Information Retrieval (CSD510)

Introduction

Ayan Das

Instructor Contacts

- Instructor: **Ayan Das**

  Email id: *ayandas@iitism.ac.in*
- TAs
  - Raj Kumar Saw
  - Neeraj Singh Dhurvey
- Mode of contact
  - Email or post in Google Classroom.
  - Classroom interaction.
  - Phone calls or Whatsapp messages shall be ignored (Exception: Class representatives).

Class timings

- **Tuesday:** 11:00 AM - 11:50 AM
- **Thursday:** 12:00 AM - 12:50 AM
- **Friday:** 10:00 AM - 10:50 AM
- Class rules:
  - **Institute rule:** 75% attendance is mandatory.
  - **Class rule:** Enter the class within 5 minutes of the commencement of the class.
  - ATTENDANCE MEANS PHYSICAL PRESENCE IN THE CLASS!!

Lecture basics

- Classes will involve both Slides + Board
- For the latest/updated slides, download them before each use.
- Use of laptops and smartphones is not allowed in the classroom.

Evaluation plan

Evaluation plan

- **Quiz 1**: 10 marks
- **Mid semester**: 32 marks
- **Quiz 2**: 10 marks
- **End semester**: 48 marks

NOTE

- There is no provision for quiz retakes or compensatory vivas !!

Course Webpage

- Google classroom link: Information Retrieval (CSD510)

  `https://classroom.google.com/c/NjUwNjMxODY4Nzg2?cjc=sgj7bhr`

- Why do I need to check the webpage?
  - Lecture Notes
  - Misc. static information about the course.
  - Announcements, Quiz schedules, and marks.

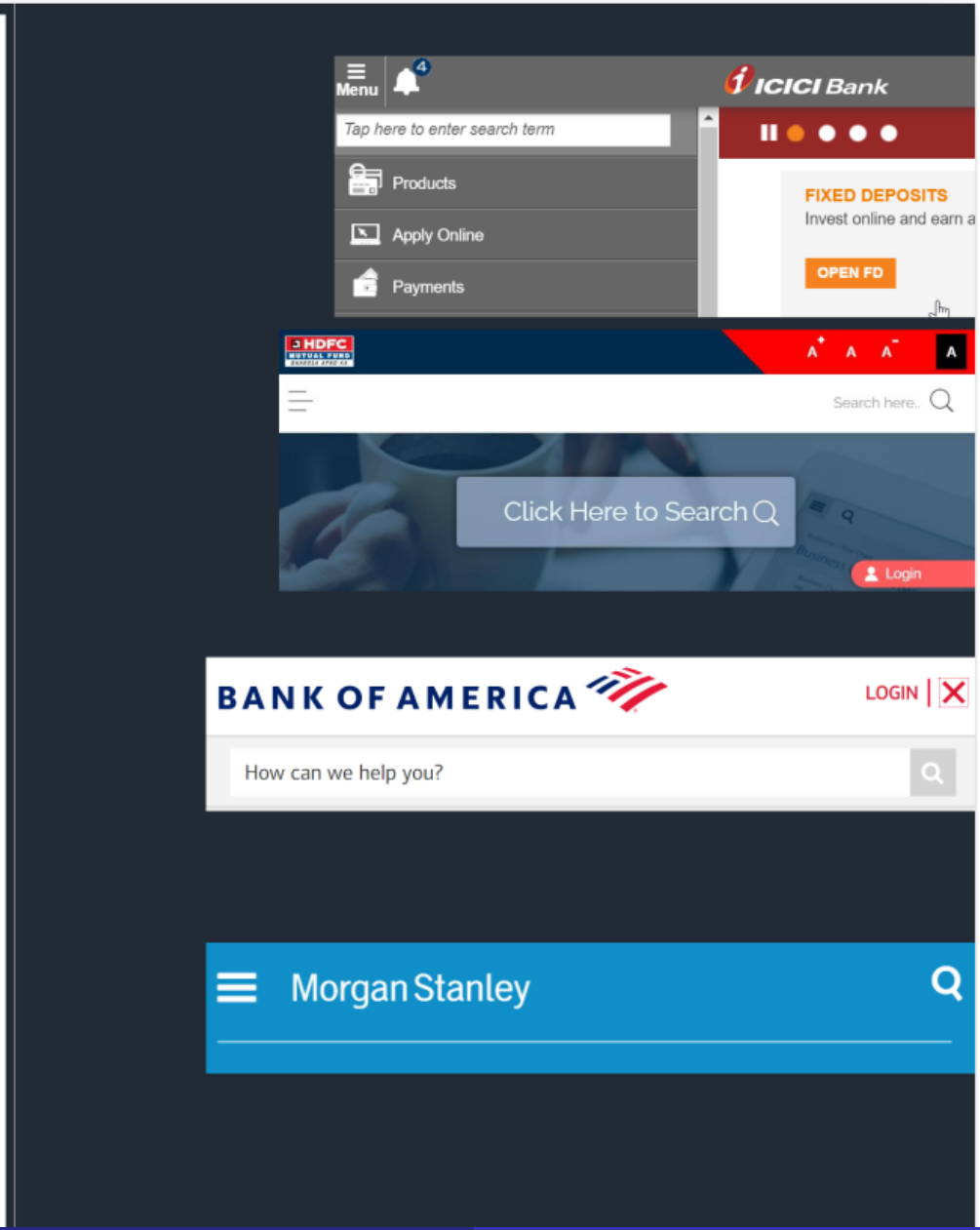
Life without search engines is difficult to imagine!

