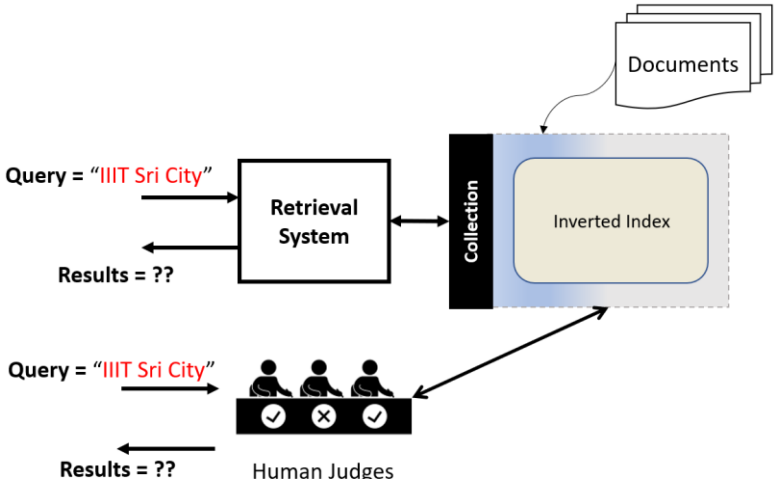


Permuterm Index



Review





Precision P = tp/(tp + fp)

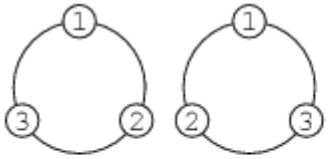
Recall R = tp/(tp + fn)

Permutation

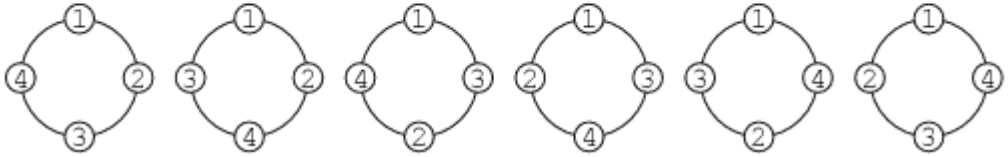
- How many ways can you arrange 123?
  - _ _ _
  - 3 options for first place.
  - 2 options for second place.
  - No option for the last place.
  - Therefore, 3 * 2 * 1 = 3! = 6 ways or 6 permutations.
- But distinct cyclic permutations are fewer.

*Remember, in permutation, order is important.

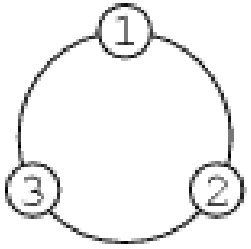
Cyclic Permutations



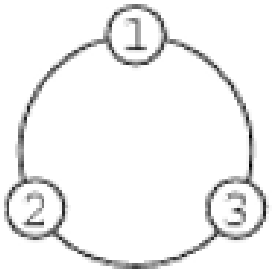
(3 – 1)! = 2! = 2 ways only.



If flipping is not allowed,



Is same as



$\frac{(3-1)!}{2}$ **= 2!/2 = 1 way only.**

Quiz

How many different necklaces can be made from 20 beads, each of a different color?

Spelling Correction

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.

