Information Retrieval (CS60092)

Course Information

(1) Marks distribution:

- Midsem (Absolute 20/100)
- Endsem (Absolute 40/100)
- Assignments (You develop parts of an IR system 20/100)
- Attendance (20/100). → Random days → aggregate to what you get on 20.

Course Materials

No fixed book.

First part will be predominantly from Manning's book.

Course Website ⟵ Single point reference to all resources (to be sent via email to you next week once the registration stabilizes)

Teaching assistants ←

Siddharth Jaiswal.

Sasthak Roy.

Course Content (Roughly)

AM
- Boolean retrieval
- Postings list & term vocab
- Skip pointers → Assignment I
- Tolerant retrieval.

NA
- Vector space models
- IR eval
- Relevance feedback
- Language Modeling.
- Link analysis → Assignment II
- Summarization → Assignment III

AM
- Open domain QA + Dense retrieval.
- Have to think.
- .....

Do I need to know ML 101?

Very little (towards the end) → Before summarization

What is Information Retrieval?

→ Finding material? ≡ documents (web documents).

↓

of an unstructured nature. (text).

↓

satisfies "information need" (query)

↓

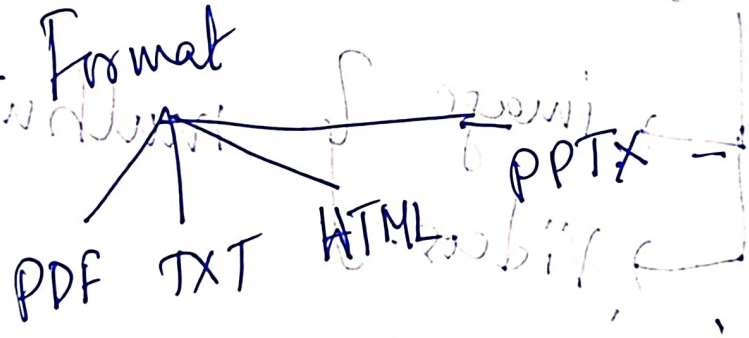
from within large collections.

Corpus of documents

What are potential documents?

- legal documents
- research papers
- Webpage (news papers, websites of orgs, people, place, travel logs ...)
- Emails / books.

- Format



Where lies the challenge in IR?

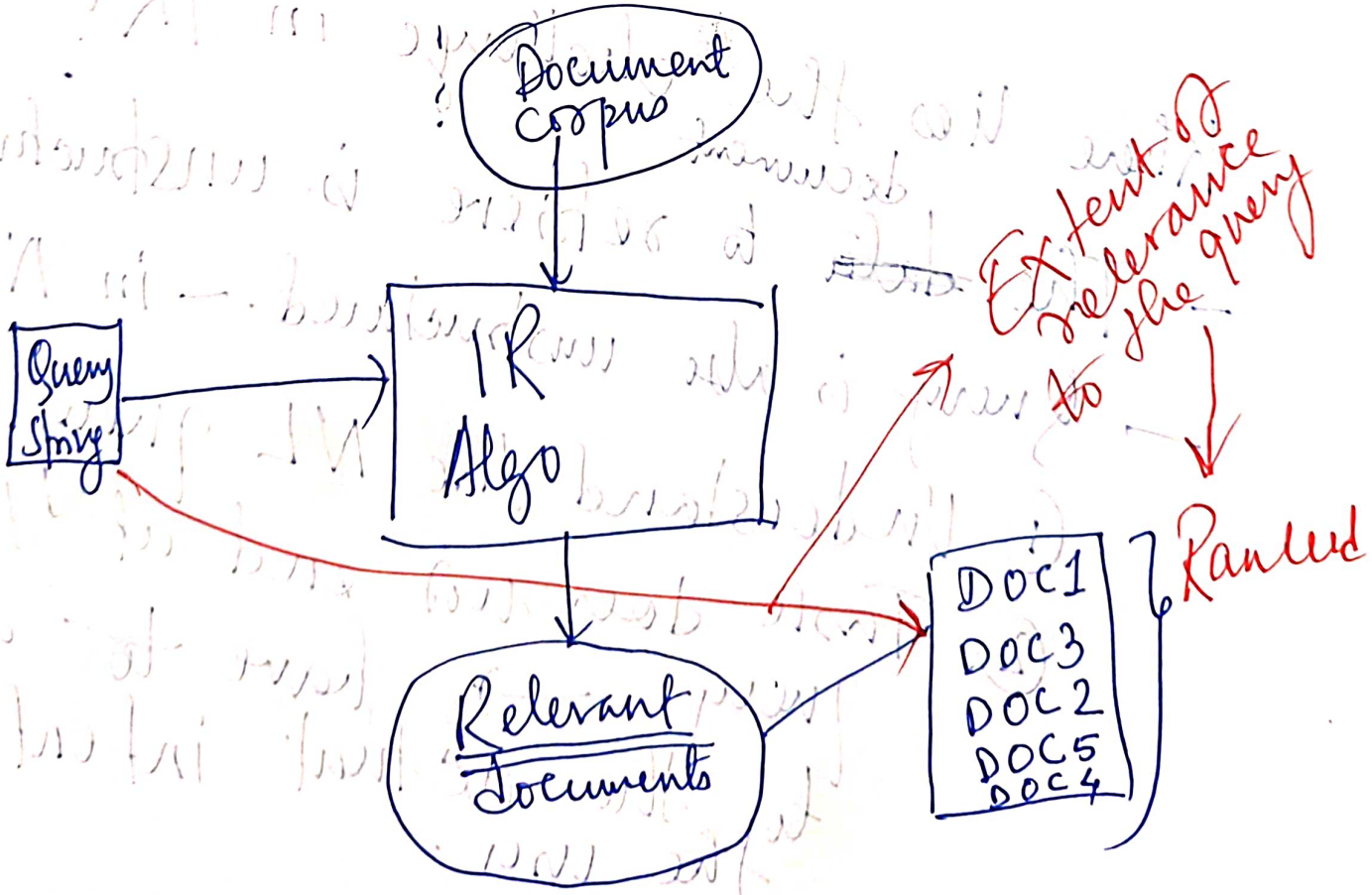
- The ~~data~~ document to retrieve is unstructured
- Query is also unstructured — in NL.
  1. Understand the NL query
  2. Task does not end at the query → have to get to the actual intent of the user.

Query: What is a "killer" app?

↳ Confusing semantics!!

Retrieve only text?

↳ image, videos → multimedia.



What do we mean by relevance?

- The response should be on the proper subject (topic).
- Timely / context
- Responses should be "authoritative".
- Relevant to the "information need" or the "intent" of the user.

Most Vanilla approach of doing IR?

Keyword search
↳ string matching
(match the query & the document string).

Pitfalls / Caveats of Keyword search?

Synonymy: ~~[scribbled out]~~ PRC & China : are probably same.

Car vs. automobile . device?

Polysemy: Apple? ⟨ company ⟨ device? / Organisation? ; fruit

Java?

→ IR → look into the intended meaning of word

↳ looking around the word (context)

" A word ~~[crossed out]~~ is known by the company it keeps" → Distributional hypothesis

Query: Colourless green ideas sleep furiously ← Noam Chomsky

Response: ? intended meaning of the sentence is "Nonsense"

A* → SIGIR (~5000) → (17% - 19%)
↳ Special Interest Group of ACM.

WWW (The Web Conference)
↳ Tim Berners Lee et al.
founder of the WWW foundation

Information extraction
Search
Retrieval & Ranking

Summarization

Legal IR

Fair IR