Search in Banking and Finance

Search in Banking and Finance

Lots of products to sell

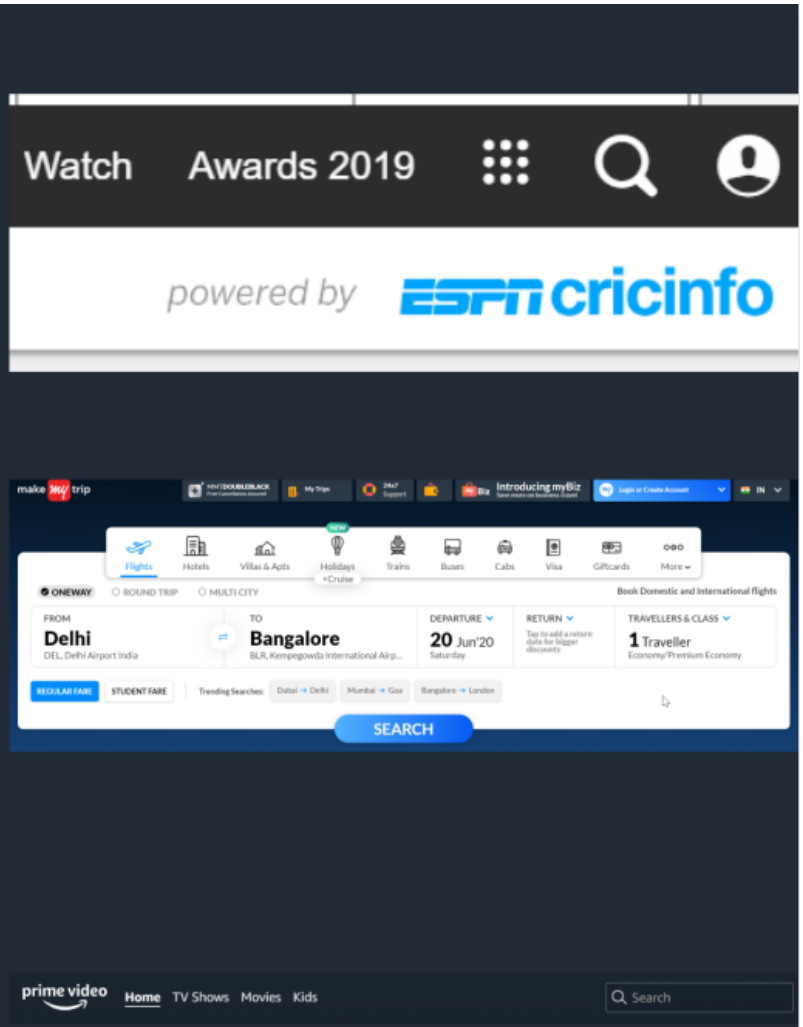
Reach a part of documentation faster



Search in Sports, travel and entertainment

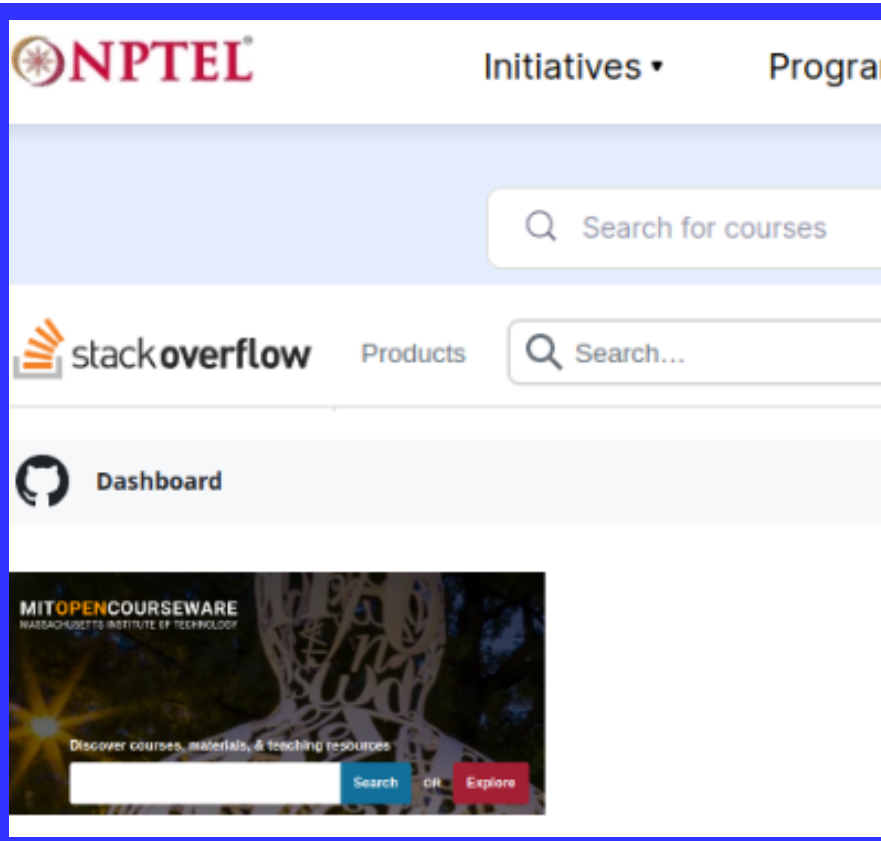
Search in Sports, Travel & Entertainment

Search events, programs, and schedules



Education

Education, coding, and study materials



Why care to learn IR and web search?

Why care to learn IR and web search?

- About 80% of business is conducted on **unstructured** information.
- About 85% of all data stored is held in an **unstructured** format.
- On an average, roughly **7 million web pages** are added everyday.
- **Unstructured** data doubles roughly every three months.