Quiz

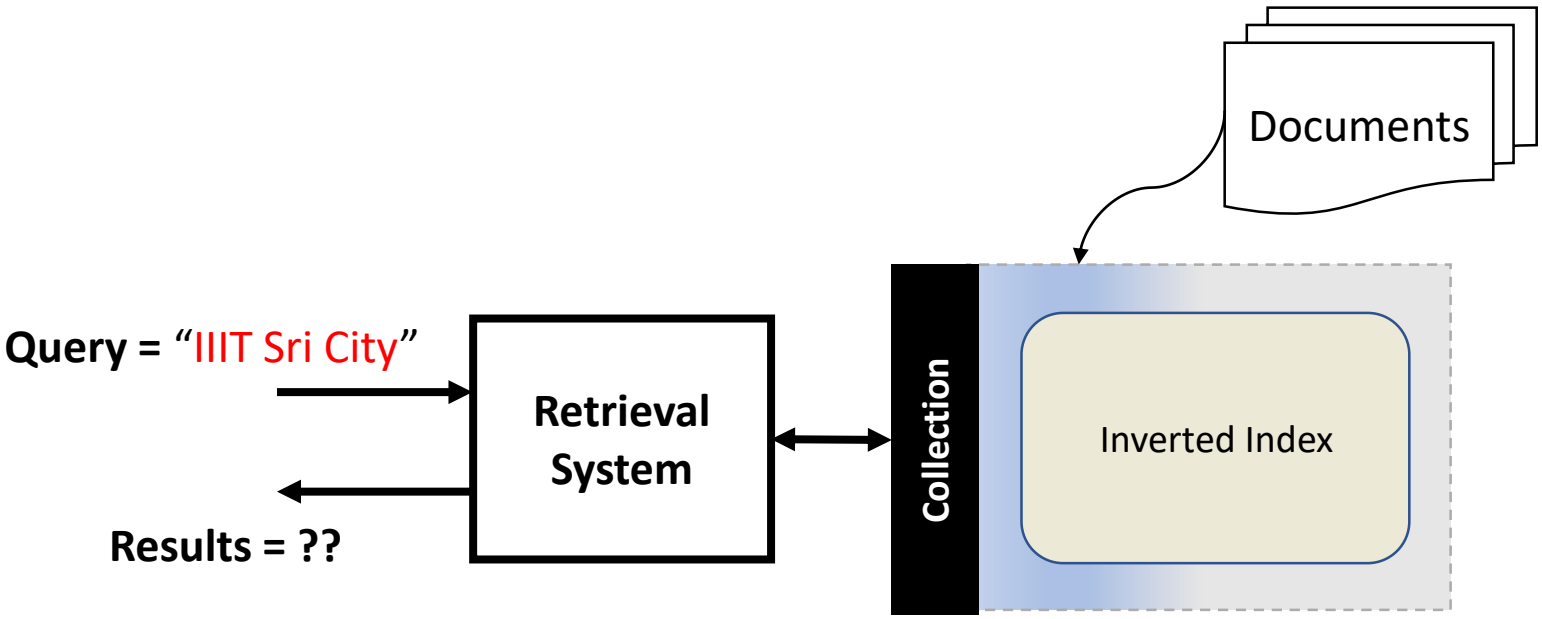
What is the edit distance between Sunday and Saturday?

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.

Phrasal Queries

- What if we do not want to match IIIT Delhi?



One (bad) Approach

- Index all biwords
  - Friends, Romans, Countrymen → Friends Romans, Romans Countrymen
- What is the problem?
- How do you match the query IIIT Sri City, Chittoor?
  - IIIT Sri AND Sri City AND City Chittoor must exist.

A Better Approach

- Store Positional Information

<term, number of docs containing term;
doc1: position1, position2 … ;
doc2: position1, position2 … ;
etc.>

Positional Index

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⟩; . . . ⟩

“to” appears six times in d1 at positions 7, 18, ….
“to” appears 993K times overall.

