https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

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

What we find changes who we become.

-Peter Morville.

Acknowledgment

Some slides are borrowed from the companion website of Manning et al.'s IR book
(https://nlp.stanford.edu/IR-book/)

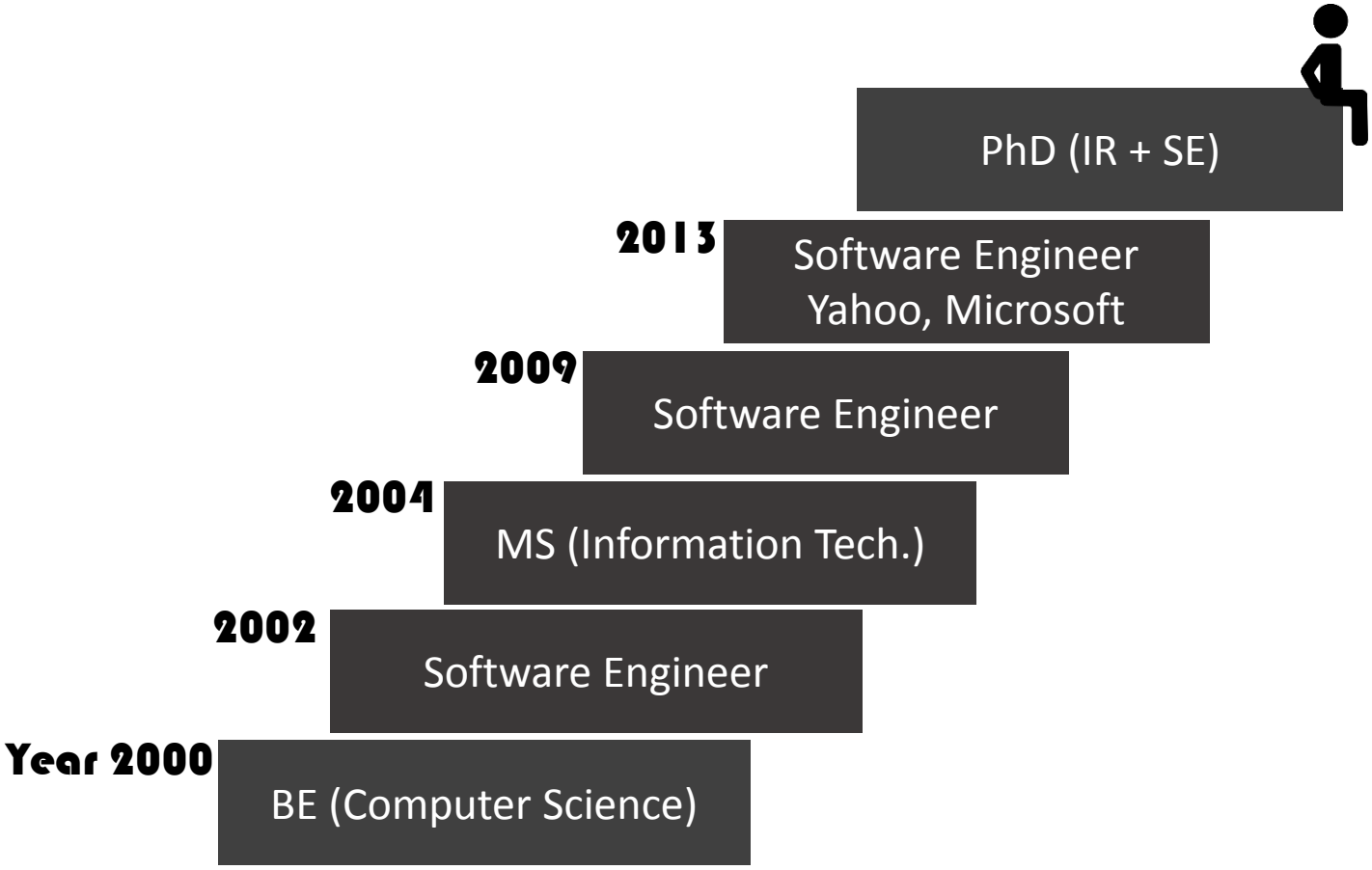
A good teacher can inspire hope, ignite the imagination, and instill a love of learning.

-Brad Henry.

Agenda

- About Me
- Introduction
- Course Dynamics
- Our First IR System
  - Linear Traversal
- Our Second IR Approach
  - Bit Vector Representation and Boolean Retrieval
- Our Third IR Approach
  - Inverted Index and Posting Lists

About Me


PhD (IR + SE)
2013
Software Engineer
Yahoo, Microsoft
2009
Software Engineer
2004
MS (Information Tech.)
2002
Software Engineer
Year 2000
BE (Computer Science)

Your Instructor Can Speak…

- Or at least understand to some extent:
  - Telugu (Son's most fluent language)
  - Tamil (Born and brought up in Chennai)
  - Marathi (Mother Tongue)
  - Malayalam (First four years of schooling in Irinjalakuda)
  - Kannada (Married to a Kannadiga)
  - Hindi (Last 5 years in Delhi)
  - English (Confused with all the above, chose to use English wherever possible)