IR as research discipline

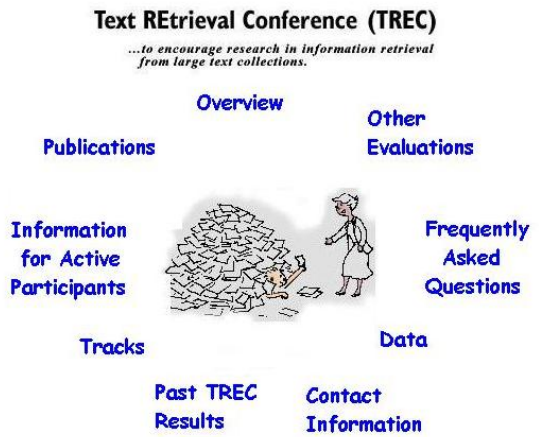
ACM's **SIGIR**

- Special Interest Group on Information Retrieval.
- Annual conferences, beginning in 1978.
- Awards the **Gerard Salton** award.

SIGIR
Special Interest Group on Information Retrieval

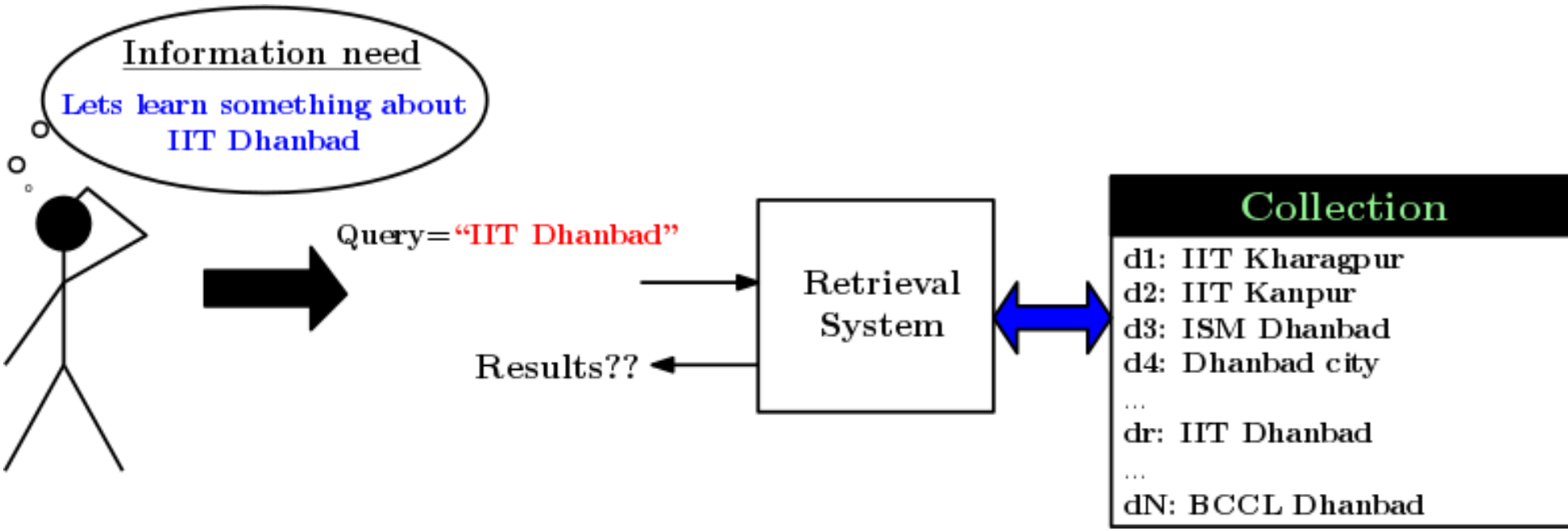
TREC

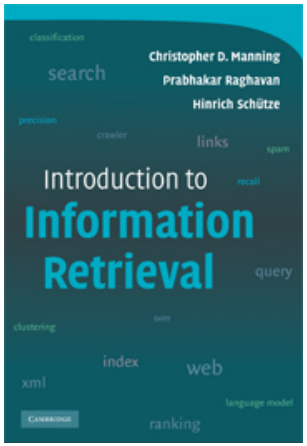
- Annual text retrieval conference, beginning in 1992.
- Sponsored by the **US National Institute of Standards and Technology** as well as **US Department of Defense**.
- Conducts different tracks, e.g. blogs, genomics, spam
- Provides **data sets** and **test problems**.



- CLEF, NTCIR and FIRE are some other major IR conferences.

Information retrieval





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

Core problems of IR

- How to store and update **large** document collections?
  - **Small !!**
  - **Scalable !**
- How to do **efficient** retrieval?
  - **Speed !**
- How to do **effective** retrieval?
  - **Ensure high result quality!**

Document vs. Database Records

Document

- A document is a collection of free text records written in some natural language.
- Web pages, emails, books, news stories, scholarly papers, text messages, Powerpoint, PDF, forum postings, patents, tweets, question-answer postings, blogs, etc.

Database records

- Database records (or tuples in relational databases) are typically made up of well-defined fields (or attributes),
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics to queries in order to find matches.

Document vs. Database Records

Example bank database query

- Find records with balance $>$ \$50,000 in branches located in Amherst, MA.
- Matches easily found by comparison with field values of records

Document vs. Database Records

Example bank database query

- Find records with balance $>$ \$50,000 in branches located in Amherst, MA.
- Matches easily found by comparison with field values of records

Example search engine query

- *financial scams since 2019 in India*

Document vs. Database Records

Example bank database query

- Find records with balance $>$ \$50,000 in branches located in Amherst, MA.
- Matches easily found by comparison with field values of records

Example search engine query

- *financial scams since 2019 in India*
- This text must be compared to the text of entire news stories

Typical IR tasks

Given

- A corpus of textual natural-language documents.
- A user query in the form of a textual string.

