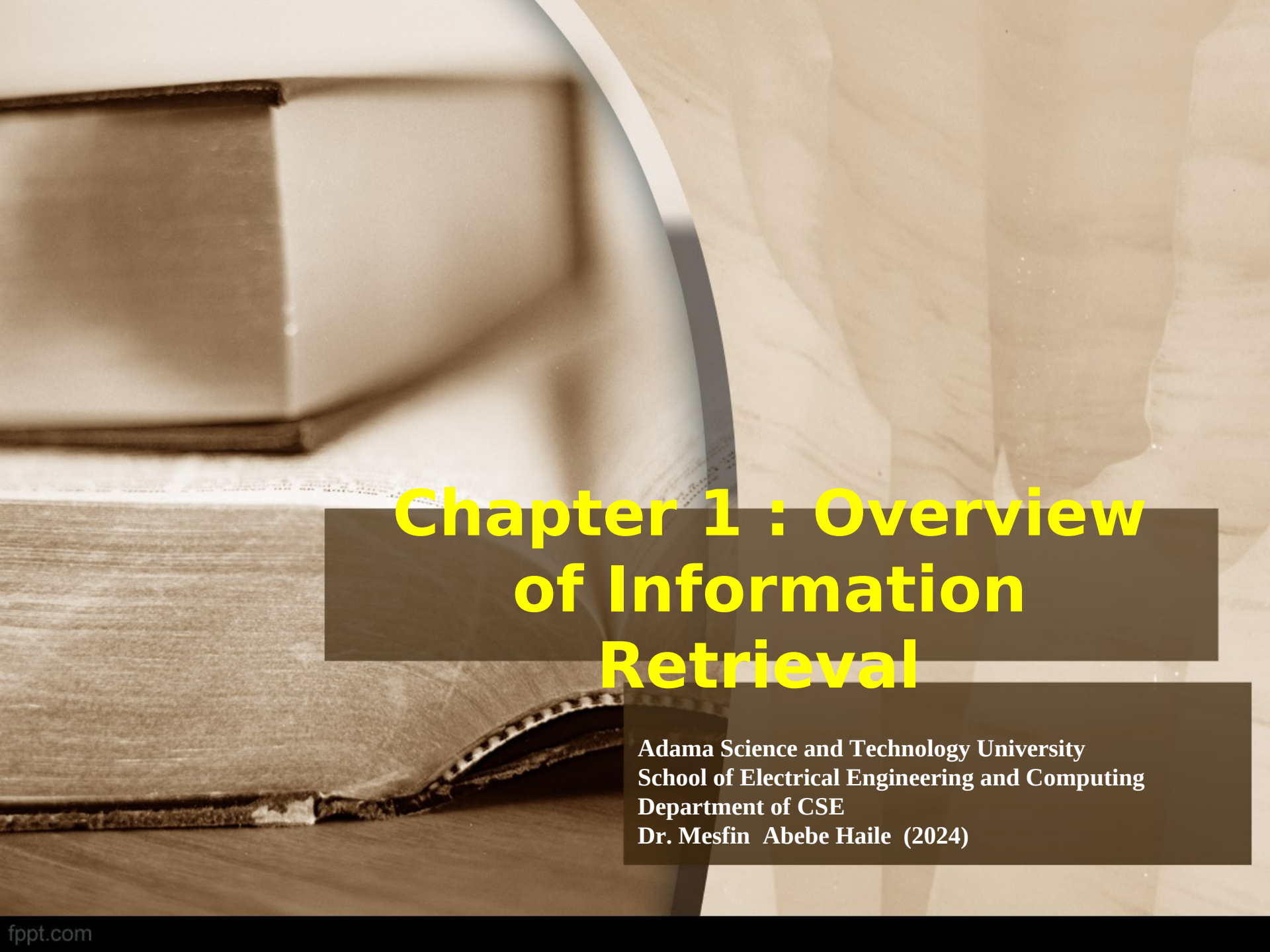




Chapter 1 : Overview of Information Retrieval

**Adama Science and Technology University
School of Electrical Engineering and Computing
Department of CSE
Dr. Mesfin Abebe Haile (2024)**

Information Retrieval

- ❖ Information retrieval (IR) is the process of finding *relevant documents* that satisfies *information need* of users from *large collections of unstructured text*.
- ❖ General Goal of *Information Retrieval*:
 - ❖ To *help users to find useful information* based on their *information needs* (with a minimum effort) despite:
 - ❖ *Increasing complexity* of Information,
 - ❖ *Changing needs* of user,
 - ❖ Provide *immediate random access* to the document collection.

Document Corpus

- ❖ *Large collections of documents* from *various sources*: news articles, research papers, books, digital libraries, Web pages, etc.

Sample *Statistics of Text Collections*:

- ❖ Dialog: (<http://www.dialog.com/>)
- ❖ Claims to have *more than 20 terabytes of data in* > 600 Databases, > 1 Billion unique records.
- ❖ LEXIS/NEXIS: (<http://www.lexisnexis.com/>)
- ❖ Claims *7 terabytes, 1.7 billion documents, 1.5 million subscribers*, 11,400 databases; > 200,000 searches per day; 9 mainframes, 300 Unix servers, 200 NT servers.

Document Corpus

- ❖ Large collections of documents from *various sources*: news articles, research papers, books, digital libraries, Web pages, etc.
- ❖ TREC (Text REtrieval Conference) collections:
 - ❖ It is *an annual information retrieval conference & competition*.
 - ❖ Total of about 10 GBs of *dataset of text for IR evaluation*.
- ❖ Web Search Engines:
 - ❖ *Google claim* to index *over 3 billion pages*.

Information Retrieval Systems ?

- ❖ Document (Web page) retrieval in *response to a query*.
 - ❖ Quite effective (at some things)
 - ❖ Commercially successful (some of them)
- ❖ But what goes on *behind the scenes*?
 - ❖ How do *they work*?
 - ❖ What *happens beyond the Web*?
- ❖ Web *search systems*:
 - ❖ Lycos, Excite, Yahoo, Google, Live, Northern Light, Teoma, HotBot, Baidu, ...



Web Search Engines

- ❖ There are *more than 2,000 general web search engines*.
- ❖ The big four are *Google, Yahoo!, Live Search, Ask*.
 - ❖ *Scientific research & selected journals* search engine: Scirus, About.
 - ❖ *Meta search engine*: Search.com, Searchhippo, Searchthe.net, Windseek, Web-search, Webcrawler, Mamma, Ixquick, AllPlus, Fazzle, Jux2
 - ❖ *Multimedia search* engine: Blinkx_
 - ❖ Visual search engine: Ujiko, Web Brain, RedZee, Kartoo, Mooter
 - ❖ Audio/sound search engine: Feedster, Findsounds
 - ❖ video search engine: YouTube, Trooker
 - *Medical search engine*: Search Medica, Healia, Omnimedicalsearch,

Web Search Engines

- ❖ There are *more than 2,000 general web search engines*.
- ❖ The big four are *Google, Yahoo!, Live Search, Ask*.
 - ❖ *Index/Directory*: Sunsteam, Supercrawler, Thunderstone, Thenet1, Webworldindex, Smartlinks, Whatusee, Re-request, DMOZ, Searchtheweb
 - ❖ *Others*: Lycos, Excite, Altavista, AOL Search, Intute, Accoona, Jayde, Hotbot, InfoMine, Slider, Selectsurf, Questfinder, Kazazz, Answers, Factbites, Alltheweb
- ❖ There are also *Virtual Libraries*: Pinakes, WWW Virtual Library, Digital-librarian, Librarians Internet Index.