Venkatesh Sir

Introduction

Information

Shannon's Definition, Fisher Information, Neumann Entropy, … 

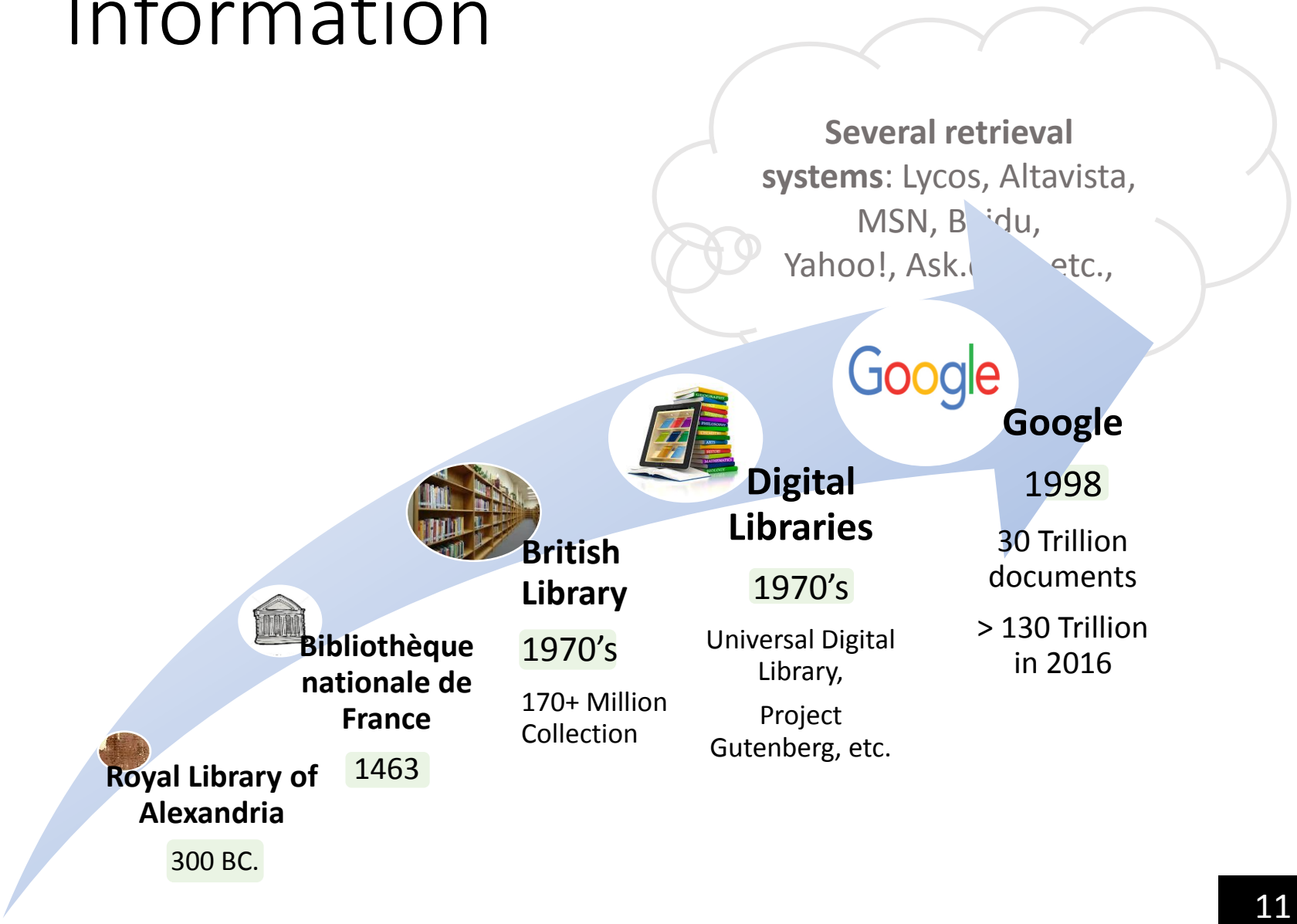
Information is any entity or form that provides the answer to a question of some kind or resolves uncertainty. – Wikipedia.