Find

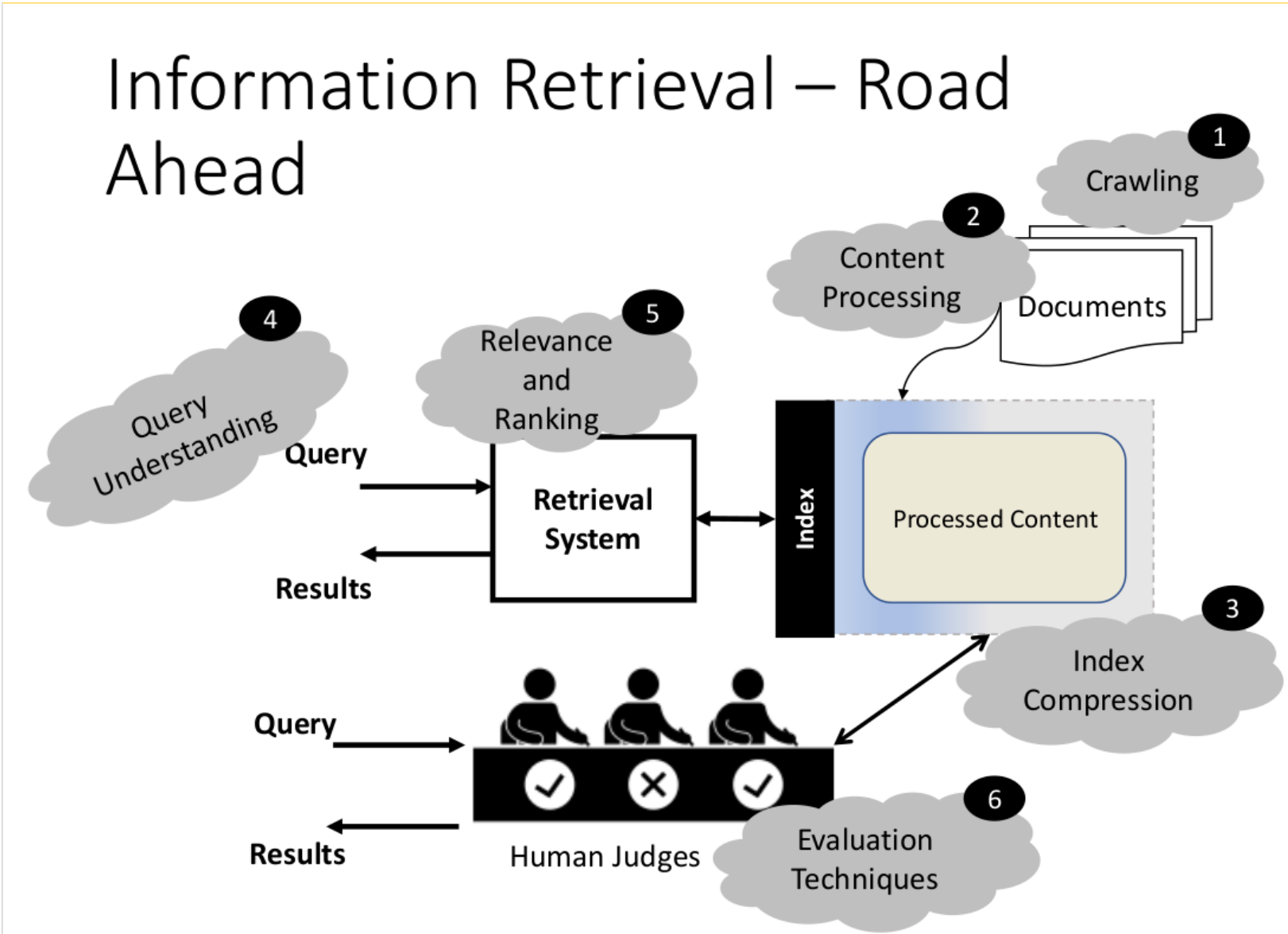
- A ranked set of documents that are relevant to the query.

So, what is relevance?

The **relevant document** contains the information that a person was looking for when they submitted the query. This may include:

- Being on the proper subject.
- Being timely (recent information).
- Being authoritative (from a trusted source).
- Satisfying the goals of the user and his/her intended use of the information (information need).

What do we do in IR??