Structure of an IR System

- ❖ An *Information Retrieval System* serves as a *bridge between* the *world of authors* and the *world of readers/users*.
 - ❖ Writers *present a set of ideas* in a document using a set of concepts.
 - ❖ Then Users seek the IR system for *relevant documents* that *satisfy their information need*.



- ❖ What is in *the Black Box*?
 - ❖ The black box is the *processing part* of the *information retrieval system*.

Information Retrieval vs. Data Retrieval

❖ Example of *data retrieval system* is a *relational database*.

	<i>Data Retrieval</i>	<i>Info Retrieval</i>
Data organization	Structured (Clear Semantics: Name, age...)	Unstructured (No fields (other than text))
Query Language	Artificial (defined, SQL)	Free text (“natural language”), Boolean
Query specification	Complete	Incomplete
Items wanted	Exact Matching	Partial & Best matching, Relevant
Accuracy	100 % (results are <i>always</i> “correct”)	< 50 %
Error response	Sensitive	Insensitive

Typical IR Task

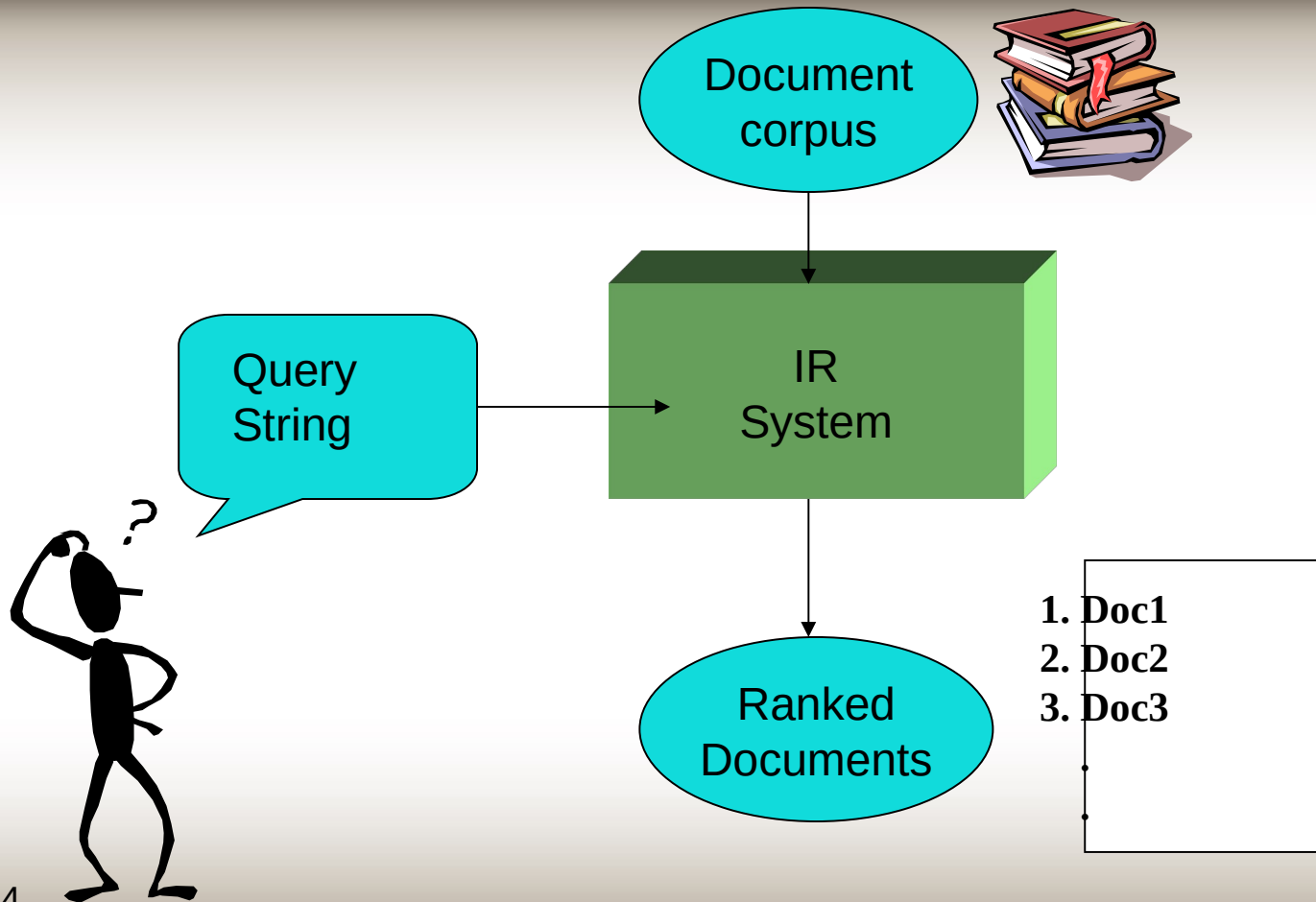
❖ Given:

- ❖ A *corpus of document collections* (text, image, video, audio) published by *various authors*.
- ❖ A *user information need* in the form of a *query*.

