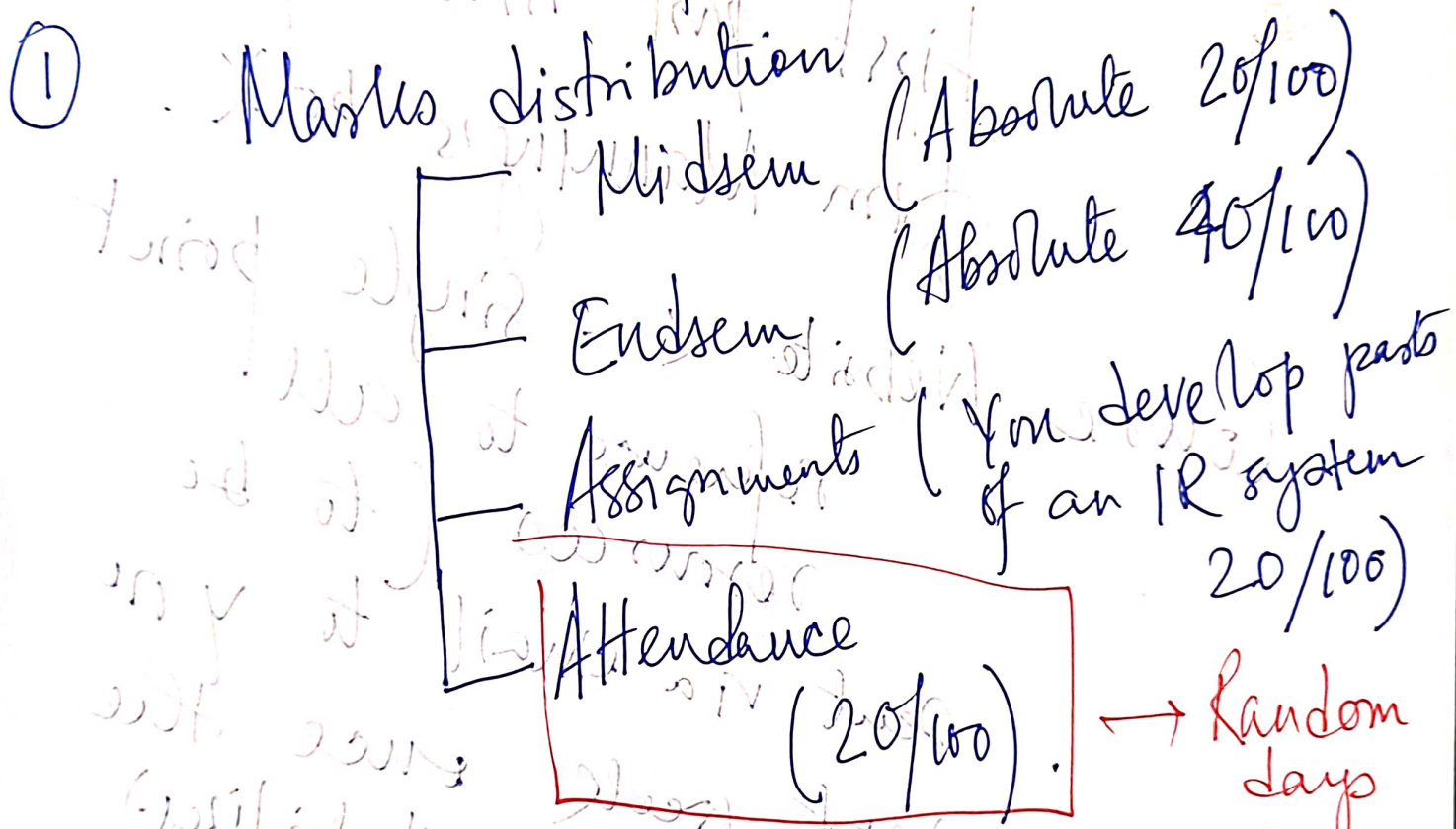


Information Retrieval (CS60092).

Course Information



→ 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 Jainmal.

Sarthak Roy.

Course Content (Roughly)

AM

Boolean retrieval

Posting list & term vocab

Skip pointers

Tolerant retrieval.

→ Assignment I

NH

Vector space models

Relevance feedback.
Language modeling.

Link analysis

→ Assignment II

Summarization

→ Assignment III

Open domain QA + dense retrieval.
Have to think.

AM

Do I need to know ML 101?
Very little (towards the ^{Before.} end) → summarization

What is Information Retrieval?

→ Finding material? (= documents (web documents)).

↓
of an unstructured nature.
(text).