Role of Information

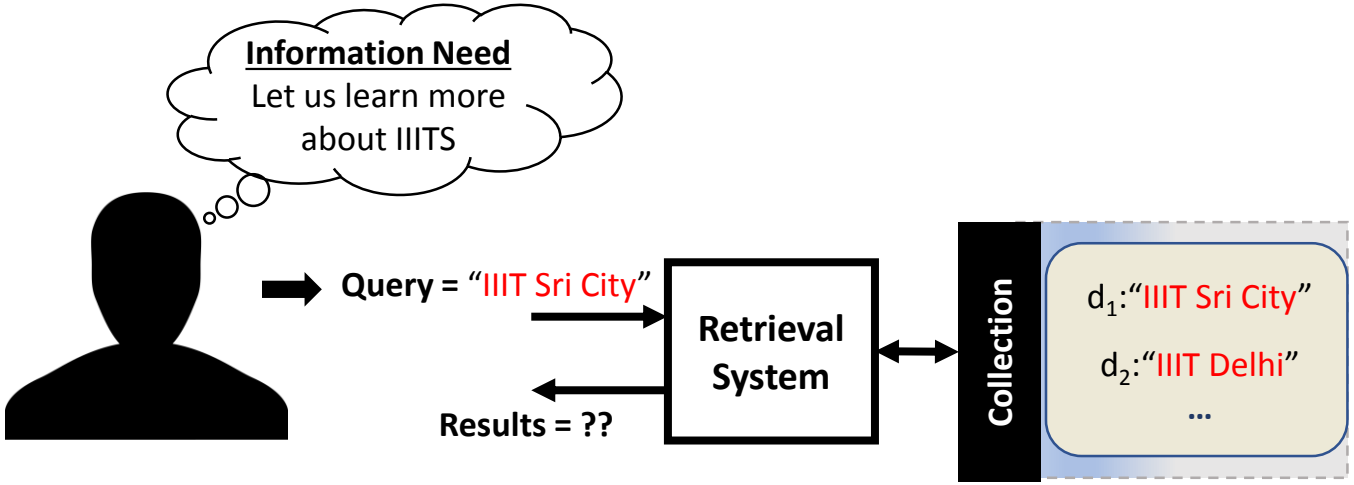
- If you only knew
  - Which stock to invest in?
  - Which faculty to work with?
  - How to get into a top college?
  - Which course to register for?
  - What to study?
  - How to prepare for job interviews?
  - …
- If only you had the information, you could rule this world!
- What happens when all the information is deprived from you?

Solitary Confinement is Cruel

Information



What is Information Retrieval?



Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an **information need** from within large collections.

– From the Manning et al. IR Book.

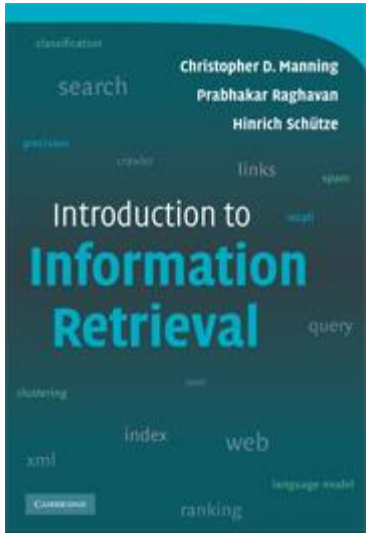
Course Dynamics

Learning Objectives

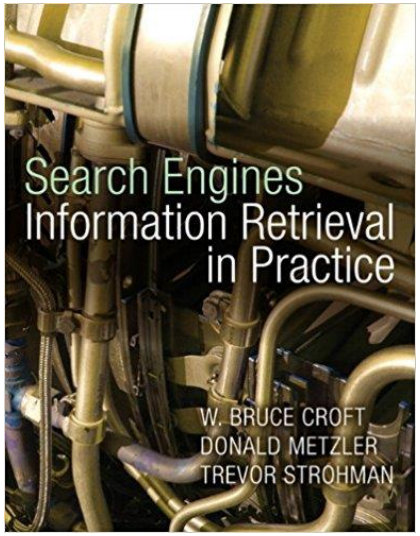
- Understand and apply text retrieval techniques to big data.
- Understand and apply text indexing techniques.
- Analyze and evaluate existing retrieval systems.

Course Website: http://vvtesh.co.in/teaching/IR.html

Resources



Course Text



Reference

Evaluation

Instrument	Max Marks
Mid-Term 1	20%
Mid-Term 2	20%
Final	30%
Assignment 1 (2%)	2%
Assignments 2, 3 (4% each)	8%
Project	20%

Assignment 1

- Make a 2 min (minimum 1 min to max 2 min) video on “Why Should we Study Information Retrieval?”
  - Rule 1: You should feature in that video. Yes, it is mandatory for all group members to show up.
  - Rule 2: Grading will be done based on “Interestingness – Richness of content discussed”, and “Clarity – quality of audio and video”.
  - Rule 3: Your mobile video is sufficient. Ensure that your file is not bigger than 50 MB.
  - Rule 4: Deadline to submit is 17th Aug 2018 9 pm Indian Standard Time.
  - Rule 5: Your TA or your instructor will announce submission details on or before 16th 9 pm.

Assignments

- Assignment 2 & 3
  - May (Not necessarily though) have a programming component.
  - Will test the concepts you study.
- All assignments (and project) can be done in groups of (max) four.
  - Remain in the same group for both projects and assignments.
- Exams are individual.

Project

- Mandatory.
- You are required to build a search engine.
  - A two page technical report covering the approach.
  - A live demo of the project to a TA.
  - A 15 minute presentation to the instructor.
- Evaluation will be based on:
  - Application of one or more concepts learned in the class.
  - Novelty of the idea.
  - Quality of implementation.
- There will be intermediate deadlines and milestones.
- Detailed guidelines will follow after Mid-Term 1.

Exams

- Closed Book.
- You may carry one page (A4 Sheet or smaller) hand-written notes to all exams. Clearly write your name and roll number on these notes.
- Expect some or even all multiple choice questions
  - Expect negative marks for wrong answers.
- Exams are designed to evaluate if you understood the concepts and if you can apply your knowledge.