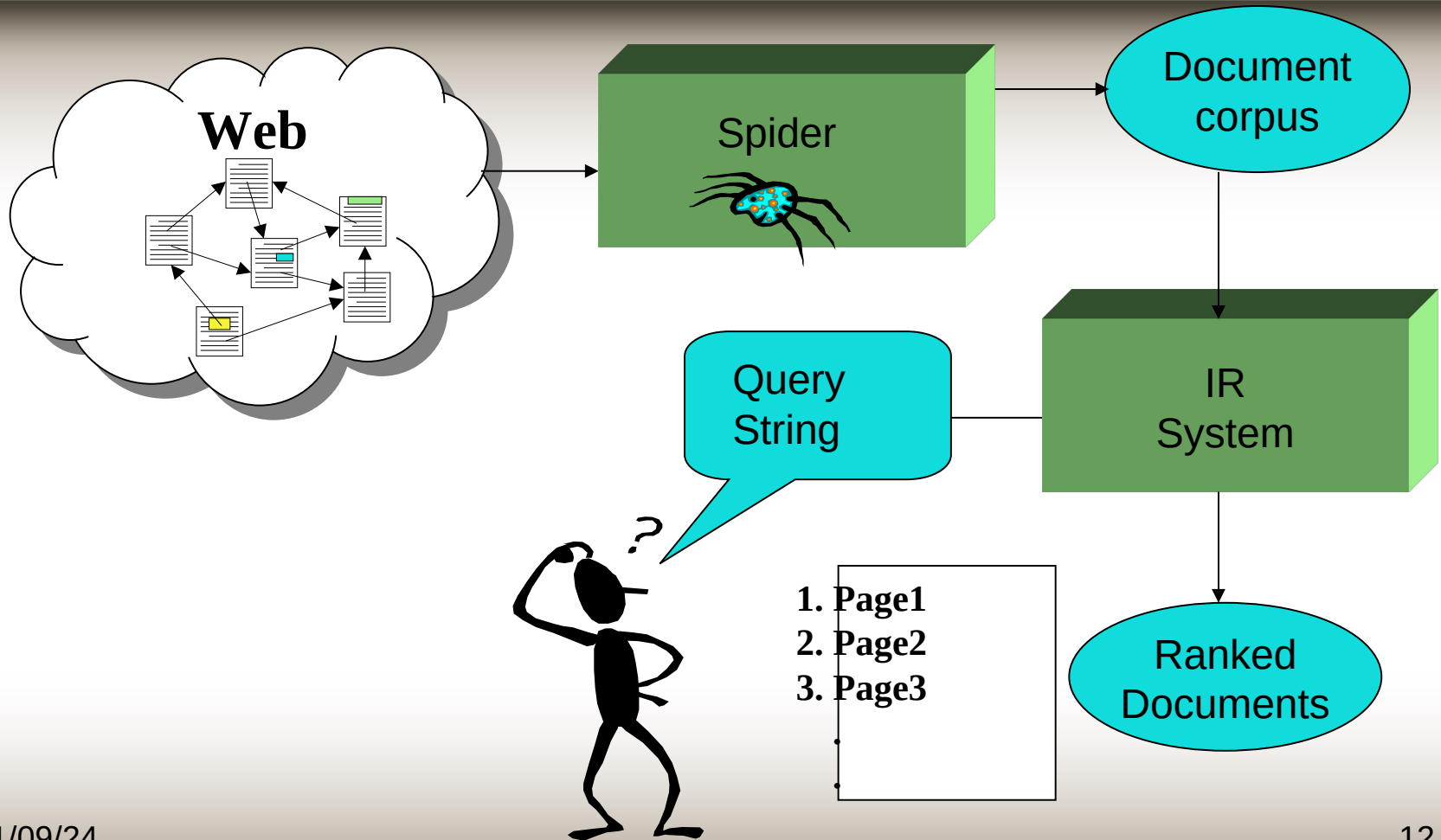
❖ An *IR system searches* for:

- ❖ A *ranked set of documents* that are *relevant* to *satisfy information need of a user*.

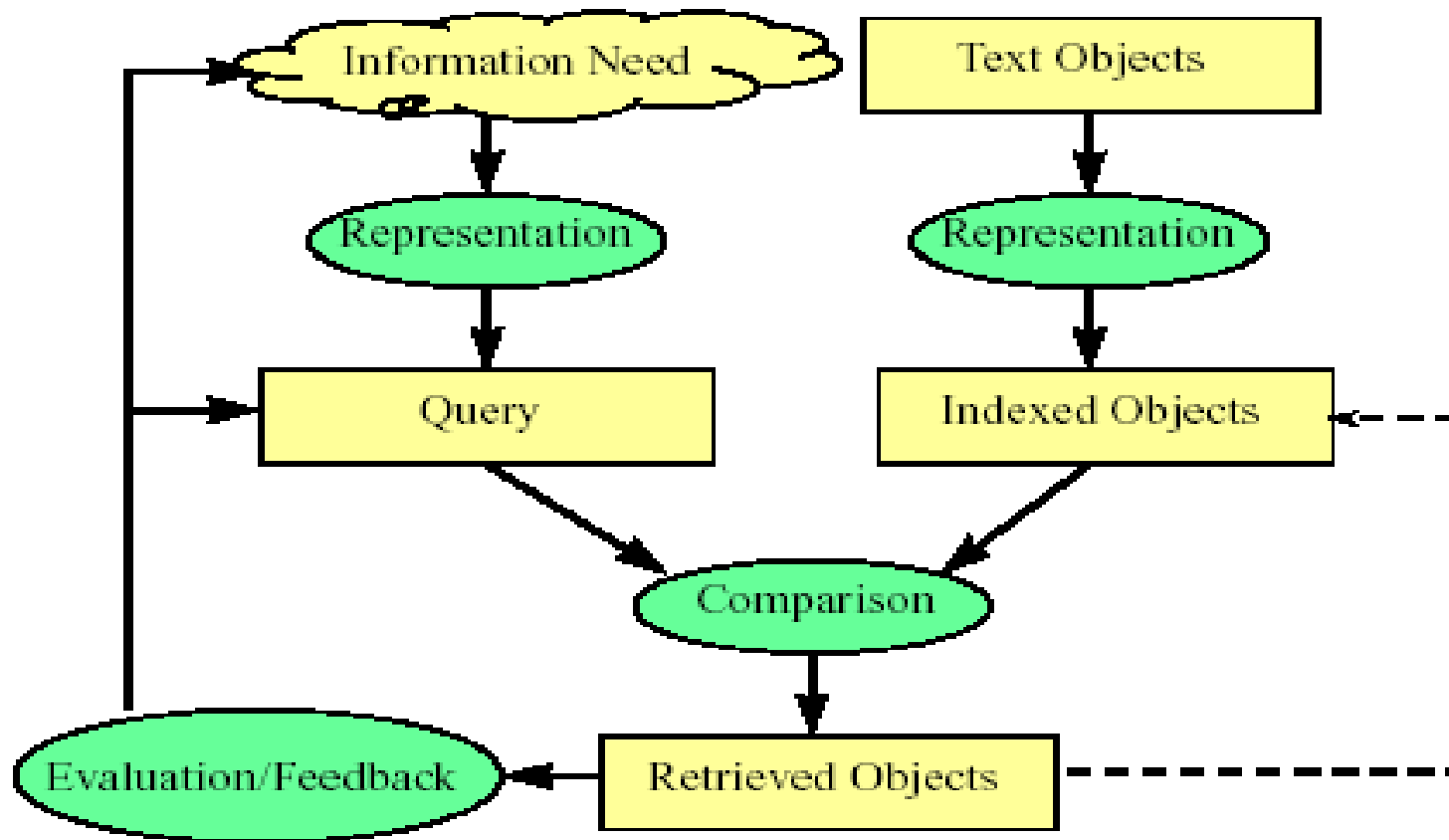
Typical IR System Architecture



Web Search System



Overview of the Retrieval Process



Issues that arise in IR

❖ Text *representation*:

- ❖ What makes a “*good*” *representation*? The use of *free-text* or *content-bearing* index-terms?
- ❖ How is a *representation generated* from text?
- ❖ What are *retrievable objects* and how are they *organized*?

❖ Information *needs representation*:

- ❖ What is an *appropriate query language*?
- ❖ How can *interactive query formulation* and *refinement be supported*?