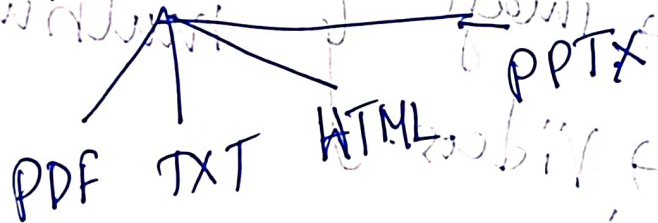
↓
satisfies "information need"
(query)

↓
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 ----)
- Email / books.

Format



Where lies the challenge in IR?

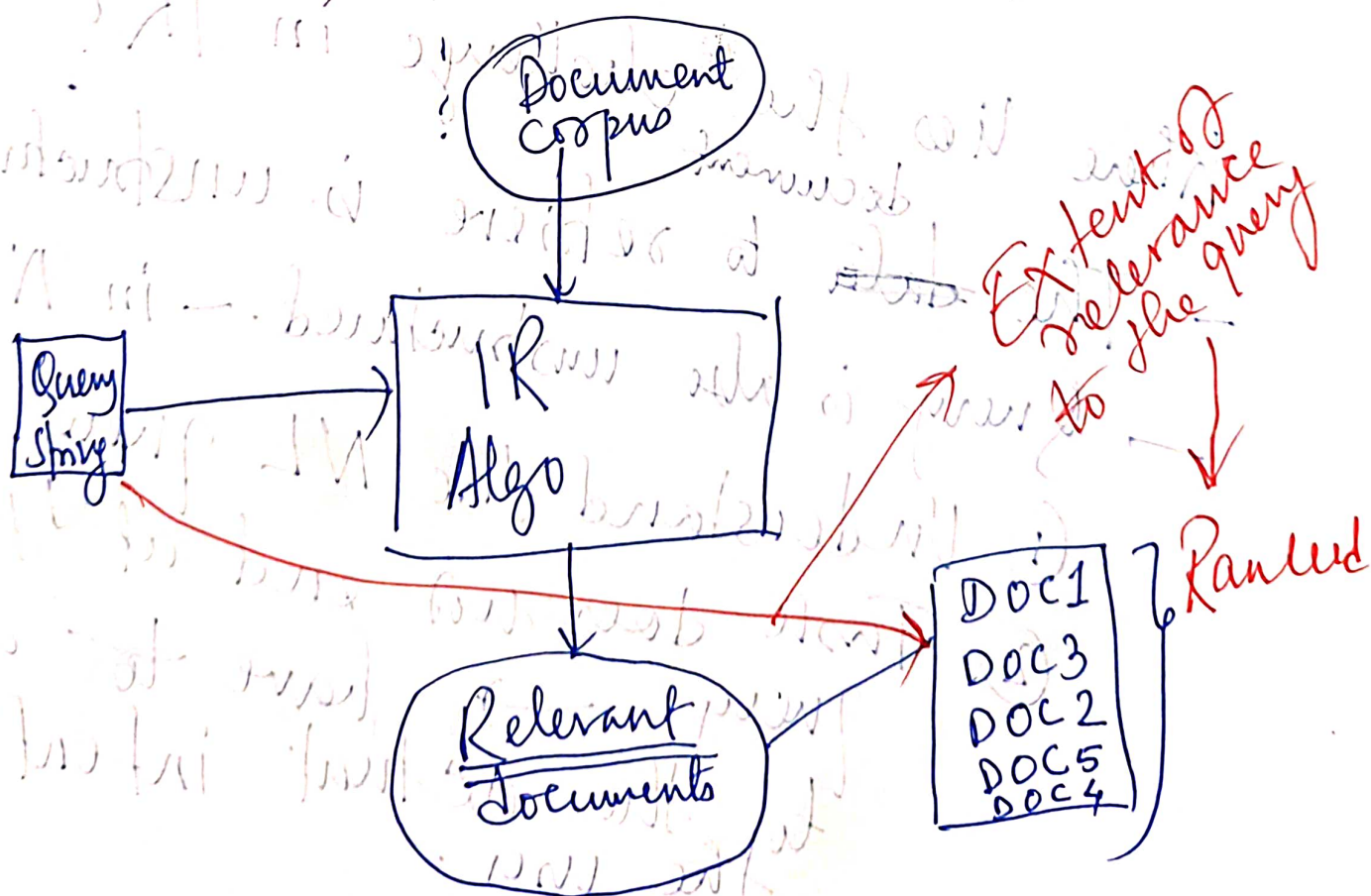
- The ~~data~~ document to retrieve is unstructured
- Query is also unstructured. - in NL.
- ① Understand the NL query
- ② Task does not end at the query → have to get to the actual intent of the user.