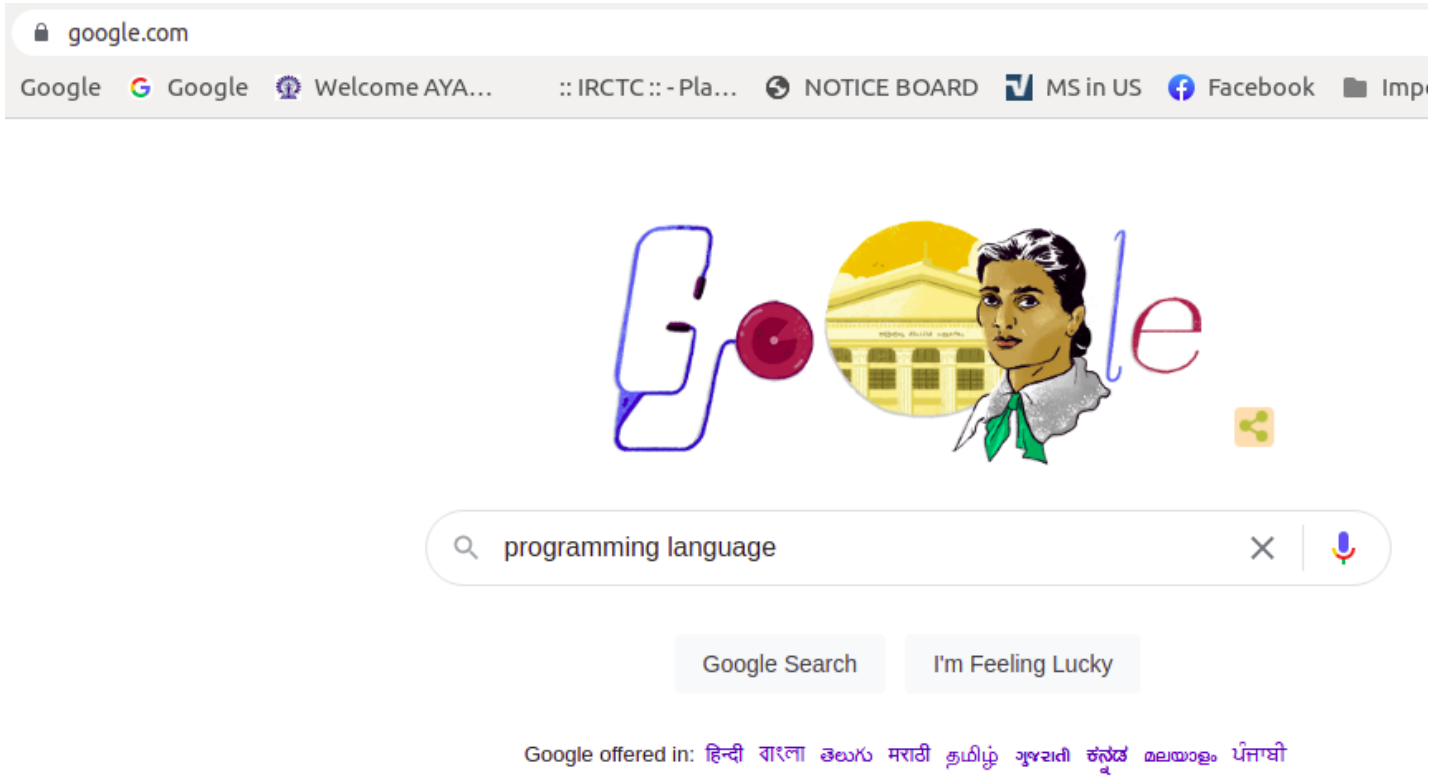


Information need

- An **information need** is the topic about which the user wants to know more.
- Refers to an individual, hidden **cognitive state.**
- Depends on what the user knows and doesn't know.
- **Ill-defined**
  - What is the capital of USA?
  - Is it really true that addictive substances are mixed in soft drinks?
  - What is "cloud computing"?

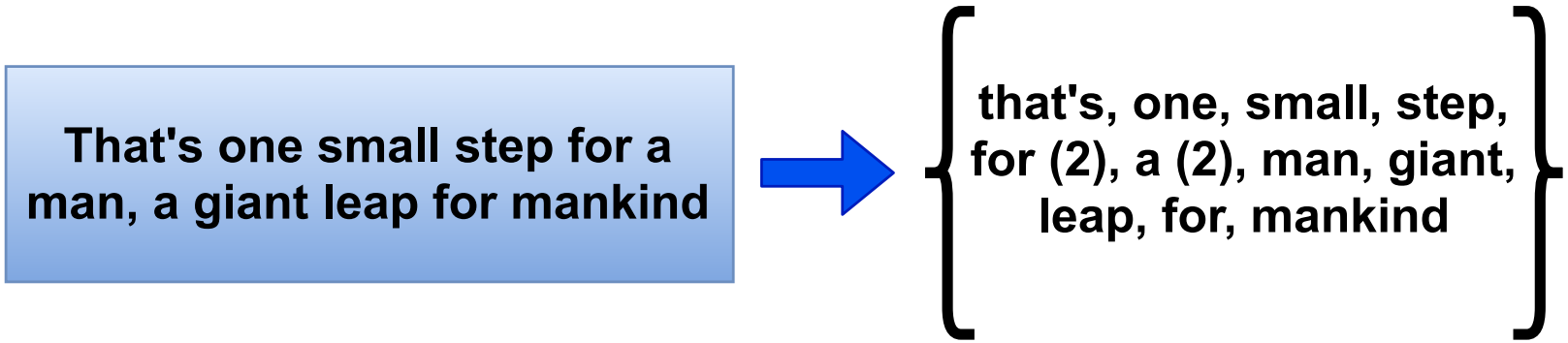
Query

- A **query** is what the user conveys to the IR system to **communicate the information need**.
- Stated using a
  - usually a list of search terms.
  - some formal query structure.



Logical view of document

- **Bag-of-words model:** Document usually treated as a **multi-set** of index terms or keywords derived from a predefined vocabulary.
  - **Index term** is a term that captures the essence of the topic of a document.
  - Keywords extracted from a document.
  - Keywords are derived automatically or generated by a specialist.
- **Text operations:** Operations involved in converting a document to a *bag of words*.
  - reduces the complexity of the document representation.
  - allows moving the logical view from that of full text to that of a set of index terms.