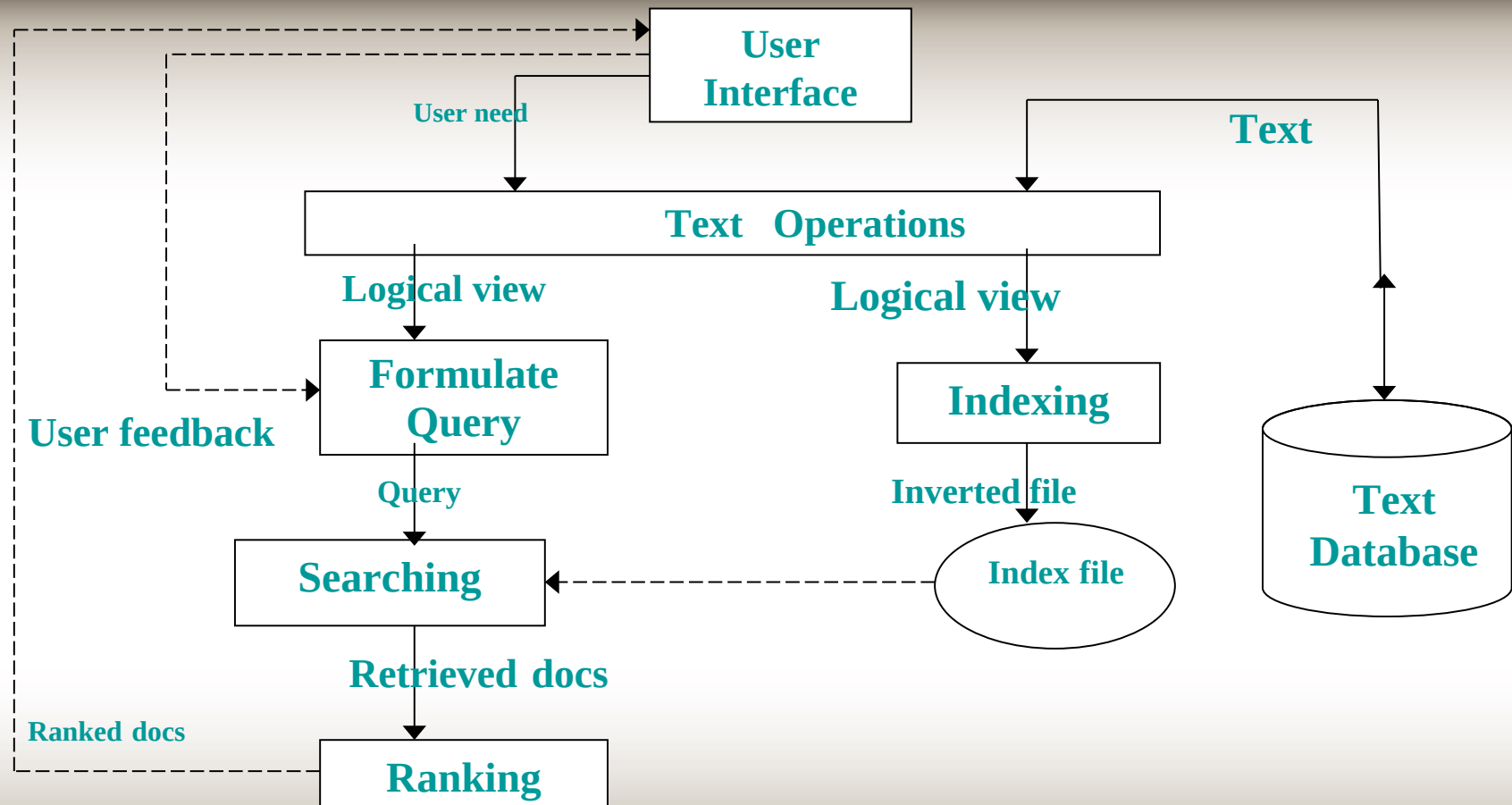
❖ *Comparing representations*:

- ❖ What is a “*good*” *model* of retrieval?
- ❖ How is *uncertainty represented*?

❖ *Evaluating effectiveness* of retrieval:

- ❖ What are *good metrics*?
- ❖ What constitutes a *good experimental test bed*?

Detail View of the Retrieval Process



Focus in IR System Design

- ❖ In improving *performance effectiveness* of the system.
 - ❖ *Effectiveness* of the system is evaluated in terms of *precision*, *recall*, ...
 - ❖ Stemming, stop words, weighting schemes, matching algorithms.
- ❖ In improving *performance efficiency*. The concern here is:
 - ❖ Storage space usage, access time, ...
 - ❖ Compression, data/file structures, space – time tradeoffs

Subsystems of IR system

- ❖ The *two subsystems* of an *IR system*:
- ❖ Indexing:
 - ❖ Is an *offline process of organizing documents* using *keywords extracted from the collection*.
 - ❖ Indexing is *used to speed up access to desired information* from document collection as per users query.
- ❖ Searching:
 - ❖ Is an *online process* that scans document corpus *to find relevant documents* that matches users query.

Statistical Properties of Text

- ❖ How is the *frequency* of *different words* distributed?
 - ❖ A few words are *very common*.
 - ❖ 2 most frequent words (e.g. “the”, “of”) can account for about *10% of word occurrences*.
 - ❖ Most words are *very rare*.
 - ❖ *Half the words* in a corpus appear *only once*, called “*read only once*”.
- ❖ How fast does *vocabulary size* grow with the *size of a corpus*?
 - ❖ Such factors affect the *performance of IR system* & can be used to select *suitable term weights* & *other aspects of the system*.