Query: What is a "killer" app?

↳ Confusing semantics!!

Retrieve only text? - word class -

image } multimedia.
Videos }



What do we mean by relevance?

The response should be on the proper subject (topic).

Timely / context

Responses should be "authoritative".

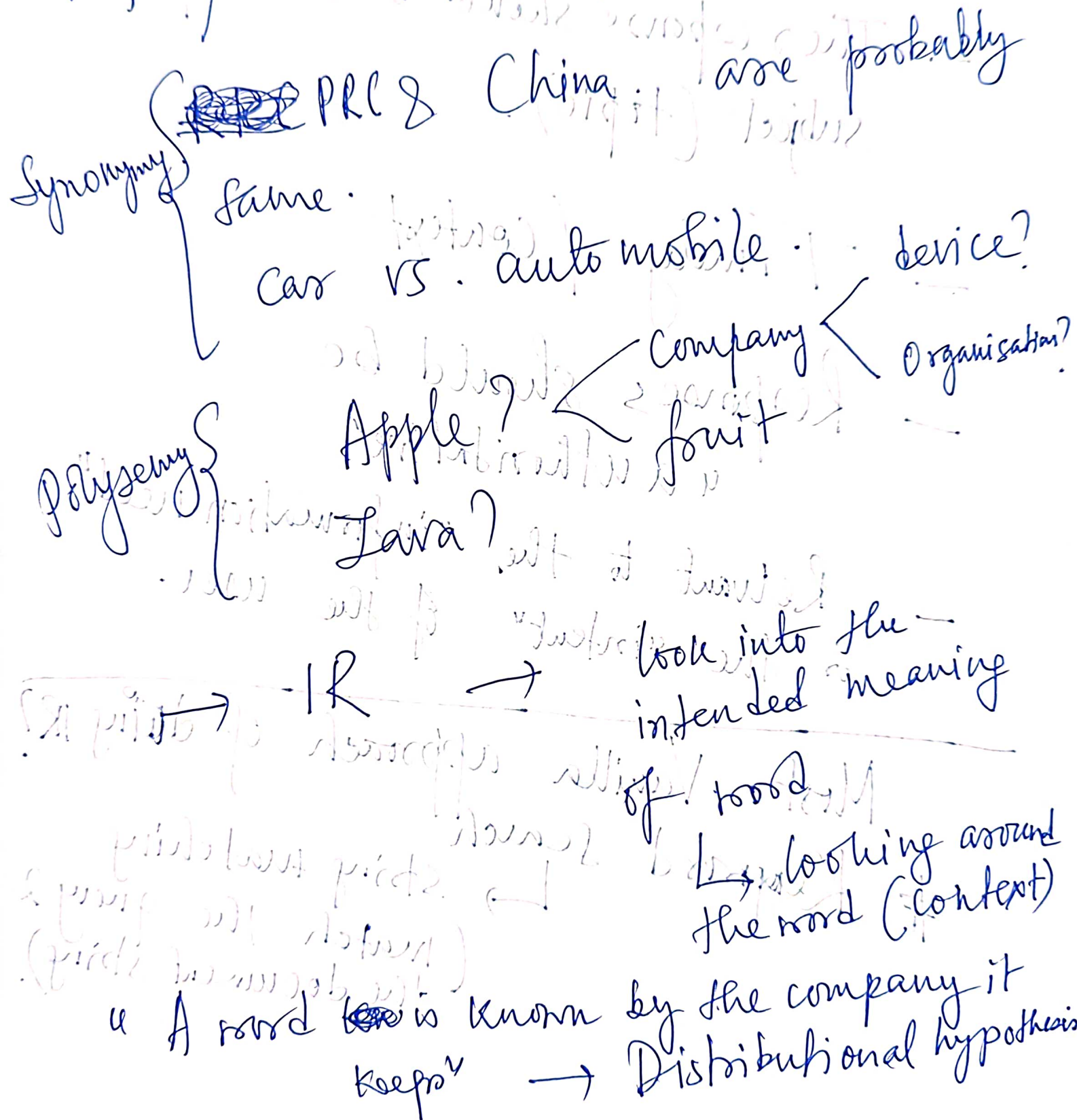
Relevant to the "information need" or the "intent" of the user.

Most Vanilla approach of doing IR?

Keyword search

→ string matching
(match the query & the document string).

Pitfalls/Caveats of Keyword Search?



Query: Colourless green ideas
sleep furiously ← Noam Chomsky

Response: ? intended meaning
of the sentence is
"Non sense"

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