Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Sep, 2019
Chennai Mathematical Institute

அட பாடல் போல தேடல் கூட ஒரு சுகமே
Search, like a song, is also a joy.

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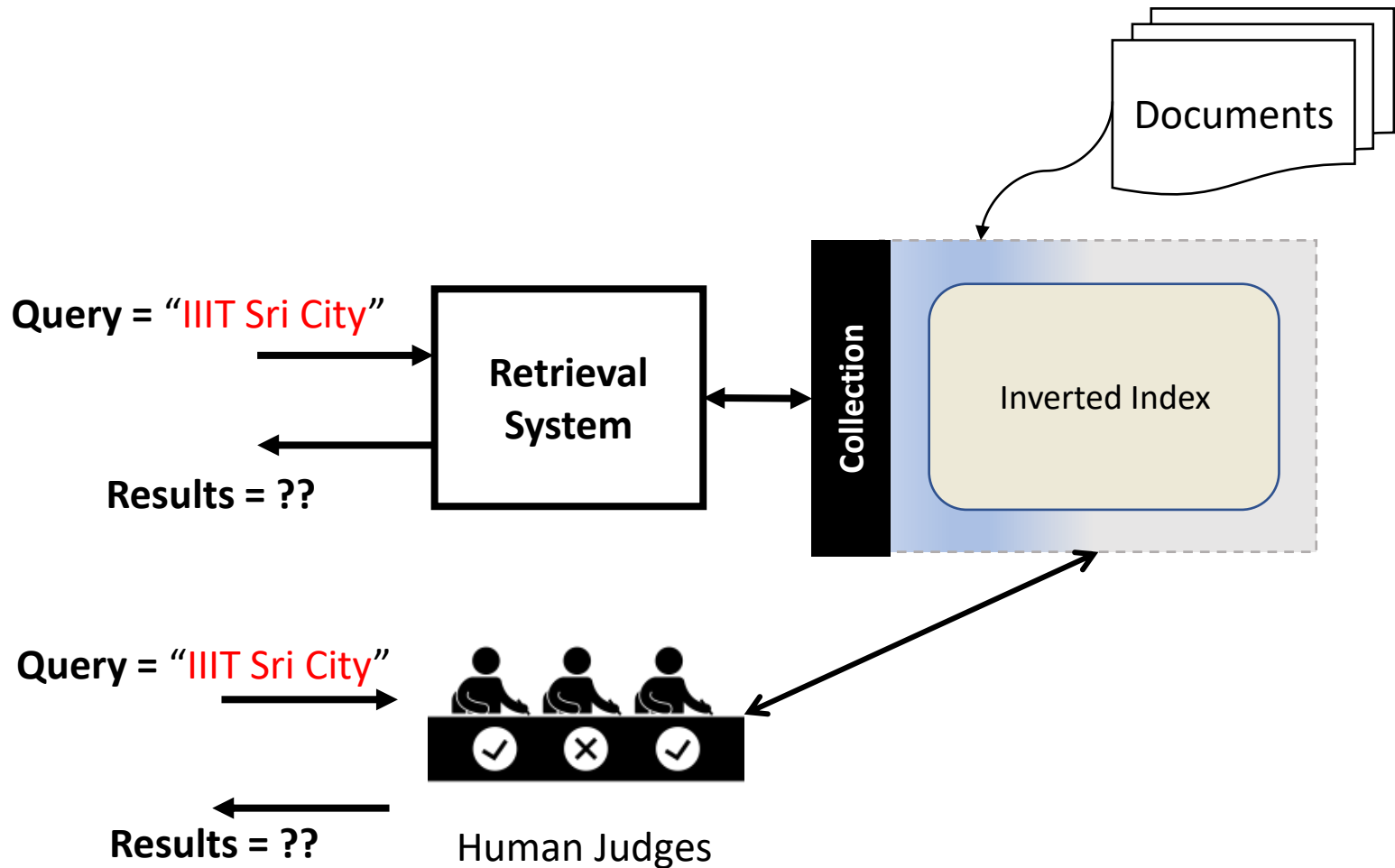
Evaluation

How Good is Our System?

- A **collection** having the following contents
  - d1: IIIT ALLAHABAD
  - d2: IIIT DELHI
  - d3: IIIT GUWAHATI
  - d4: ISI
  - d5: IIIT SRI CITY
  - d6: KREA SRI CITY
- **Query** is
  - SRI CITY
- **Result** is
  - IIIT SRI CITY
  - KREA SRI CITY

Very Good!