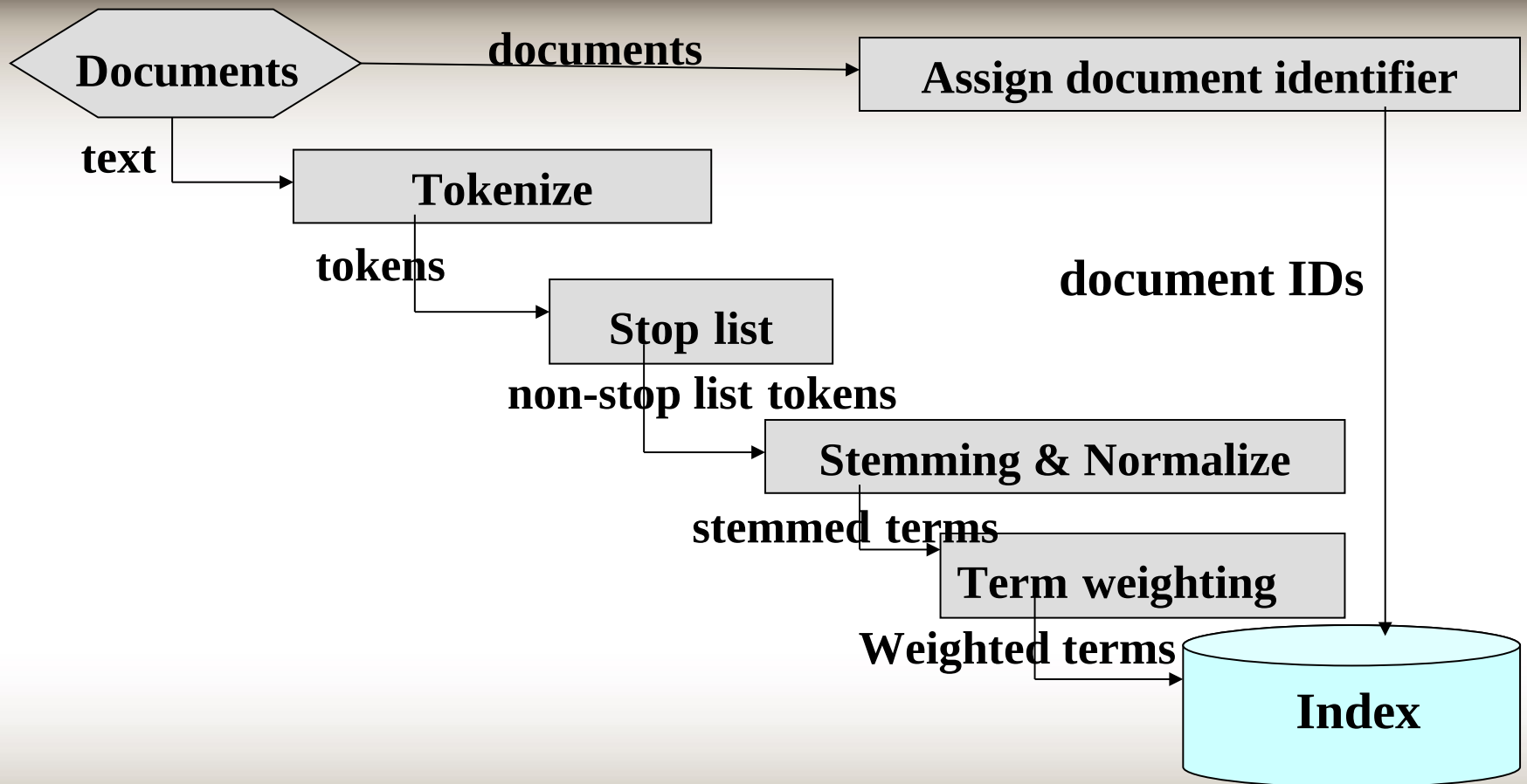
Text Operations

- ❖ Not all words in a document are *equally significant* to represent the *contents/meanings of a document*.
 - ❖ Some word carry *more meaning* than others.
 - ❖ Noun words are the *most representative* of a document content.
 - ❖ Therefore, need to *preprocess the text of a document* in a collection to be used as *index terms*.
- ❖ *Text operations* is the process of *text transformations* in to *logical representations*.
- ❖ It (text operation) generated a set of *index terms*.

Text Operations

- ❖ Main operations for *selecting index terms*:
 - ❖ *Tokenization*: identify a set of words used to describe the content of text document.
 - ❖ *Stop words removal*: filter out frequently appearing words.
 - ❖ *Stemming words*: remove prefixes, infixes & suffixes.
 - ❖ *Design term categorization structures* (like thesaurus), which *captures relationship* for allowing the *expansion* of the original query with related terms.

Indexing Subsystem



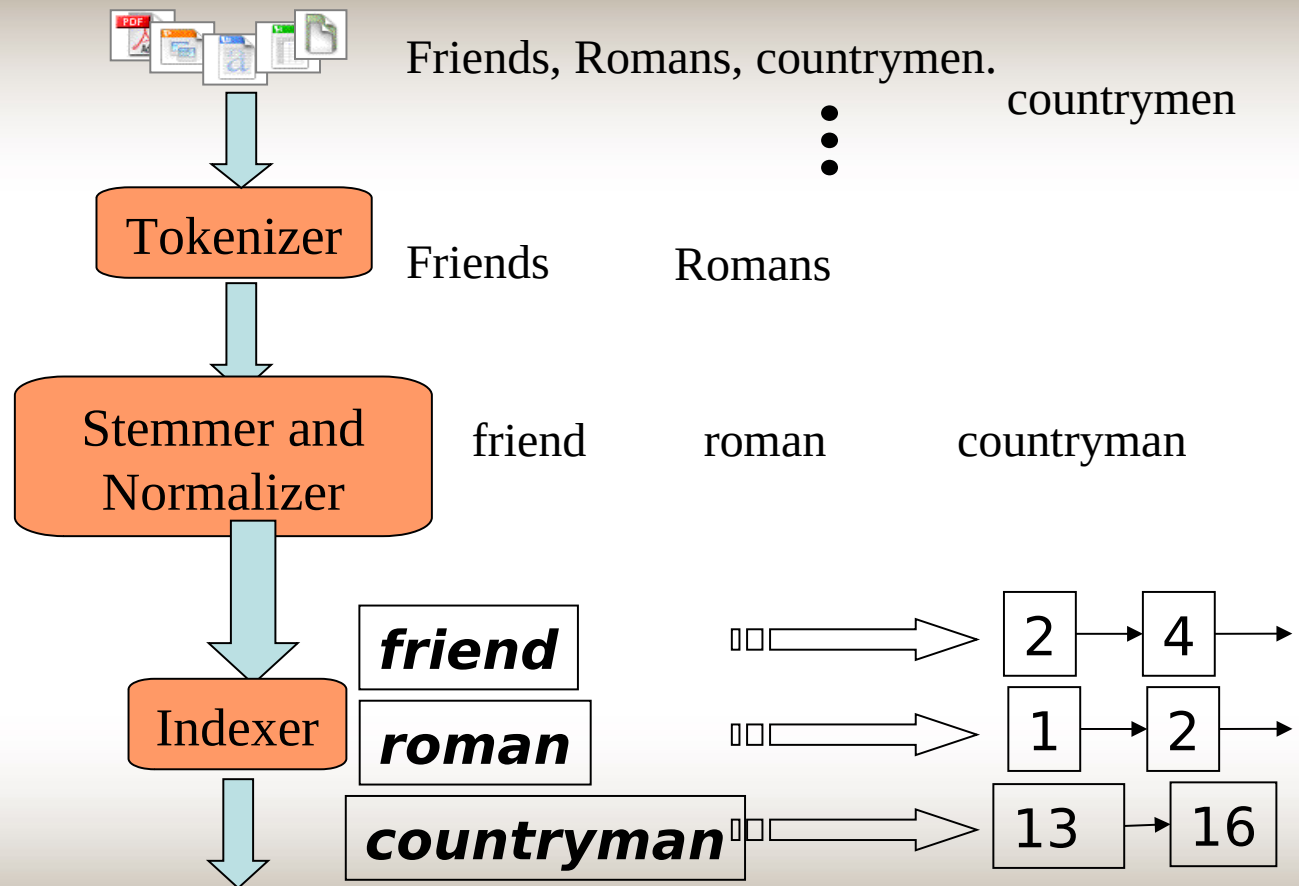
Example: Indexing

Documents to
be indexed.

Token
stream.

Modified
tokens.

Index File
(Inverted file).



Index File

- ❖ An index file *consists of records*, called *index entries*.
- ❖ Index files are *much smaller than* the *original file*.
 - ❖ For 1 GB of TREC text collection the vocabulary *has a size of only 5 MB* (Ref: Baeza-Yates and Ribeiro-Neto, 2005)
 - ❖ This *size may be further reduced* by Linguistic *pre-processing* (like *stemming* & *other normalization* methods).
- ❖ The usual *unit for text indexing* is a *word*.
 - ❖ Index *terms - are* used to look up *records in a file*.
- ❖ Index file usually has *index terms* in a *sorted order*.
 - ❖ The *sort order of the terms* in the index file provides an order on a *physical file*.