Exams

- No make-up exams for mid-sems even for extra-ordinary circumstances (as per institute policy).
- End-sem make-up happens only on Dean's approval
  - Even if it happens, it will be significantly tougher than the actual exam.
  - Students are strongly discouraged from considering a make-up.

Office Hours

- Up to 15 minutes after each class.
  - Find me in Room 101.
- Otherwise, by appointment.
  - Send me an email.
  - Keep “[IR Class]” on subject line.
  - Preferred Venue: CSA Cellar Canteen. Preferred Time: Evening 6pm to 7pm on Fridays. Please drop a mail if you need to meet.

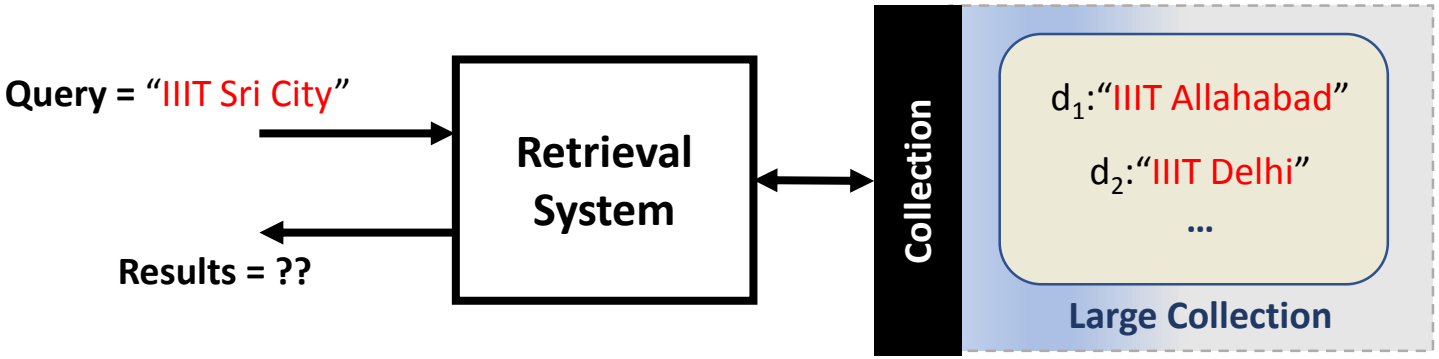
A Simple Retrieval System

Our first IR system.

Simple Retrieval Problem

- A **collection** with 5 **documents** having the following contents
  - d1: IIIT ALLAHABAD
  - d2: IIIT DELHI
  - d3: IIIT GUWAHATI
  - d4: IIIT KANCHIPURAM
  - d5: IIIT SRI CITY
- **Query** is
  - IIIT SRI CITY
- Which **document** will you match and why?

The Problem


Query = "IIIT Sri City"
Results = ??
Retrieval System
Collection
d_1:"IIIT Allahabad"
d_2:"IIIT Delhi"
...
Large Collection

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.

Three iterations!
Quiz: Can we do better?

A Simple Retrieval Exercise

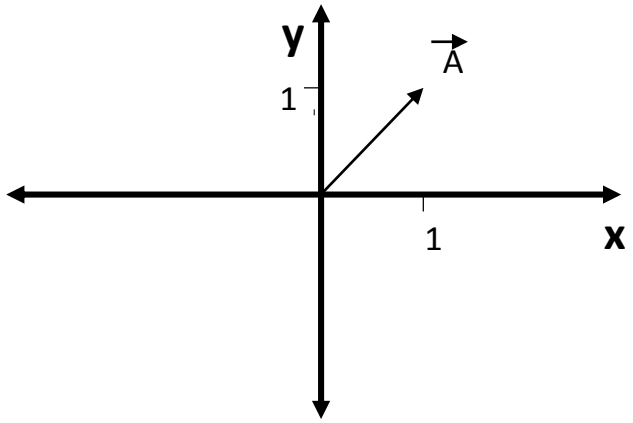
Our second IR system.