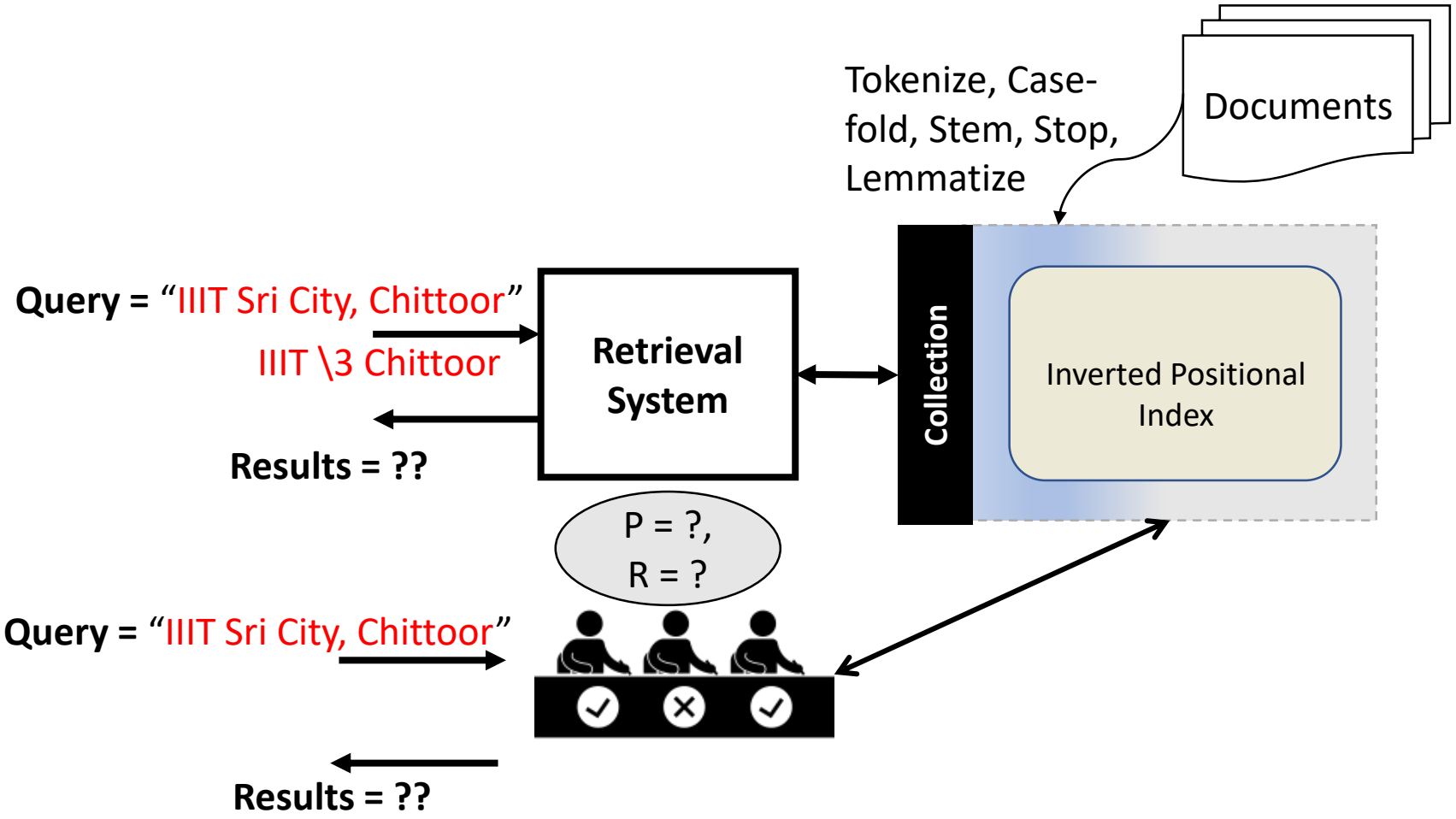
Which document is likely to contain “to be”?

Proximity Search

- IIIT /3 Chittoor
  - /k means “within k words of (on either side)”

- Merging postings is expensive
  - Index well-known phrases such as “Taj Mahal”

How does a Search Engine Work?



Project Ideas 

- Develop a Inverted **Positional Index** for Tamil/Telugu.
  - Dataset: Use any content from web which uses these languages.
- Develop a **Search Engine** (using Apache Lucene) with inverted positional index and demonstrate a phrasal query working.

Remember, all projects must implement a search engine.

Indexing Big Data!

REUTERS

You are here: Home > News > Science > Article

Go to a Section: U.S. International Business Markets Politics Entertainment Technology Sports Oddly Enoug

Extreme conditions create rare Antarctic clouds

Tue Aug 1, 2006 3:20am ET

Email This Article | Print This Article | Reprints

[-] Text [+]

SYDNEY (Reuters) - Rare, mother-of-pearl colored clouds caused by extreme weather conditions above Antarctica are a possible indication of global warming, Australian scientists said on Tuesday.

Known as nacreous clouds, the spectacular formations showing delicate wisps of colors were photographed in the sky over an Australian meteorological base at Mawson Station on July 25.