Bag-of-words model

Vocabulary (Index terms)	that's	one	small	step	for	a	man	giant	leap	mankind	Abdul	Kalam's	is	India
That's one small step for a man, a giant leap for mankind	1	1	1	1	2	2	1	1	1	1	0	0	0	0
Abdul Kalam's small step is a giant leap for India	0	0	1	1	0	1	0	1	1	0	1	1	1	1

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
D1	1	1	1	1	2	2	1	1	1	1	0	0	0	0
D2	0	0	1	1	0	1	0	1	1	0	1	1	1	1

The bag-of-words model

- **Pros**
  - Simple set-theoretic representation of documents.
  - Efficient storage and retrieval of individual terms.
  - IR models using the bag-of-words representation have been found to perform reasonably well.
- **Cons**
  - Word order not maintained
  - Very different documents could have similar representation.
    - "advantages of C over Java"
      AND
    - "advantages of Java over C"
  - Document structure information or metadata is ignored.

Logical view of a document

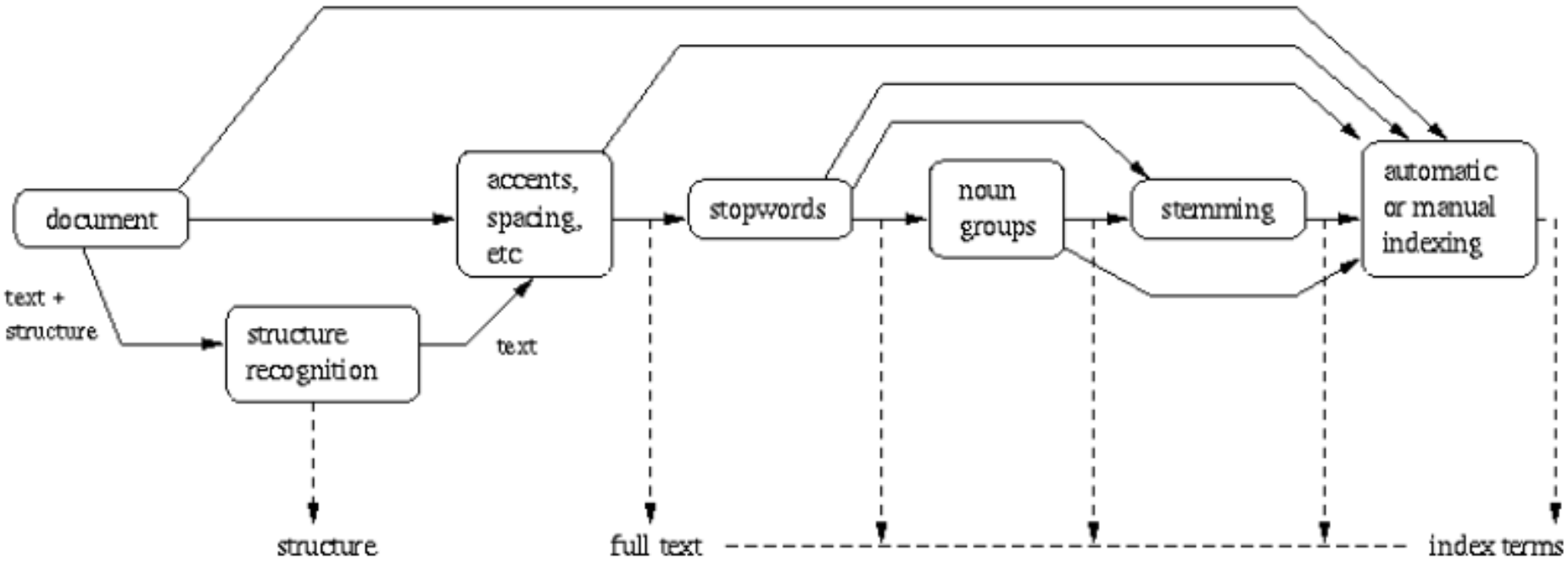


Figure: Logical view of the document: from full text to a set of index terms

Logical view of a document (contd.)

Stop-word removal

- **Word categories**
  - Content words: Nouns, verbs, adjectives, adverbs.
  - Function words: Other parts-of-speeches.
- **Stop-words**
  - Function words do not bear useful information for IR.
    - of, in, about, with, I, although
  - Reduce the set of representative keywords from large collection.
  - The removal of stop-words usually **improves** IR effectiveness.
- **Stop-lists**
- PoS tagging is usually not an integral component of an IR system.
- **Stop-lists**: Lists of stop-words consisting of function words and very frequent words not to be indexed.

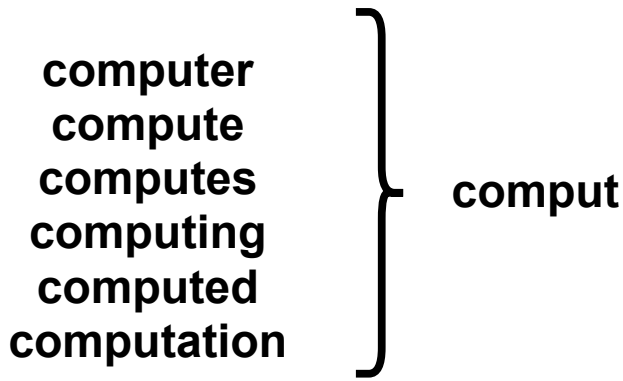
Logical view of a document (contd.)