A Better Approach

Revisiting
Linear Algebra

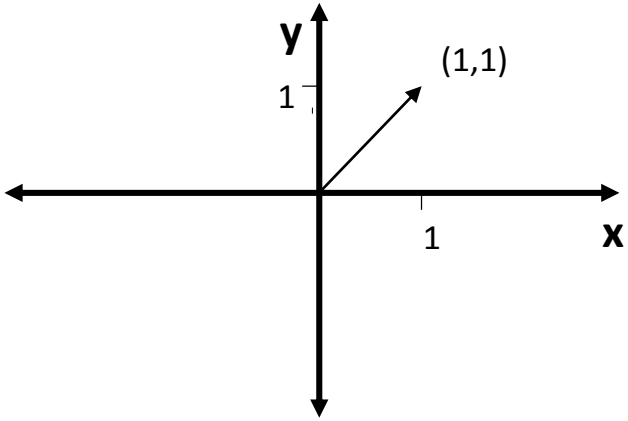
Vectors

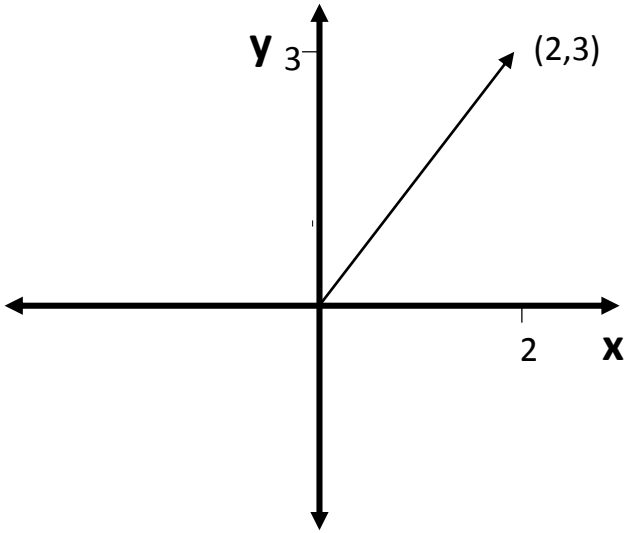
- Geometric entity which has magnitude and direction



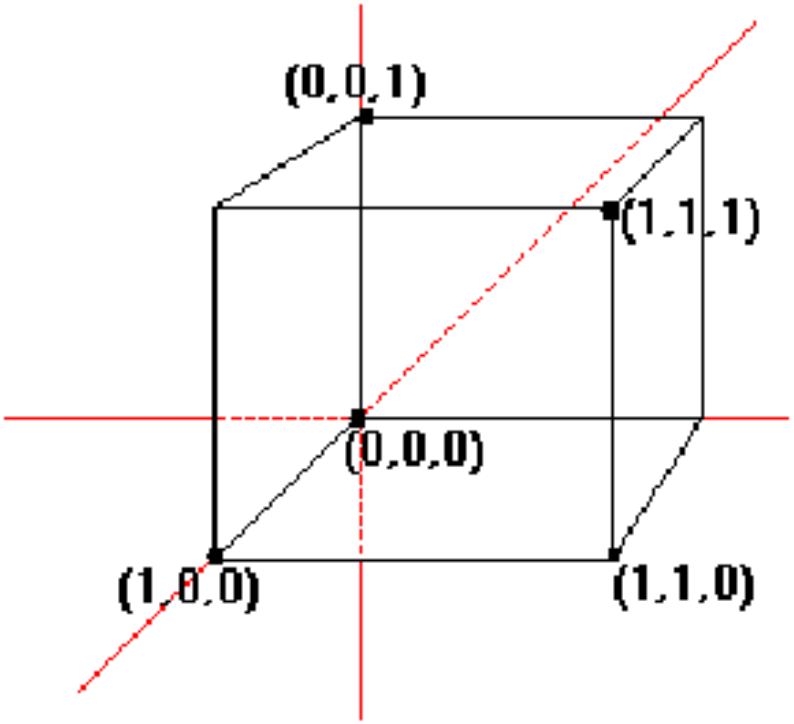
- If (x,y) is our vector of interest, this figure shows $\vec{A}$ vector = (1,1).

How is (2,3) Different?





What is (1,1,1) ?

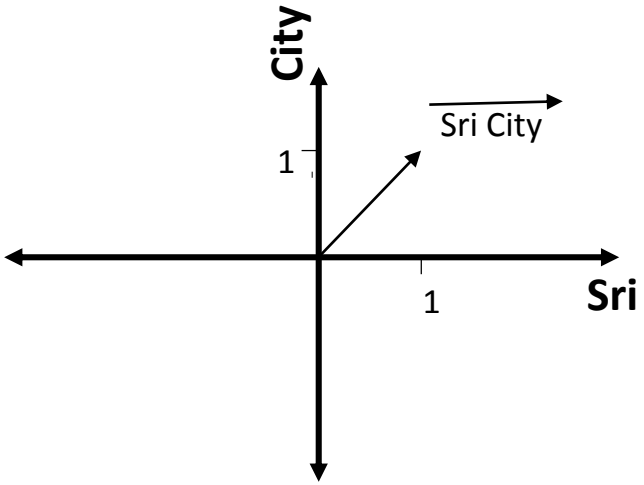


Remember!

A number is just a mathematical object. We give meaning to it!