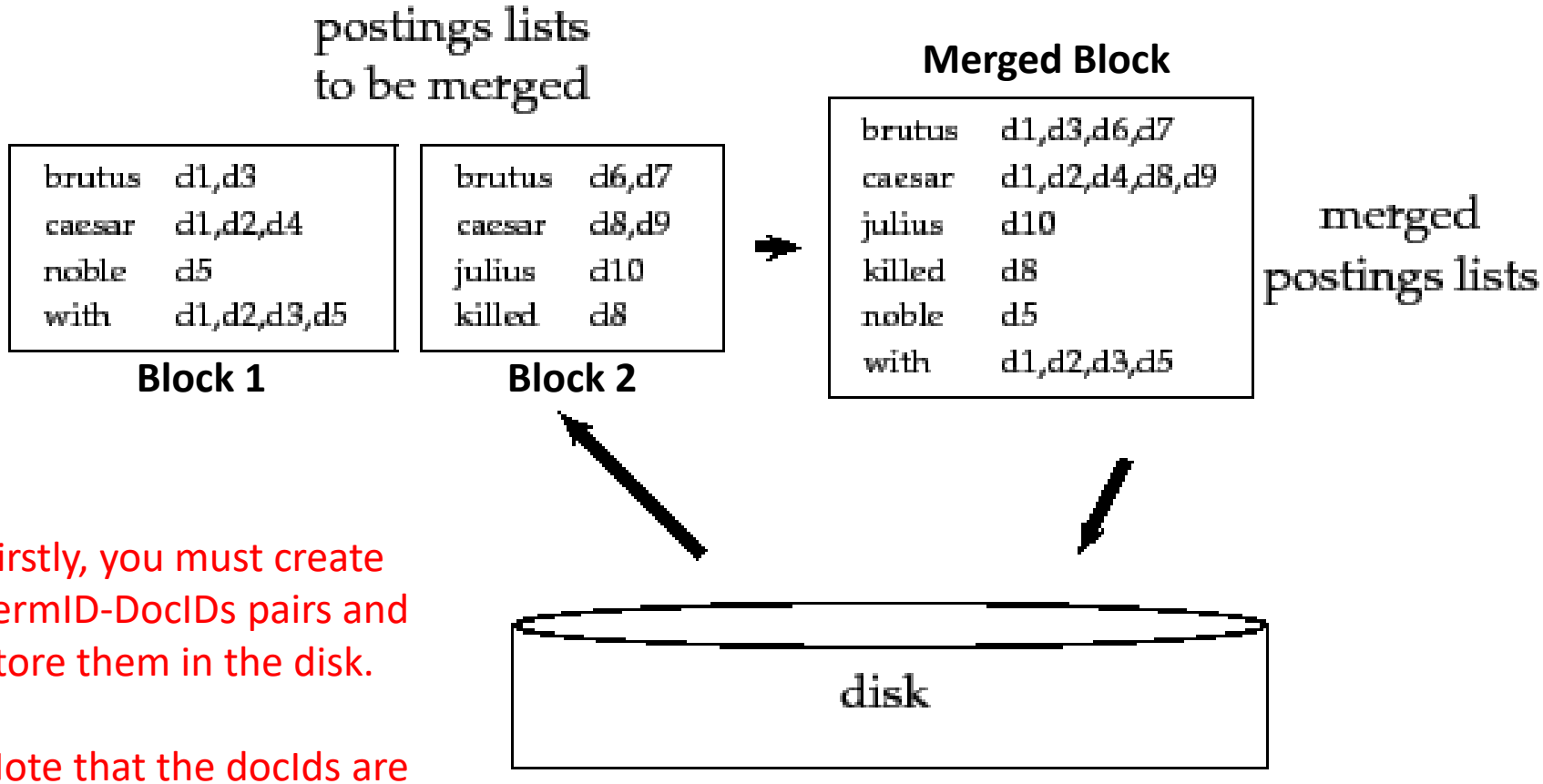
Reuters RCV1 Corpus

symbol	statistic	value
N	documents	800,000
L	avg. # tokens per doc	200
M	terms (= word types)	400,000
	avg. # bytes per token (incl. spaces/punct.)	6
	avg. # bytes per token (without spaces/punct.)	4.5
	avg. # bytes per term	7.5
	non-positional postings	100,000,000

Indexing

- How can we construct an index for very large collections?
  - Hard disk space is limited.
  - RAM is limited.
- Easier Questions
  - How to sort huge amount of data?
    - External sorting (Merge Sort) [Complexity is O(nlogn)
  - How to leverage clusters of machines?
    - Map-Reduce

Blocked Sort-Based Indexing



Firstly, you must create termID-DocIDs pairs and store them in the disk.

Note that the docIds are already sorted.