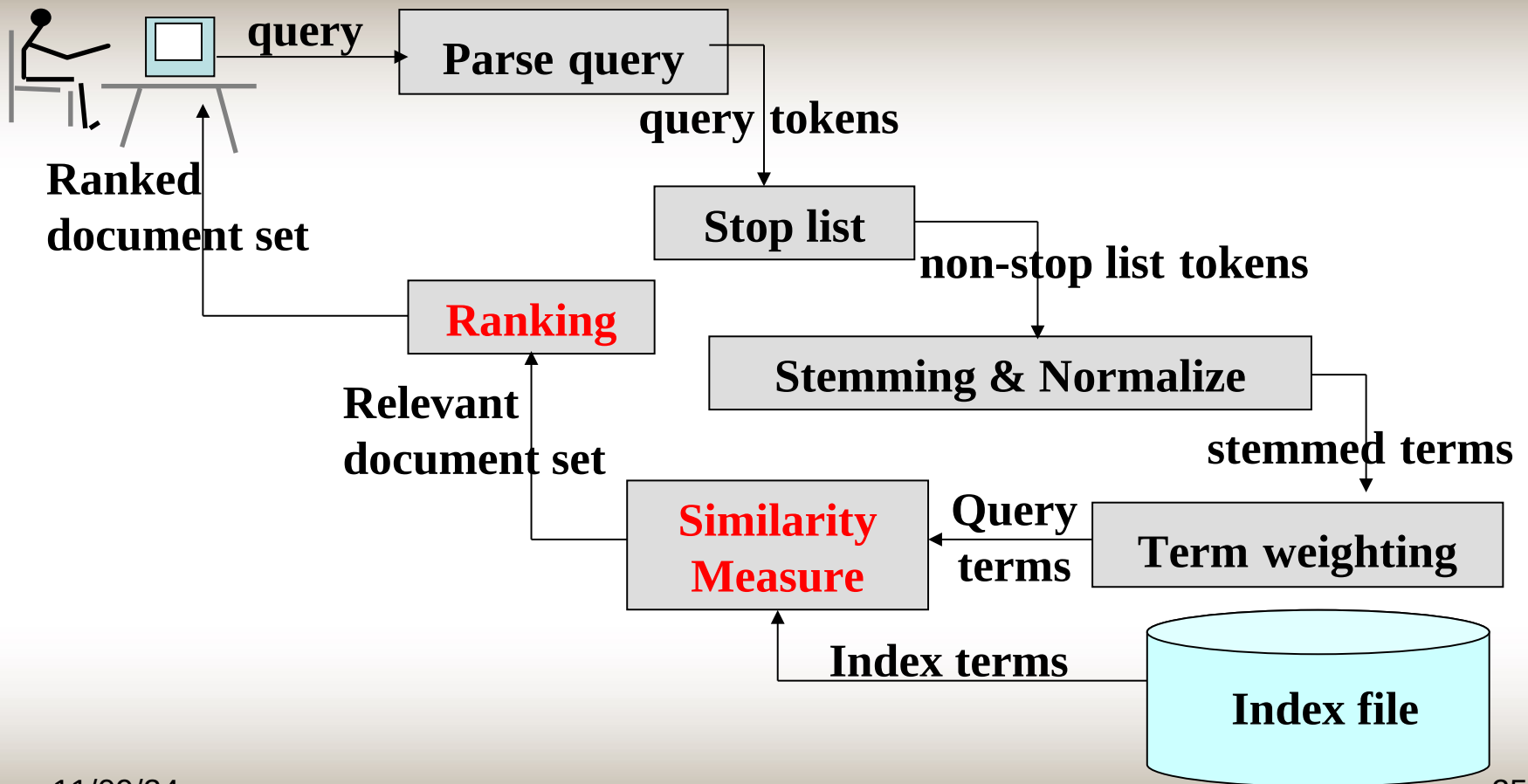
Building Index file

- ❖ An *index file* of a document is a file consisting of a *list of index terms* and a *link to one or more documents* that has *the index term*.
 - ❖ A *good index file maps* each keyword K_i to *a set of documents* D_i that *contain the keyword*.



- ❖ An index file is *list of search terms* that are organized for associative look-up, i.e., to *answer user's query*:
 - ❖ In which documents does a *specified search term appear*?
 - ❖ Where within each document *does each term appear*?
- ❖ For *organizing index file* for a collection of documents, there are *various options available*:
 - ❖ Decide what *data structure* and/or *file structure to use*. Is it *sequential file*, *inverted file*, *suffix array*, *signature file*, etc. ?

Searching Subsystem



IR Models - Basic Concepts

- ❖ One central problem regarding IR systems is the issue of predicting which documents are *relevant* and *which are not*.
 - ❖ Such a decision is usually dependent on a *ranking algorithm* which attempts to *establish a simple ordering* of the *documents retrieved*.
 - ❖ Documents appearing *at the top of this ordering* are considered to be *more likely to be relevant*.
- ❖ Thus *ranking algorithms* are at the *core of IR systems*.
 - ❖ The IR models determine the *predictions of what is relevant* and *what is not*, based on the notion of *relevance implemented* by the system.

IR Models - Basic Concepts

- ❖ After preprocessing, *N distinct terms remain* which are *Unique terms* that form the *VOCABULARY*.
- ❖ Let k_i be an *index term* i & d_j be a *document* j .
- ❖ Each *term*, i , in a *document* or *query* j , is given a *real-valued weight*, w_{ij} .
 - ❖ w_{ij} is a *weight associated* with (k_i, d_j) . If $w_{ij} = 0$, it indicates that term does not *belong to document* d_j .
- ❖ The *weight* w_{ij} quantifies the importance of *the index term* for describing the *document contents*.
 - ❖ $\text{vec}(d_j) = (w_{1j}, w_{2j}, \dots, w_{tj})$ is a *weighted vector associated* with the *document* d_j .

Mapping Documents & Queries

- ❖ Represent *both documents & queries* as N-dimensional vectors in a *term-document matrix*, which shows occurrence of terms in the *document collection/query*.

❖ E.g. $d_j = (t_{1,j}, t_{2,j}, \dots, t_{N,j})$; $q_k = (t_{1,k}, t_{2,k}, \dots, t_{N,k})$

- ❖ An entry in *the matrix corresponds* to the “*weight*” of a term in the document; *zero means the term doesn't exist* in the document.

	T_1	T_2		T_N
D_1	w_{11}	w_{12}	...	w_{1N}
D_2	w_{21}	w_{22}	...	w_{2N}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
D_M	w_{M1}	w_{M2}	...	w_{MN}

- ❖ Document collection is mapped to *term-by-document matrix*.

- ❖ View as vector in *multidimensional space*.

- ❖ Nearby *vectors* are *related*.