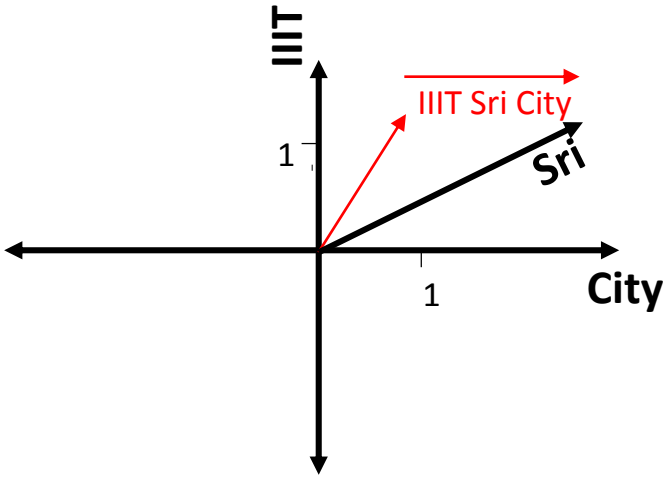
Sentences are Vectors

- “Sri City” as a vector



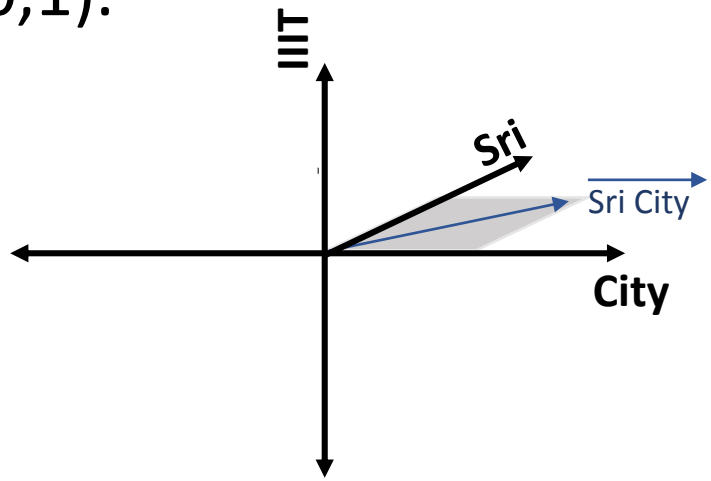
Sentences are Vectors

- "IIIT Sri City" is a 3-dimensional vector



Sentences are Vectors

- On this 3D space, “Sri City” vector will lie on the x (City) and z (Sri) plane. If (x,y,z) denotes the vector, “Sri City” is (1,0,1).



More Linear Algebra…

- So, we learned to represent English phrases on the vector space.
- We need something more!

Revisiting Matrices

Natural Language Phrases as Vectors

Let query q = "IIIT Sri City".

Let document, d_1 = "IIIT Sri City" and d_2 = "IIIT Delhi".

	IIIT	Sri	City	Delhi
q	1	1	1	0
d_1	1	1	1	0
d_2	1	0	0	1

q = (1,1,1,0), d_1= (1,1,1,0) and d_2 = (1,0,0,1)

Quiz

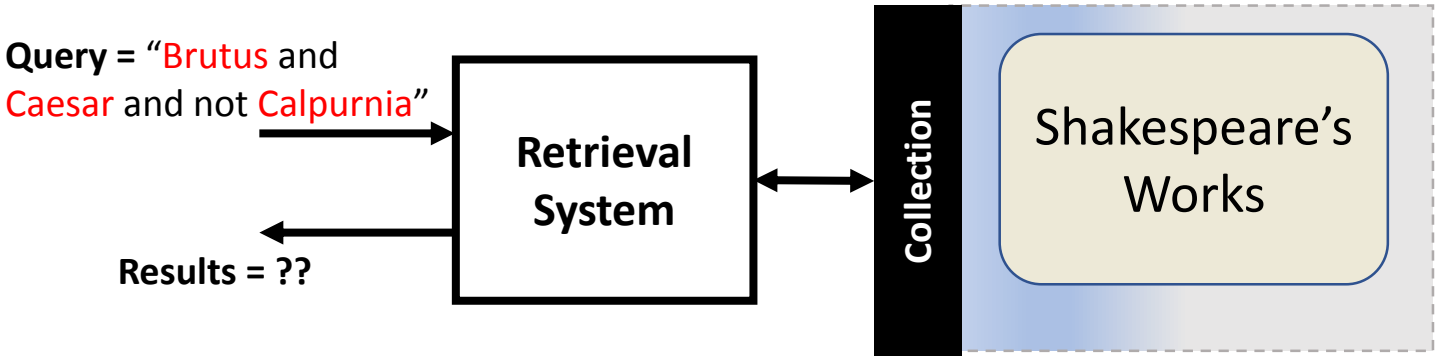
- Considering the following vectors:

	IIIT	Sri	City	Delhi
q	1	1	1	0
d_1	1	1	1	0
d_2	1	0	0	1

- What is the Natural Language (NL) equivalent of (0,1,1,0) ?
- What is the NL equivalent of (1,0,0,1) ?
- What is the vector for Delhi?
- If q represents query, d1 and d2 are documents, what is the NL query here?

Boolean Retrieval

The Problem


Query = "Brutus and Caesar and not Calpurnia"
Results = ??
Retrieval System
Collection
Shakespeare's Works

A term-document Matrix Example

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

"Brutus and Caesar and not Calpurnia"

Revisiting Boolean Algebra

What is the best way to get to the answer?

The Answer

"Brutus and Caesar and not Calpurnia"

1	1	0	1	0	0
1	1	0	1	1	1
1	0	1	1	1	1
AND					
1	0	0	1	0	0

Document 1 and 4 satisfy our query.

Our Third IR Approach

Inverted Index & Posting Lists Merging

Disadvantages of term-document Matrix

- When a new document is added to collection:
  - More distinct words are added to the matrix i.e., new columns get added.
- If the collection is very large (say Millions of documents),
  - Each document has far fewer words from the dictionary.
  - So, the matrix is sparse.