Single Pass In-Memory Indexing

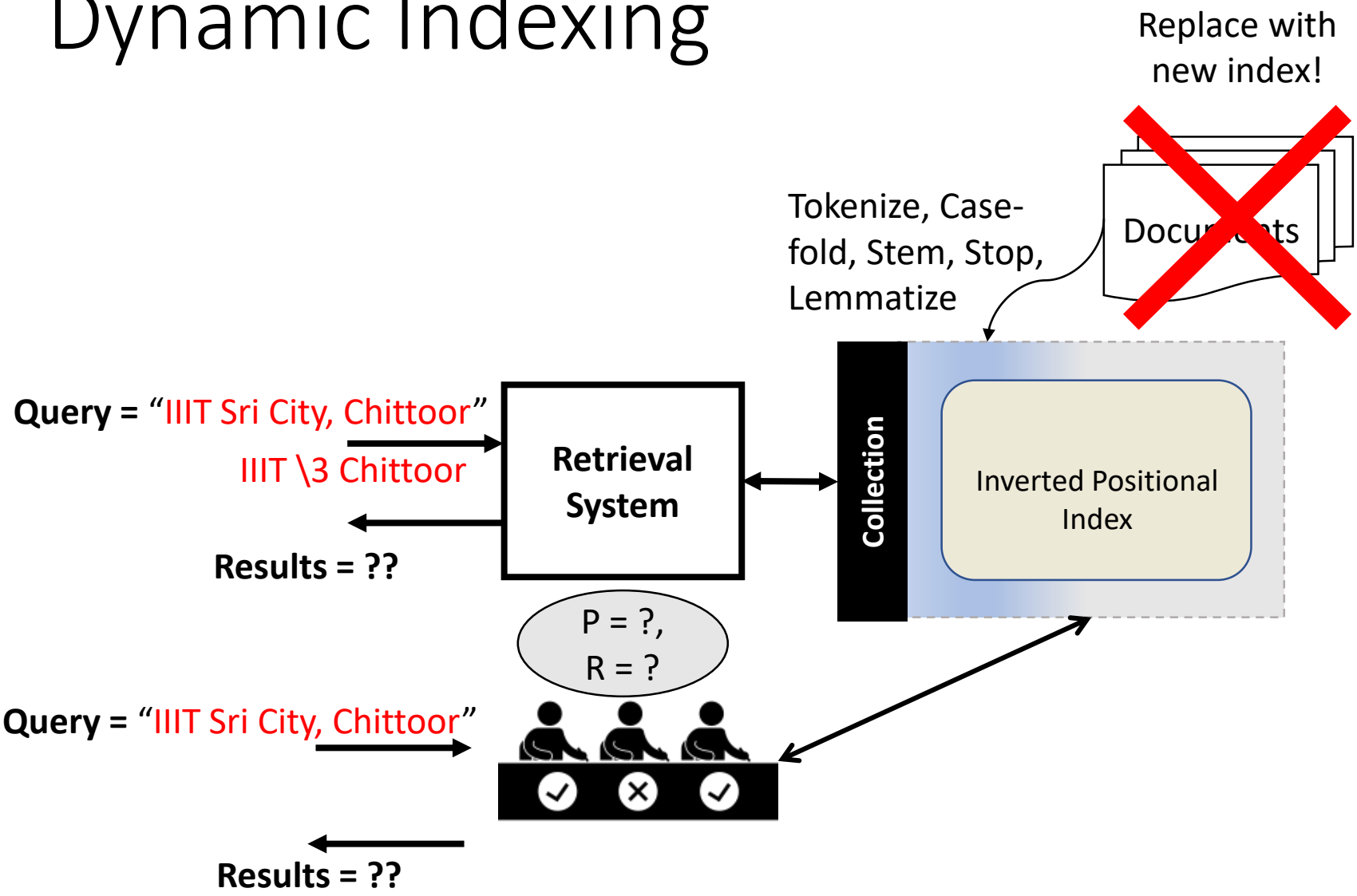
- No upfront creation of termID-DocIDs pairs.
- Parse each document
  - Identify and add terms to dictionary.
  - Create a posting list or add docID to the existing posting list.

Google Data Centers

- Contain commodity machines distributed around the world.
- Estimate: 1 million servers = 3 million processors/cores (Gartner 2007)
- Estimate: Google installs 100,000 servers each quarter.
  - Based on expenditures of 200–250 million dollars per year
- This would be 10% of the computing capacity of the world!?!

Source: cecs.wright.edu/~tkprasad/

Dynamic Indexing



One Approach

- One “big” main index
- New documents go into “small” auxiliary index
- Search across both, merge results

Can we do better?