IR Models: Matching function

- ❖ IR models measure the *similarity between documents and queries*.
- ❖ Matching function is a *mechanism used to match query with a set of documents*.
- ❖ For example, the *vector space model* considers *documents and queries as vectors* in term-space and measures *the similarity of the document to the query*.
- ❖ Techniques for matching include *dot-product, cosine similarity, dynamic programming...*

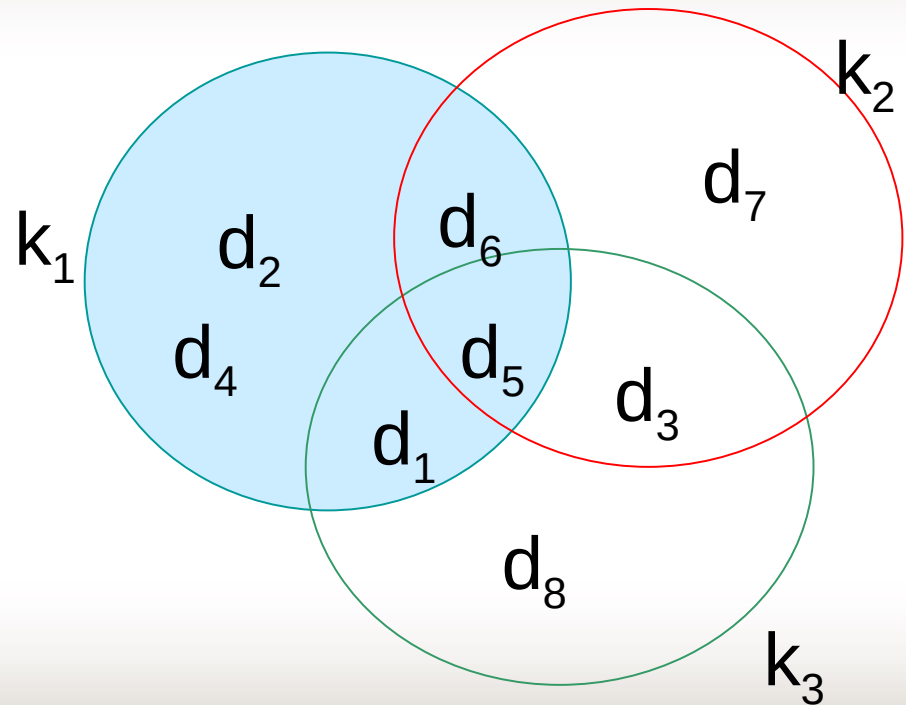
IR Models

- ❖ A number of *major models* have been developed to *retrieve information*:
 - ❖ The *Boolean* model,
 - ❖ The *vector space* model,
 - ❖ The *probabilistic* model, and
 - ❖ Other models.
- ❖ *Boolean model*: is often referred to as the "*exact match*" *model*;
- ❖ Others are the "*best match*" *models*.

The Boolean Model: Example

❖ Generate the *relevant documents retrieved* by the *Boolean model* for *the query*:

$$❖ q = k_1 \wedge (k_2 \vee \neg k_3)$$



IR System Evaluation?

- ❖ It provides the ability to *measure the difference between IR systems*.
 - ❖ How well do our *search engines work*?
 - ❖ Is system A *better than* B?
 - ❖ Under *what conditions*?
- ❖ *Evaluation drives* what to research:
 - ❖ Identify *techniques* that *work* and *do not work*,
 - ❖ There are many *retrieval models/ algorithms/ systems*.
 - ❖ Which one is *the best*?
 - ❖ What is the *best method for*:
 - ❖ *Similarity measures* (dot-product, cosine, ...)
 - ❖ *Index term selection* (stop-word removal, stemming...)
 - ❖ *Term weighting* (TF, TF-IDF,...)

Types of Evaluation Strategies

❖ *System-centered* studies:

- ❖ Given *documents*, *queries*, and *relevance judgments*.
- ❖ Try *several variations* of the system.
- ❖ Measure which *system returns the “best” hit list*.

❖ *User-centered* studies:

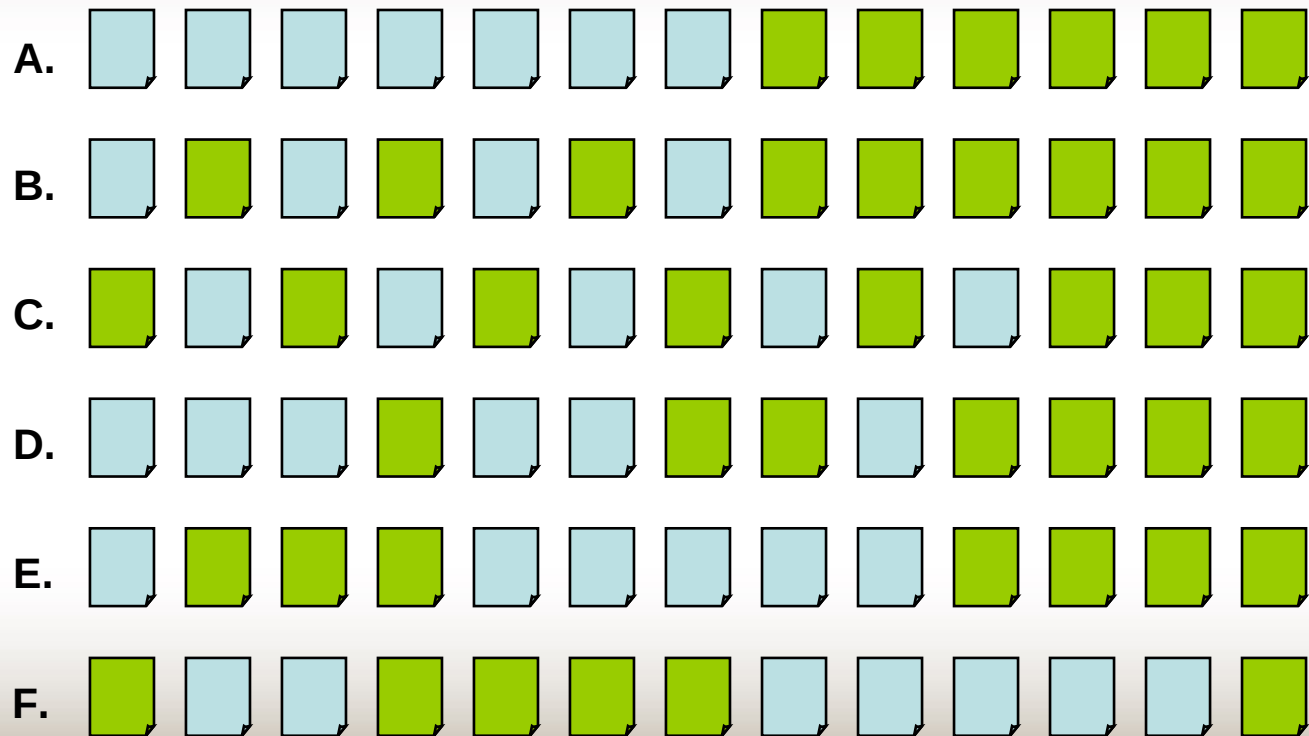
- ❖ Given *several users*, and *at least two retrieval systems*.
- ❖ Have each user try the *same task on both systems*.
- ❖ Measure which *system satisfy the “best” for users information need*.

Evaluation Criteria

- ❖ What are *some main measures* for evaluating an *IR system's performance*?
- ❖ Measure *effectiveness of the system*:
 - ❖ How is a *system capable of retrieving relevant documents* from the *collection*?
 - ❖ Is a *system better than another one*?
 - ❖ *User satisfaction*: How “good” are the documents that are returned as a *response to user query*?
 - ❖ “Relevance” of results *to meet information need of users*.

Retrieval scenario

❖ The *scenario where 13 results* retrieved by *different search engine* for a *given query*. Which *search engine* you *prefer*? Why?



Measuring Retrieval Effectiveness

- ❖ Metrics often used to *evaluate effectiveness of the system*.