Can we do better?

Instead of handling both 1s and 0s, can we only have the 1s?

Revisiting Data Structures

Arrays Vs. Linked Lists

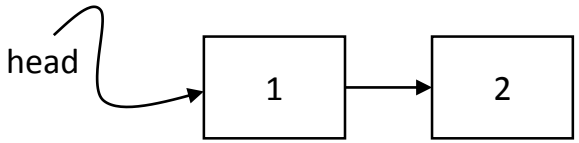
The Problem

- An n-Dimensional Vector can be represented as
  - an array of n elements.
    - Example: (1,1,1) is int[] A = {1,1,1}; in Java.
- So, a large vector {1,1,0,0,0,0,0,0,0,…. 10K elements} is
  - an array with 10K elements where only first two elements are 1s.

Is there a better way to represent this data?

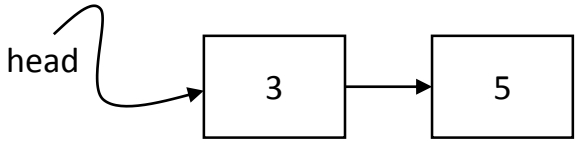
The Answer

- {1,1,0,0,0,0,0,0,0,…. 10K elements} is



A Linked List!

- {0,0,1,0,1,0,0,…..10K elements} is

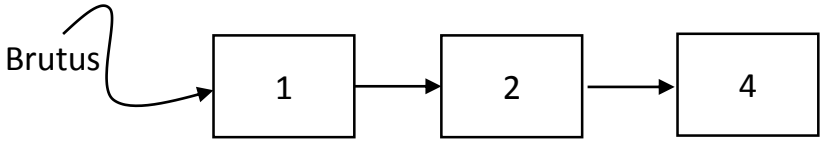


A Linked List!

Representing term-document Data

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



Tokenization

- Task
  - Chop documents into pieces.
  - Throw away characters such as punctuations.
  - Remaining terms are called tokens.
- Example
  - Document 1
    - I did enact Julius Caesar. I was killed i' the Capitol; Brutus killed me.
  - Document 2
    - So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Sort

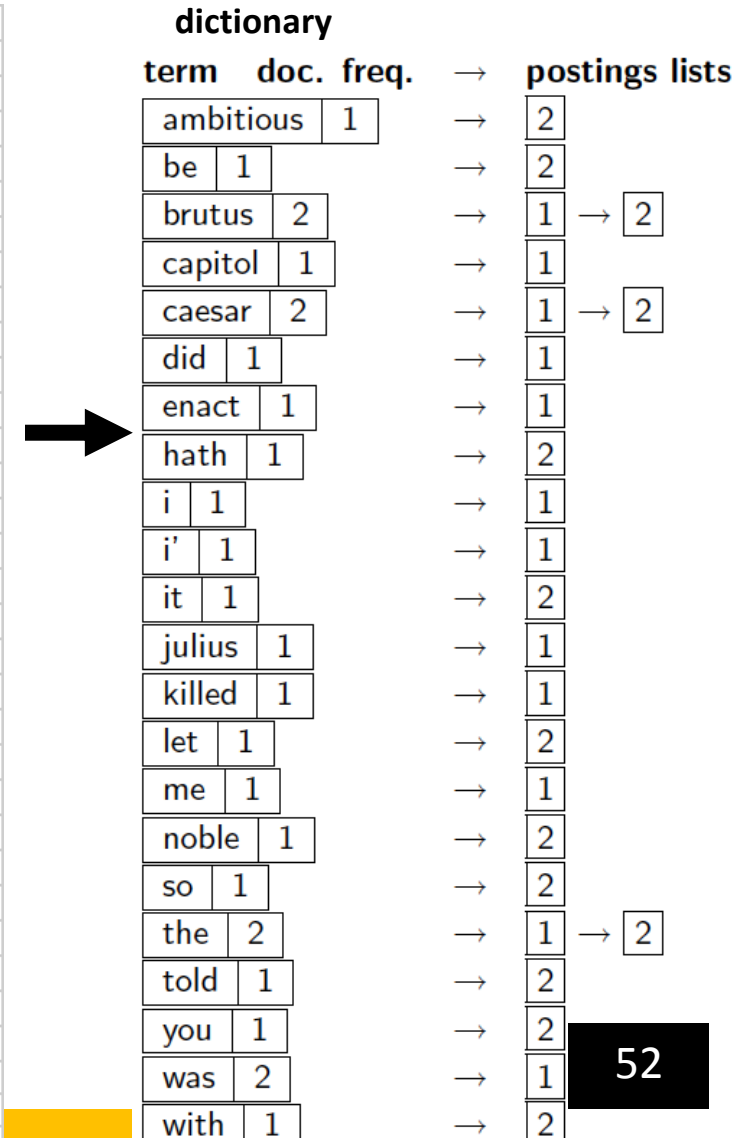
Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2

Dictionary & Postings

- Multiple term entries in a single document are merged.
- Split into Dictionary and Postings