Logarithmic Merge

Index Compression

The Rule of 30

The 30 most common words account for 30% of the tokens in written text.

Add a, an, the, … to the stop words list.

Quiz

Given a collection (of documents), how to estimate the number of terms?

One (bad) Approach

- Use Oxford Dictionary
  - Oxford English Dictionary defines 600K+ words.
- The Problem
  - Does not contain people, places, products which users may query.

Heap's Law

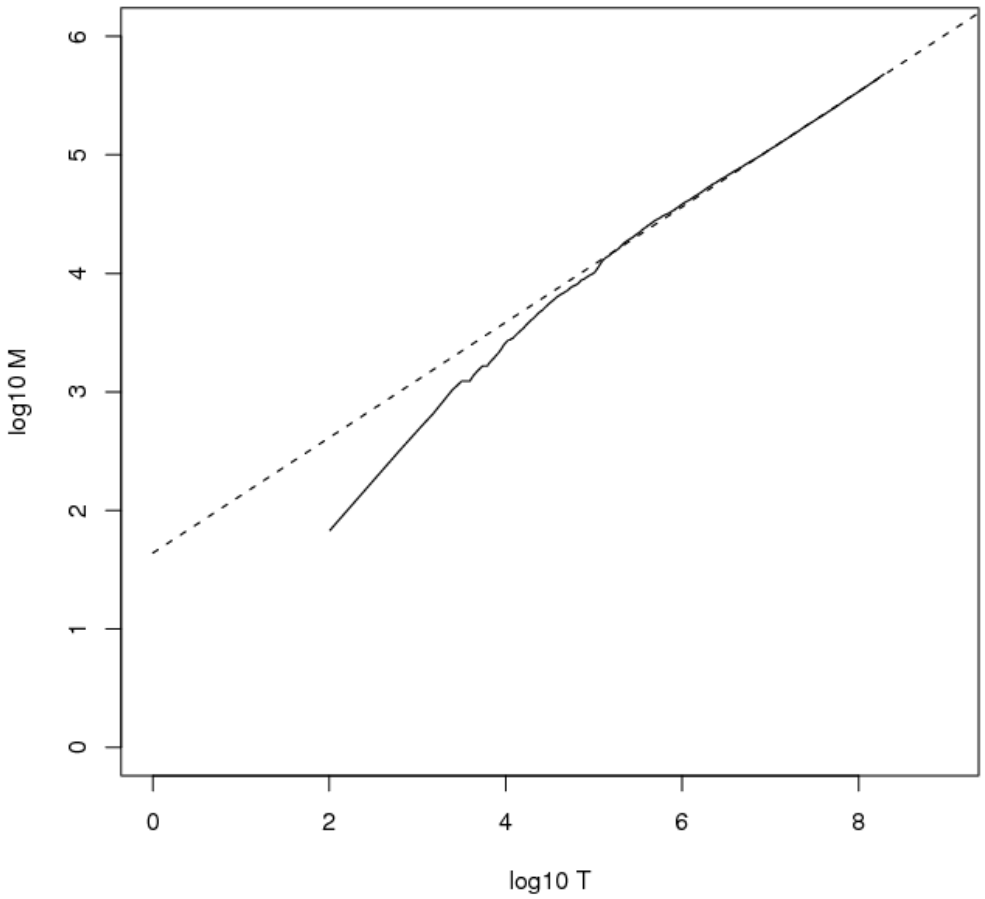
- An Empirical Finding (from experiments on several datasets)

$$M = kT^b$$

- M is the size of the vocabulary, T is the number of tokens in the collection

Heap's Law

- $\log_{10}M = 0.49 \log_{10}T + 1.64$ is the best least squares fit for RCV1.
- So $k = 10^{1.64} \approx 44$ and $b = 0.49$.
- For first 1,000,020 tokens, law predicts 38,323 terms; actually, 38,365 term



Takeaways from Heap's Law

- Dictionary Size grows with collection size.
- Size of dictionary can get "really" large!

Project Ideas 

- Using few large **non-English** collections, estimate the number of terms in semi-automated ways. Validate if Heap's law holds.
- Create a stop-words list for this language.
- Develop a **Search Engine** (using Apache Lucene) which indexes these terms and demonstrate simple term queries.

Remember, all projects must implement a search engine.

Term Frequency

- What are top-3 frequent terms in the text given below? Give the frequency of those terms.