	Relevant	Irrelevant
Retrieved	A	B
Not retrieved	C	D

- ❖ *Recall:*

- ❖ Is percentage of *relevant documents retrieved* from the *database* in response to *users query*. $(A / A + C)$

- ❖ *Precision:*

- ❖ Is percentage of *retrieved documents that are relevant* to the *query*. $(A / A + B)$

Query Language

- ❖ How *users query*?
 - ❖ The basic IR approach is *Keyword-based search*.
 - ❖ *Queries* are combinations of *words*.
- ❖ The *document collection* is searched for documents that contain *these words*.
- ❖ Word queries are intuitive, *easy to express* and *provide fast ranking*.
- ❖ There are *different query language*:
 - ❖ Single query,
 - ❖ Multiple query,
 - ❖ Boolean query, ... etc

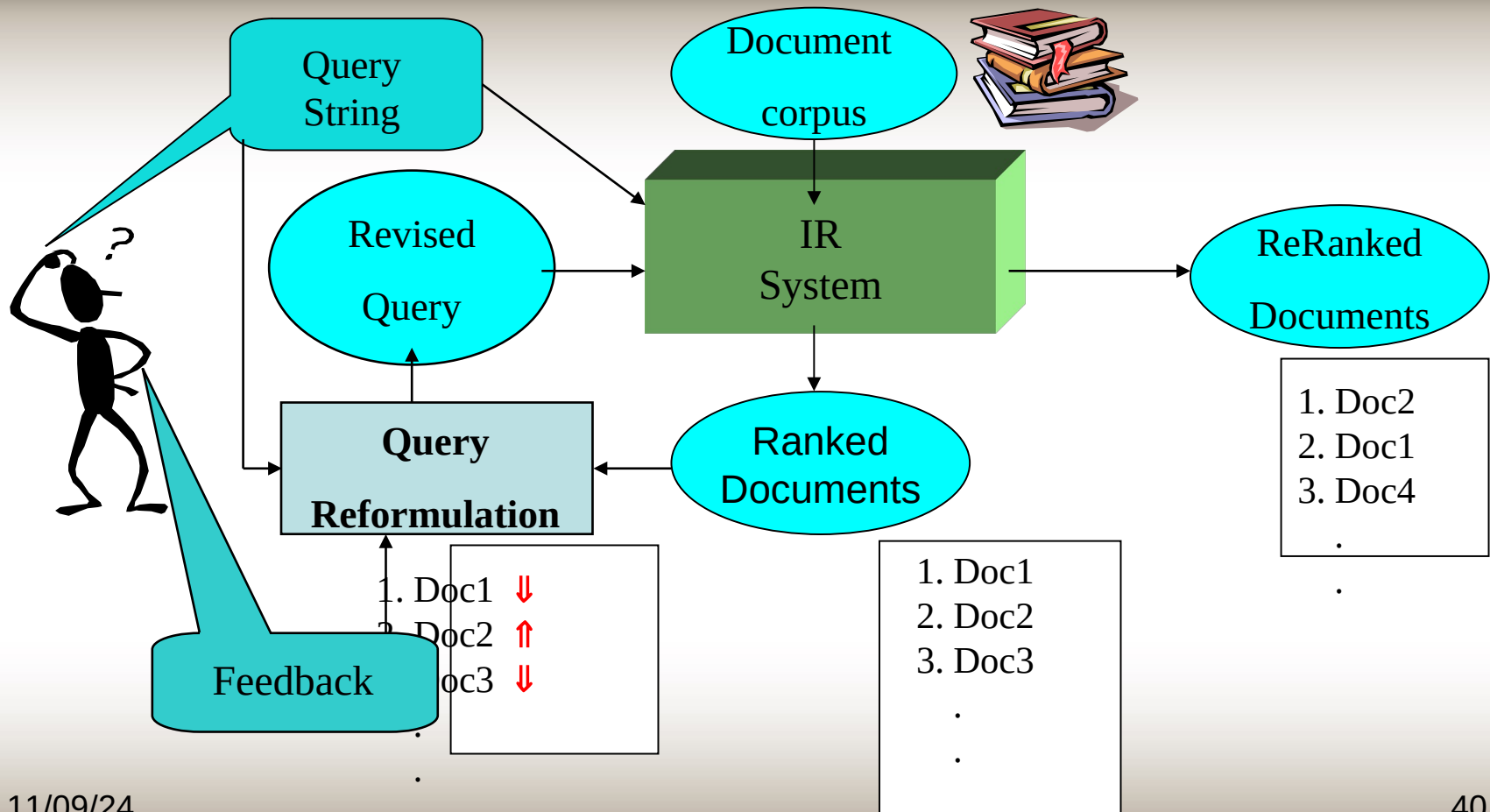
Problems with Keywords

- ❖ May *not retrieve relevant documents* that include *Synonymous terms* (words with *similar meaning*).
 - ❖ “restaurant” vs. “café”
 - ❖ “Ethiopia” vs. “Abyssinia”
 - ❖ “Car” vs. “automobile”
 - ❖ “Buy” vs. “purchase”
 - ❖ “Movie” vs. “film”
- ❖ May retrieve *irrelevant documents* that include *Polysemy terms* (terms with *multiple meaning*).
 - ❖ “Apple” (company vs. fruit)
 - ❖ “Bit” (unit of data vs. act of eating)
 - ❖ “Bat” (baseball vs. mammal)
 - ❖ “Bank” (financial institution vs. river bank)

Relevance Feedback

- ❖ After *initial retrieval results are presented*, allow the user to *provide feedback on the relevance* of one or more of the *retrieved documents*.
- ❖ Use this *feedback information* to *reformulate the query*.
- ❖ Produce *new results* based on *reformulated query*.
- ❖ Allows *more interactive, multi-pass process*.
- ❖ *Relevance feedback* can be automated in such a way that it allows:
 - ❖ Users relevance feedback,
 - ❖ Pseudo relevance feedback.

Users Relevance Feedback Architecture



Challenges for IR researchers and practitioners

- ❖ *Technical challenge*: what tools should IR systems provide to allow *effective* and *efficient manipulation of information* within such diverse media of *text, image, video* and *audio*?
- ❖ *Interaction challenge*: what *features should IR systems provide* in order to support a wide variety of users in their search for *relevant information*.
- ❖ *Evaluation challenge*: how can we *measure effectiveness* of retrieval? which tools and features are *effective* and *usable*, given the increasing diversity of *end-users* and *information seeking situations*?

Assignments - One

❖ Pick *three of the following concept* (which is not taken by other students). Review literatures (books, articles & Internet) (concerning the *meaning, function, pros and cons & application* of the concept).

- | | |
|-----------------------------|-----------------------------|
| 1. Information Retrieval | 11. Normalization |
| 2. Search engine | 12. Thesaurus |
| 3. Data retrieval | 13. Searching |
| 4. Cross language IR | 14. IR models |
| 5. Multilingual IR | 15. Term weighting |
| 6. Document image retrieval | 16. Similarity measurement |
| 7. Indexing | 17. Retrieval effectiveness |
| 8. Tokenization | 18. Query language |
| 9. Stemming | 19. Relevance feedback |
| 10. Stop words | 20. Query Expansion |

Question & Answer



Thank You !!!