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2



Query Processing with Inverted Index

Boolean queries: Exact match

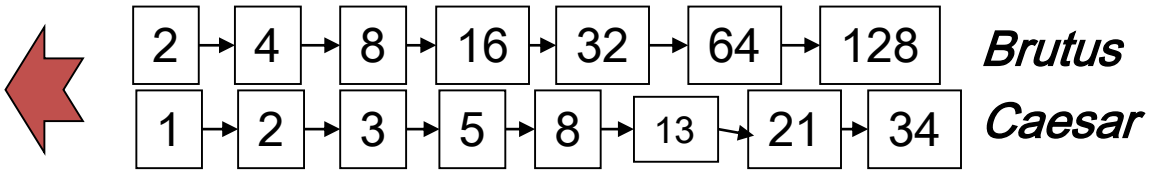
- The Boolean retrieval model is being able to ask a query that is a Boolean expression:
  - Boolean Queries are queries using *AND, OR* and *NOT* to join query terms
    - Views each document as a set of words
    - Is precise: document matches condition or not.
  - Perhaps the simplest model to build an IR system on

Query processing: AND

- Consider processing the query:

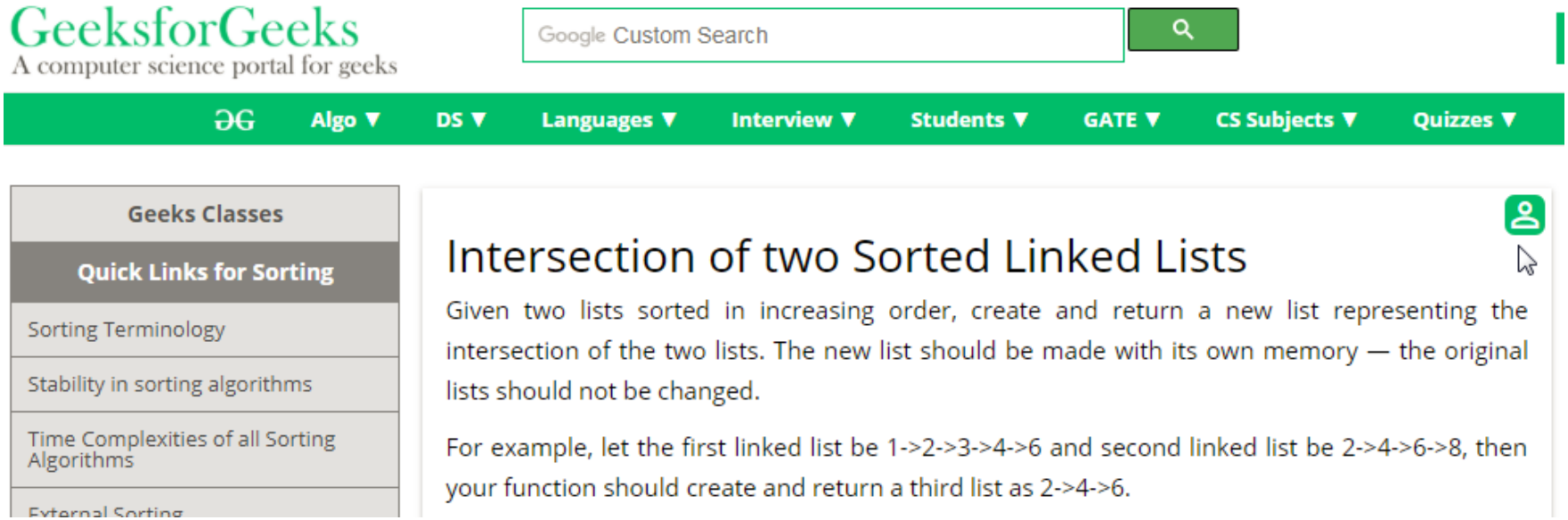
  Brutus *AND* ***Caesar***

  - Locate ***Brutus*** in the Dictionary;
    - Retrieve its postings.
  - Locate ***Caesar*** in the Dictionary;
    - Retrieve its postings.
  - "Merge" the two postings (intersect the document sets):



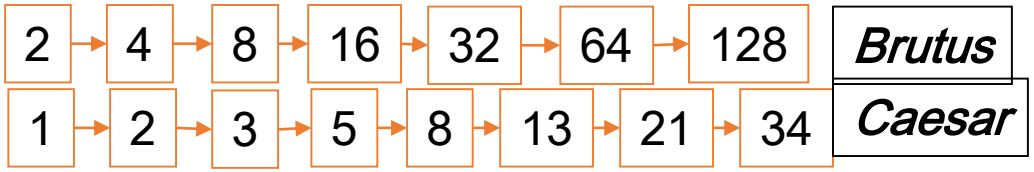
Common Interview Question

- https://www.geeksforgeeks.org/intersection-of-two-sorted-linked-lists/



The Merge

- Walk through the two postings simultaneously
  - Clue: Use two pointers



If the list lengths are *x* and *y*, the merge takes O(*x*+*y*) operations.
Crucial: postings sorted by docID.

The Big Picture

- Content Processing
  - Build Term Document Matrix or Build Inverted Index
- Query Handling
  - Boolean AND or Intersect the Posting Lists (called merging process)