Noun groups

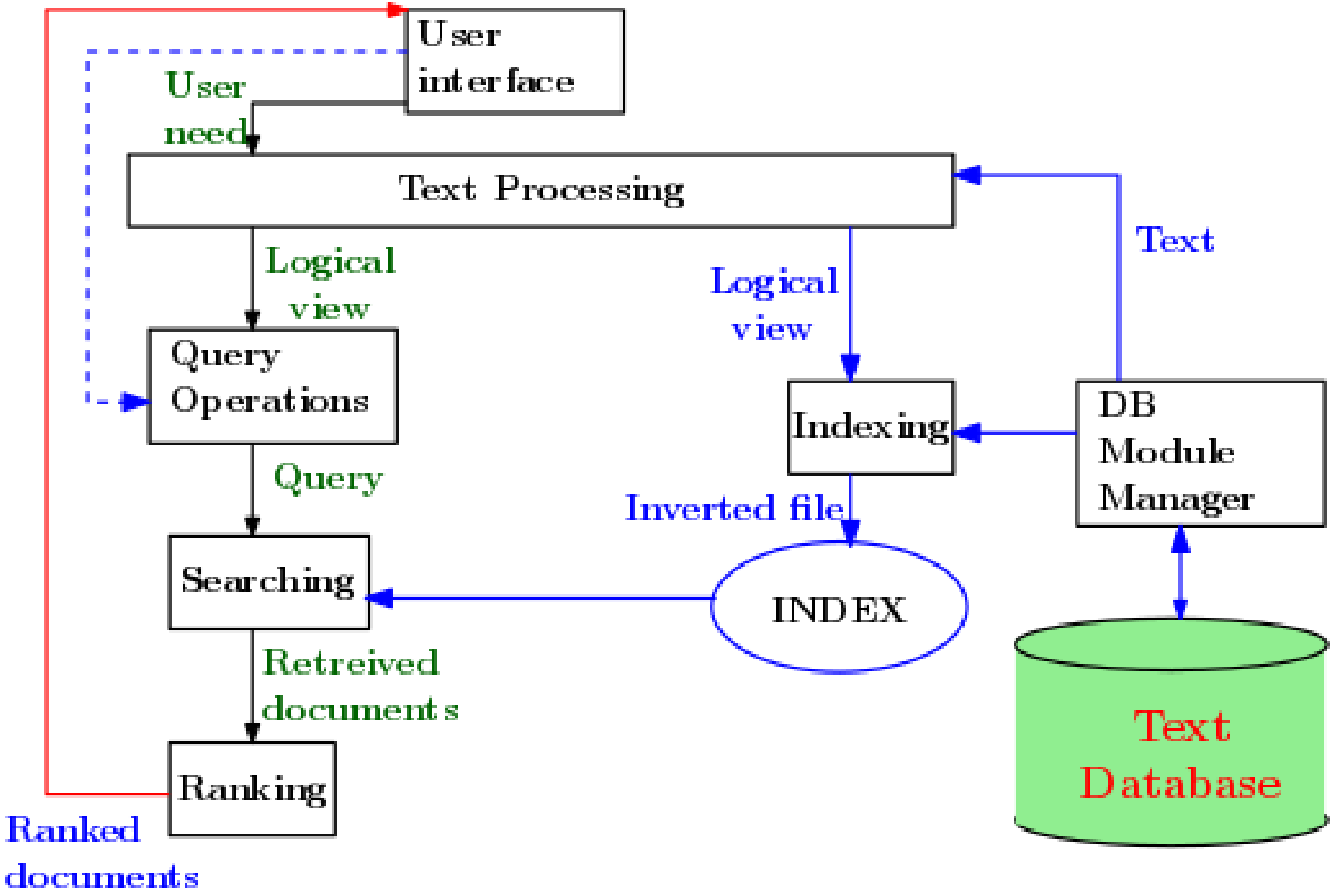
- Word retention module.
- Required when only NOUNs are needed by the retrieval system.
- To identify the noun groups - gazetteer list, list of nouns updated constantly.
- Which eliminates the adjectives, adverbs and verbs.

Stemming

- A **root word** may take different **word forms** based on their usage in a context.
- **Stemming** used to *normalize* the different word forms to a standard form.



