INFORMATION RETRIEVAL

- IR is concerned w/ the organization & retrieval of info from a large no. of text-based documents.
- Due to the abundance of text info, there are many app's of IR, like — online library catalog systems, online document management systems & more.
- A typical IR problem is to locate relevant documents in a document collection based on a user's query, which often consists of keywords describing the info needed or an example relevant doc.
- In such a search problem, the user takes the initiative to "pull" the relevant info out from the collection — this is apt when the user has some short-term info need.
- When a user has a long-term info need, a retrieval system may take the initiative to "push" any newly arrived info to the user. Such systems are called recommender systems.

TEXT RETRIEVAL METHODS

- They fall into 2 categories — they either view the retrieval problem as document selection or document ranking.

- **DOCUMENT SELECTION:**
  - ↳ The query is regarded as specifying constraints for selecting relevant documents.
  - ↳ A typical method of this type is — Boolean Retrieval model — a document is represented by a set of keywords & a user provides a Boolean expression of keywords, like, "tea or coffee", "blue and Tshirt", etc.
  - ↳ The retrieval system would take the boolean query & return a set of relevant docs.
  - ↳ The Boolean retrieval method only works well when the user knows a lot about the document collection & can formulate a good query.

- **DOCUMENT RANKING:**
  - ↳ These methods use the query to rank all documents in the order of relevance.
  - ↳ These methods are more apt for ordinary users & exploratory queries than document selection.
  - ↳ There are many different ranking methods based on a large spectrum of mathematical foundations.
  - ↳ The common intuition blw all these methods is that we match the keywords in a query wl those in the doc & score each doc based on how well it matches the query.
  - ↳ The goal is to approximate the degree of relevance wl a score computed based on info such as freq. of words in the doc & the whole collection.