Evaluation

How Good is Our System?

- A **collection** having the following contents
    - d1: IIIT ALLAHABAD
    - d2: IIIT DELHI
    - d3: IIIT GUWAHATI
    - d4: ISI
    - d5: IIIT SRI CITY
    - d6: KREA SRI CITY
- **Query** is
    - IIIT
- **Result** is
    - IIIT SRI CITY
    - KREA SRI CITY

Not so Good!

Objective

We want all relevant documents and only relevant documents

Relevance

- How many **relevant** documents?
  - Four (IIIT SRI CITY, IIIT ALLAHABAD, IIIT DELHI, IIIT GUWAHATI)

- How many **retrieved** documents?
  - Two (IIIT SRI CITY, KREA SRI CITY)

How to quantify the "goodness" of our system?

Terminology

- Documents we see in results are "**positive**"
  - Positive
    - + IIIT SRI CITY,
    - + KREA SRI CITY
  - Negative
    - - IIIT ALLAHABAD
    - - IIIT DELHI
    - - IIIT GUWAHATI
    - - ISI

Terminology

- Documents that we correctly classify are "**true**"
  - Positive
    - \+ IIIT SRI CITY (true)
    - \+ KREA SRI CITY
  - Negative
    - \- IIIT ALLAHABAD
    - \- IIIT DELHI
    - \- IIIT GUWAHATI
    - \- ISI (true)

Here, query is "IIIT"

Quiz

- All retrieved results =
    1. tp + fp
    2. tp + fn
    3. tn + fp
    4. tn + fn

Legend
tp = true positive
tn = true negative
fp = false positive
fn = false negative

Quiz

- All retrieved results =
  1. **tp + fp**
  2. tp + fn
  3. tn + fp
  4. tn + fn

Legend
tp = true positive
tn = true negative
fp = false positive
fn = false negative

Quiz

- All relevant results =
   1. tp + fp
   2. tp + fn
   3. tn + fp
   4. tn + fn

Legend
tp = true positive
tn = true negative
fp = false positive
fn = false negative

Quiz

- All relevant results =
    1. tp + fp
    2. **tp + fn**
    3. tn + fp
    4. tn + fn

Legend
tp = true positive
tn = true negative
fp = false positive
fn = false negative

You have 100% Precision

- Everything you retrieved were relevant.
  - tp + fp = tp
  - fp = 0

You have 100% Recall when

- You retrieved everything that were relevant. (Note: You could have retrieved more).
  - fn = 0
  - tp = all relevant documents

Quiz

- R refers to Relevant Document

- N refers to Nonrelevant Document.

- Collection has 10,000 documents.

- Assume that there are 8 relevant documents in total in the collection. Calculate Precision and Recall.

- Retrieved Documents:

    RRNNN NNN**R**N **R**NNN**R** NNNN**R**

Precision and Recall

- Precision = 6/20
- Recall = 6/8

Precision and Recall

Precision: fraction of retrieved docs that are relevant = P(relevant|retrieved)

Recall: fraction of relevant docs that are retrieved
= P(retrieved|relevant)

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision P = tp/(tp + fp)
- Recall   R = tp/(tp + fn)

Exercise

Suppose, a document is relevant only if both judges agree that it is relevant. Assume (0 = nonrelevant, 1 = relevant). What is the Precision and Recall?

Query = "Taj"

Document ID	Judge 1	Judge 2	Our System
d1 = Bru	0	0	Retrieved
d2 = 3Roses	0	0	No
d3 = Taj	1	1	Retrieved
d4 = Taj Tea	1	1	No
d5 = Taj Mahal	1	0	No

Exercise

Suppose, a document is relevant only if both judges agree that it is relevant. Assume (0 = nonrelevant, 1 = relevant). What is the Precision and Recall?

Query = "Taj"

Document ID	Judge 1	Judge 2	Our System	
d1 = Bru	0	0	Retrieved	False positive
d2 = 3Roses	0	0	No	
d3 = Taj	1	1	Retrieved	True positive
d4 = Taj Tea	1	1	No	
d5 = Taj Mahal	1	0	No	

Exercise

Suppose, a document is relevant only if both judges agree that it is relevant. Assume (0 = nonrelevant, 1 = relevant). What is the Precision and Recall?