Retrieval process



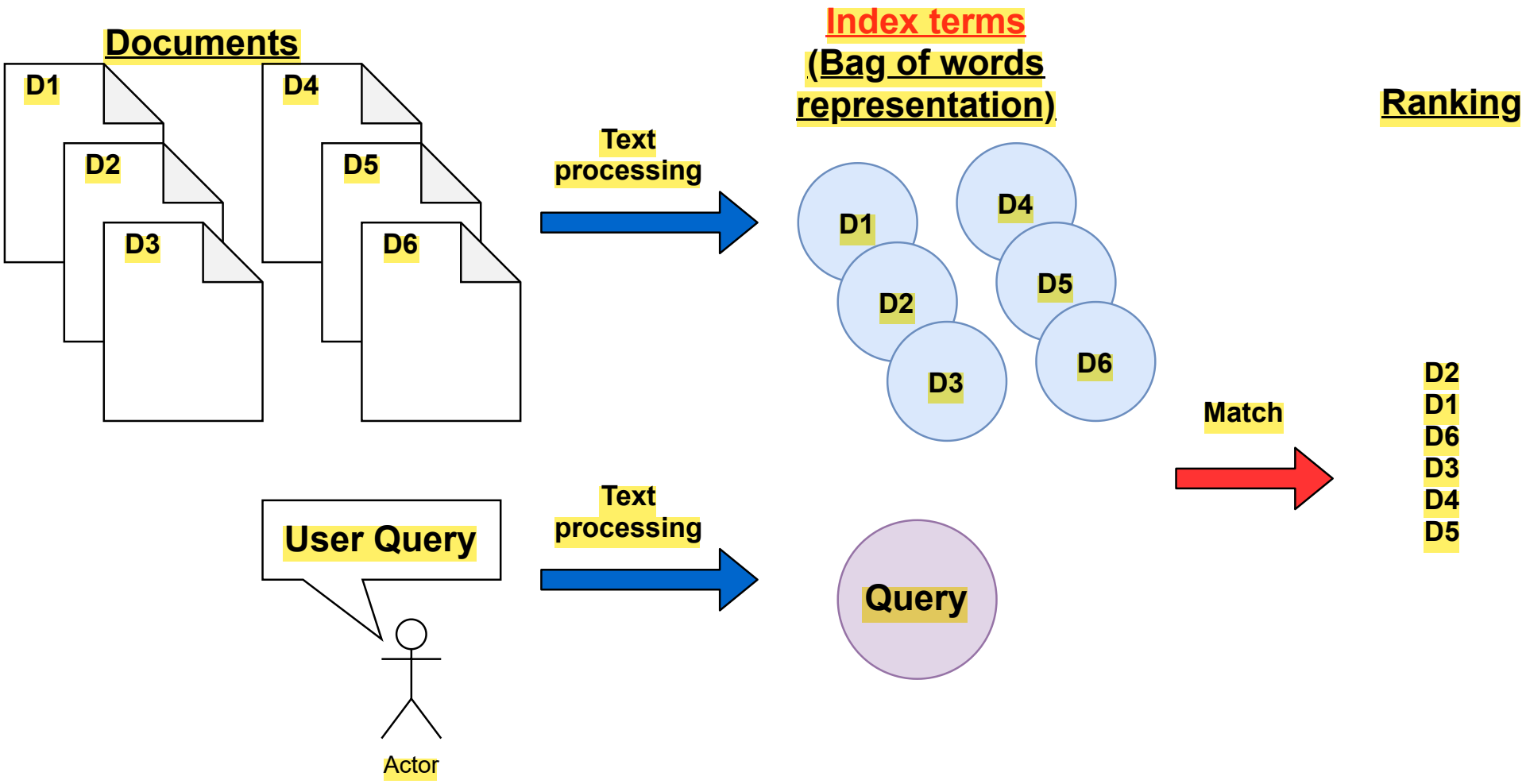
The retrieval process

- The RP can be initiated, it is necessary to define the text DB.
- This is done the DB manager, which specifies the following,
  - The documents to be used
  - The operations to be performed on the text
  - The text model, i.e. the text structure and what elements can be retrieval.
- Text operations transform the original documents and generate a logical view of them.
- The database manager builds an index of the text i.e. "inverted file",
- Query operations used to generate actual "query" based on the used needs To retrieve the relevant document for processing the query
- The retrieved document ranked, before sent to the user

The retrieval process (contd.)

- Text Operations forms index words (tokens).
  - Stop-word removal
  - Stemming
- Indexing constructs an inverted index of word to document pointers.
- Searching retrieves documents that contain a given query token from the inverted index.
- Ranking scores all retrieved documents according to a relevance metric.
- User Interface manages interaction with the user:
  - Query input and document output.
  - Relevance feedback.
  - Visualization of results.
- Query Operations transform the query to improve retrieval:
  - Query expansion
  - Query transformation using relevance feedback.

Modelling


Documents
D1
D2
D3
D4
D5
D6
Text processing
Index terms (Bag of words representation)
D1
D2
D3
D4
D5
D6
Match
Ranking
D2
D1
D6
D3
D4
D5
User Query
Actor
Text processing
Query

Modelling

- IR systems usually adopt index terms to process queries.
- Index term:
  - a keyword or group of selected words
  - any word (more general)
- Stemming might be used
  - connect: connecting, connection, connections
- An inverted file is built for the chosen index terms.
- A *ranking* is an ordering of the documents retrieved to the user query.
- A ranking is based on fundamental premises regarding the notion of relevance, such as:
  - common sets of index terms
  - sharing of weighted terms
  - likelihood of relevance
- Each set of premises leads to a distinct *IR model*.

Simplest notion of Relevance from Retrieval Models' Perspective

Keyword Search

- Simplest notion of relevance is that the query string appears verbatim in the document.
- Slightly less strict notion is that (most of) the words in the query appear frequently in the document, in any order (*bag of words*).

Problems with Keywords Search

Term mismatch

May not retrieve relevant documents that include synonymous terms

- PRC vs. China
- car vs. automobile

Ambiguity

May retrieve irrelevant documents that include ambiguous terms (due to polysemy)

- 'Apple' (company vs. fruit)
- 'Java' (programming language vs. Island)
- 'Python' (programming language vs. Snake)

Topics to be covered in the course

1. Boolean retrieval
2. The term vocabulary & postings lists
3. Dictionaries and tolerant retrieval
4. Index construction and compression
5. Scoring, term weighting & the vector space model
6. Computing scores in a complete search system
7. Evaluation in information retrieval.
8. Relevance feedback & query expansion
9. Probabilistic information retrieval
10. Language models for information retrieval
11. Text classification.
12. Link analysis – HITS, PageRank
13. Learning to Rank
14. Neural IR - Word embeddings, Semantic Matching - DSSM

Books

- Textbooks
  1. **Introduction to Information Retrieval** - Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze: *Cambridge University Press.*
- Reference books
  1. **Mining of Massive Datasets** - Jure Leskovec, Anand Rajaraman, Jeff Ullman: *Cambridge University Press.*
  2. **Mining the Web: Discovering Knowledge from Hypertext Data** - Soumen Chakrabarti: *Morgan Kaufmann Series in Data Management Systems*
  3. **An Introduction to Neural Information Retrieval** - Bhaskar Mitra, Nick Craswell: *NOW publishers*
  4. other materials (if required) shall be made available in the Google classroom....

Technologies & Frameworks

Apache

Apache

Apache

Univ. of Glasgow

Galago, Indri

UMass & CMU

There are many more....