Being an excellent student has more benefits than just getting good grades. In the short term, it will make you a more appealing college candidate and, in many cases, can earn you some fairly hefty scholarships. Big picture, the skills you learn at school will stick with you for the rest of your life, helping you tackle any problem that comes your way.

Repeating Terms

- Give me the term frequency for each repeating term.

Being an excellent student has more benefits than just getting good grades. In the short term, it will make you a more appealing college candidate and, in many cases, can earn you some fairly hefty scholarships. Big picture, the skills you learn at school will stick with you for the rest of your life, helping you tackle any problem that comes your way.

Repeating Tokens are Highlighted

- you = 6, the = 3, in = 2, your = 2 (case-folded)

Being an excellent student has more benefits than just getting good grades. In the short term, it will make you a more appealing college candidate and, in many cases, can earn you some fairly hefty scholarships. Big picture, the skills you learn at school will stick with you for the rest of your life, helping you tackle any problem that comes your way.

Zipf's Law

The i^{th} most frequent term has frequency proportional to $\frac{1}{i}$

Do not apply on small examples…

Being an excellent student has more benefits than just getting good grades. In the short term, it will make you a more appealing college candidate and, in many cases, can earn you some fairly hefty scholarships. Big picture, the skills you learn at school will stick with you for the rest of your life, helping you tackle any problem that comes your way.

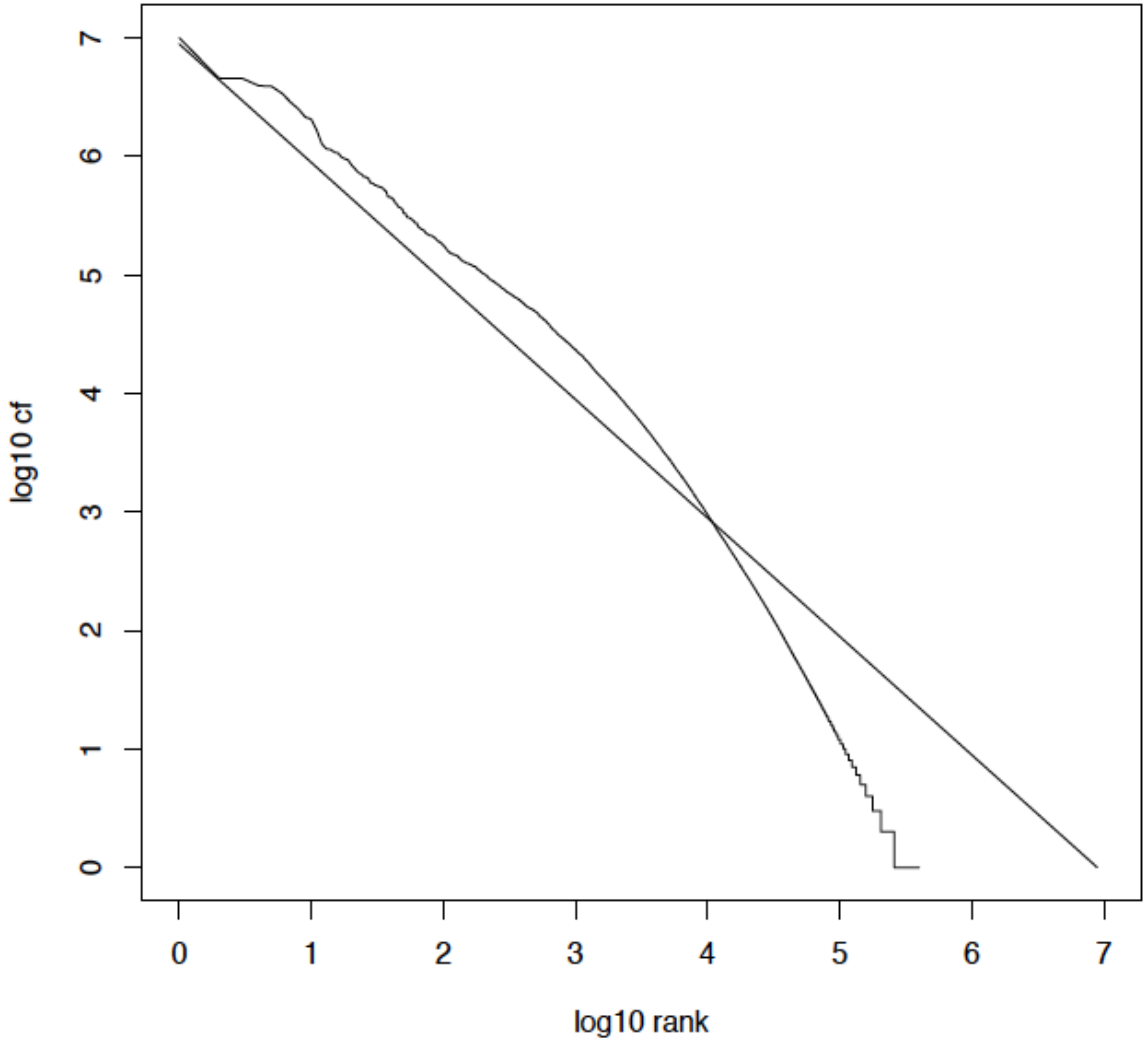
Only for illustration.