Query = "Taj"

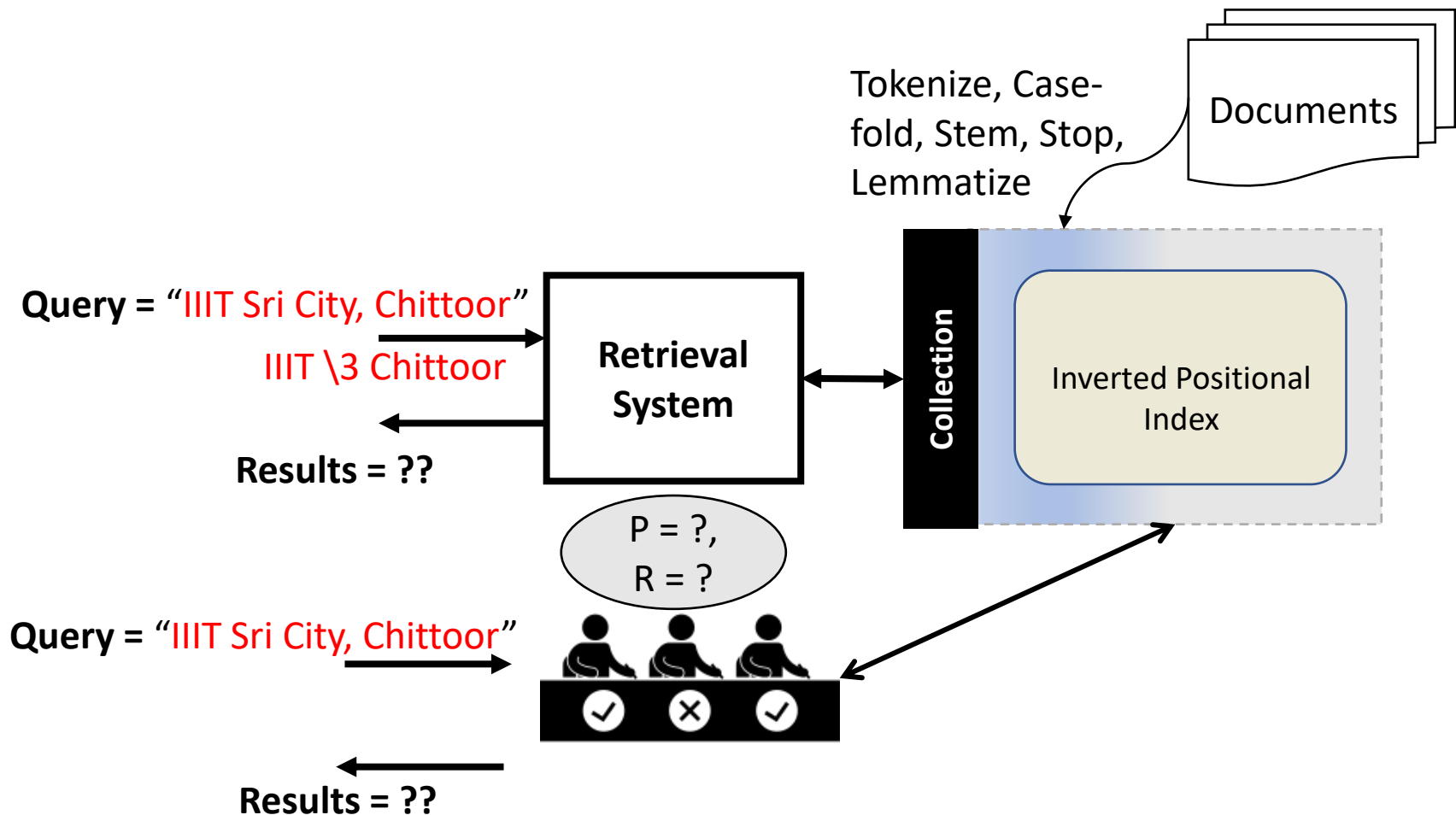| Document ID	Judge 1	Judge 2	Our System	
d1 = Bru	0	0	Retrieved	
d2 = 3Roses	0	0	No	True Negative
d3 = Taj	1	1	Retrieved	
d4 = Taj Tea	1	1	No	False Negative
d5 = Taj Mahal	1	0	No	True Negative

Answer

- Precision = 1/2
- Recall = 1/2

How does a Search Engine Work?

Tokenize, Case-fold, Stem, Stop, Lemmatize

Documents

Query = "IIIT Sri City, Chittoor"

IIIT \3 Chittoor

Retrieval System

Collection

Inverted Positional Index

Results = ??

P = ?,
R = ?

Query = "IIIT Sri City, Chittoor"

Results = ??

Thank You