Rank	Frequency
1 (you)	6
2 (the)	3
3 (in)	2

$cf_i \propto 1/i = k/i$

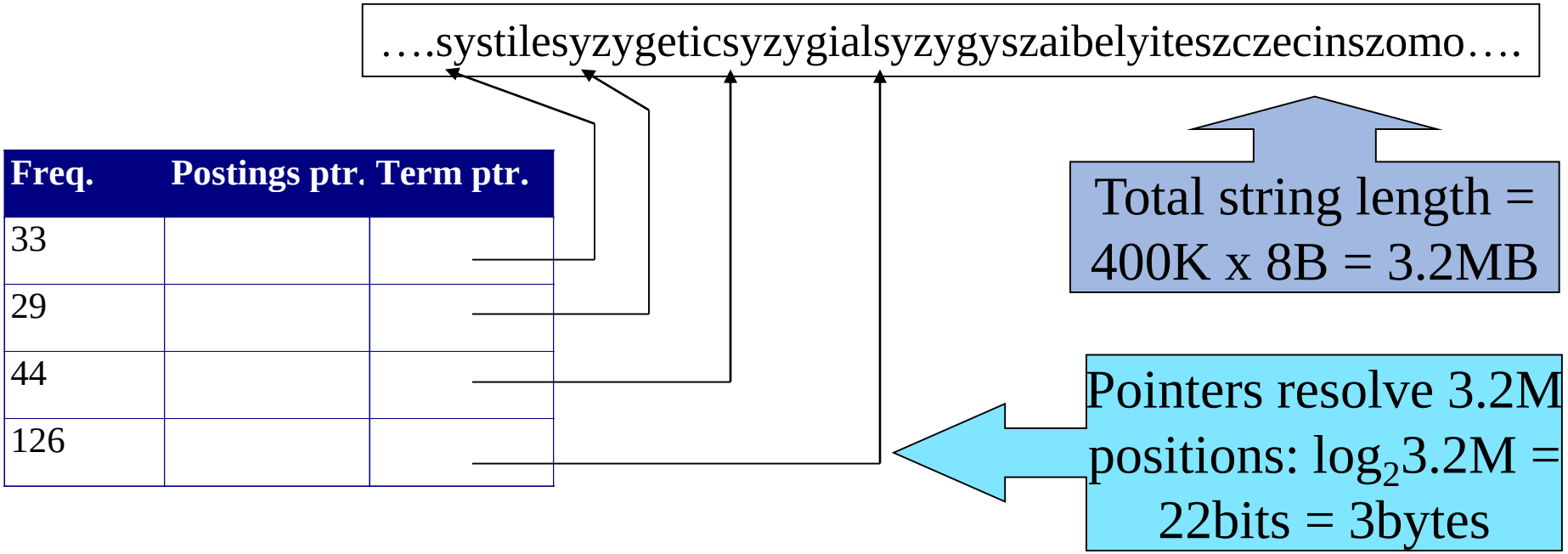
k = 6 in our case!

Zipf's law for Reuters RCV1



Compressing the term list: Dictionary-as-a-String

- Store dictionary as a (long) string of characters:
  - Pointer to next word shows end of current word
  - Hope to save up to 60% of dictionary space.



Compressing Postings

How to store a large number of "numbers" efficiently?

Store Gaps

- 33,47,154,159,202 …
- 33,14,107,5,43 …
- most gaps can be encoded/stored with far fewer than 20 bits.

Questions