# Information Retrieval

# Information Retrieval

# Domains, Applications, and Tasks

- Web search
- Vertical search
- Enterprise search
- Media search
- Question answering
- Recommender systems
- Advertising
- Personal item search
- Passage retrieval
- Filtering
- Summarization
- Clustering
- Topic detection
- Cross-language
- Federated search
- Metasearch
- Social search
- Novel-item retrieval

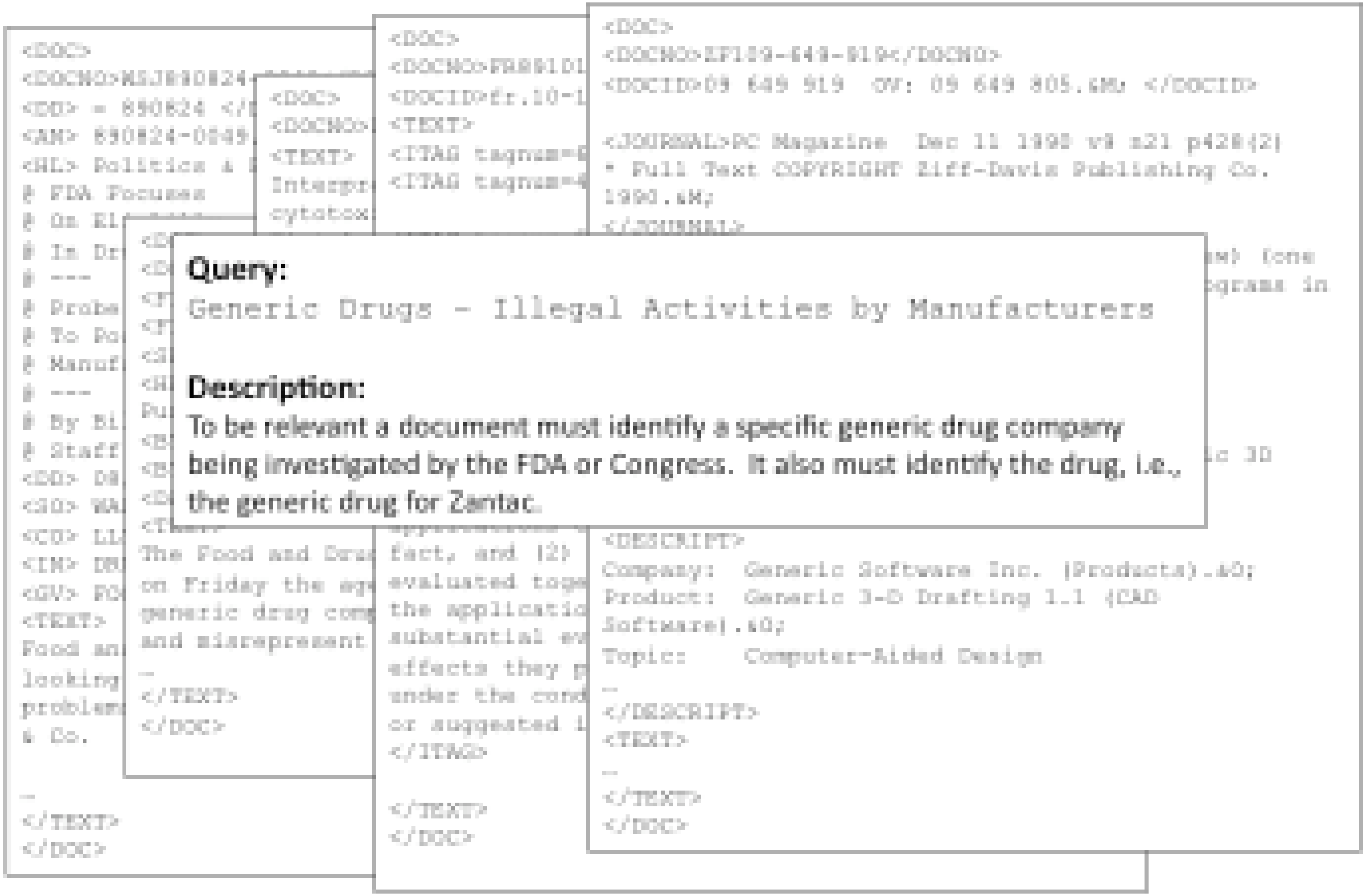
# What is IR?

- Gerard Salton, 1968:
  - *Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information*.
- This class is about computational methods for the structure, analysis, organization, storage, searching, and retrieval of information.
  - And primarily about *text documents*.

# What is a Document?

- Examples:
  - web pages, email, books, news stories, scholarly papers, text messages, Word$^{TM}$, Powerpoint$^{TM}$, PDF, forum postings, patents, IM sessions, etc.
- Common properties:
  - Significant text content.
  - Some structure (e.g., title, author, date for papers; subject, sender, destination for email).

# Examples of Documents



```
<DOC>
<DOCNO>WSJ890824-
<DD> = 890824 </
<AN> 890824-0049.
<HL> Politics & 
@ FDA Focuses
@ On El
@ In Dr
@ ---
@ Probe
@ To Po
@ Manuf
@ ---
@ By Bi
@ Staff
<DD> 08
<SO> WA
<CO> LL
<IN> DR
<GV> FO
<TEXT>
Food an
looking
problem
& Co.

...
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>
<TEXT>
Interpr
cytotox
```

```
The Food and Drug
on Friday the age
generic drug comp
and misrepresent
...
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>FR88101
<DOCID>fr.10-1
<TEXT>
<ITAG tagnum=
<ITAG tagnum=
```

```
fact, and (2)
evaluated toge
the applicatio
substantial ev
effects they p
under the cond
or suggested i
</ITAG>

</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>ZF109-649-919</DOCNO>
<DOCID>09 649 919  OV: 09 649 805.&M; </DOCID>

<JOURNAL>PC Magazine  Dec 11 1990 v9 n21 p428(2)
* Full Text COPYRIGHT Ziff-Davis Publishing Co.
1990.&M;
</JOURNAL>
```

```
<DESCRIPT>
Company:  Generic Software Inc. (Products).&O;
Product:  Generic 3-D Drafting 1.1 (CAD
Software).&O;
Topic:    Computer-Aided Design
...
</DESCRIPT>
<TEXT>
...
</TEXT>
</DOC>
```

**Query:**

Generic Drugs - Illegal Activities by Manufacturers

**Description:**

To be relevant a document must identify a specific generic drug company being investigated by the FDA or Congress. It also must identify the drug, i.e., the generic drug for Zantac.

# Documents vs. Database Records

- Database records are typically made up of well-defined fields (or *attributes)*
  - e.g. company names ,addresses ,account numbers, drug names, patent numbers, investigation file numbers.
- Easy to compare fields with well-defined semantics to queries in order to find matches.
- Our query has no fields and our documents have little structure.

# IR vs. Databases

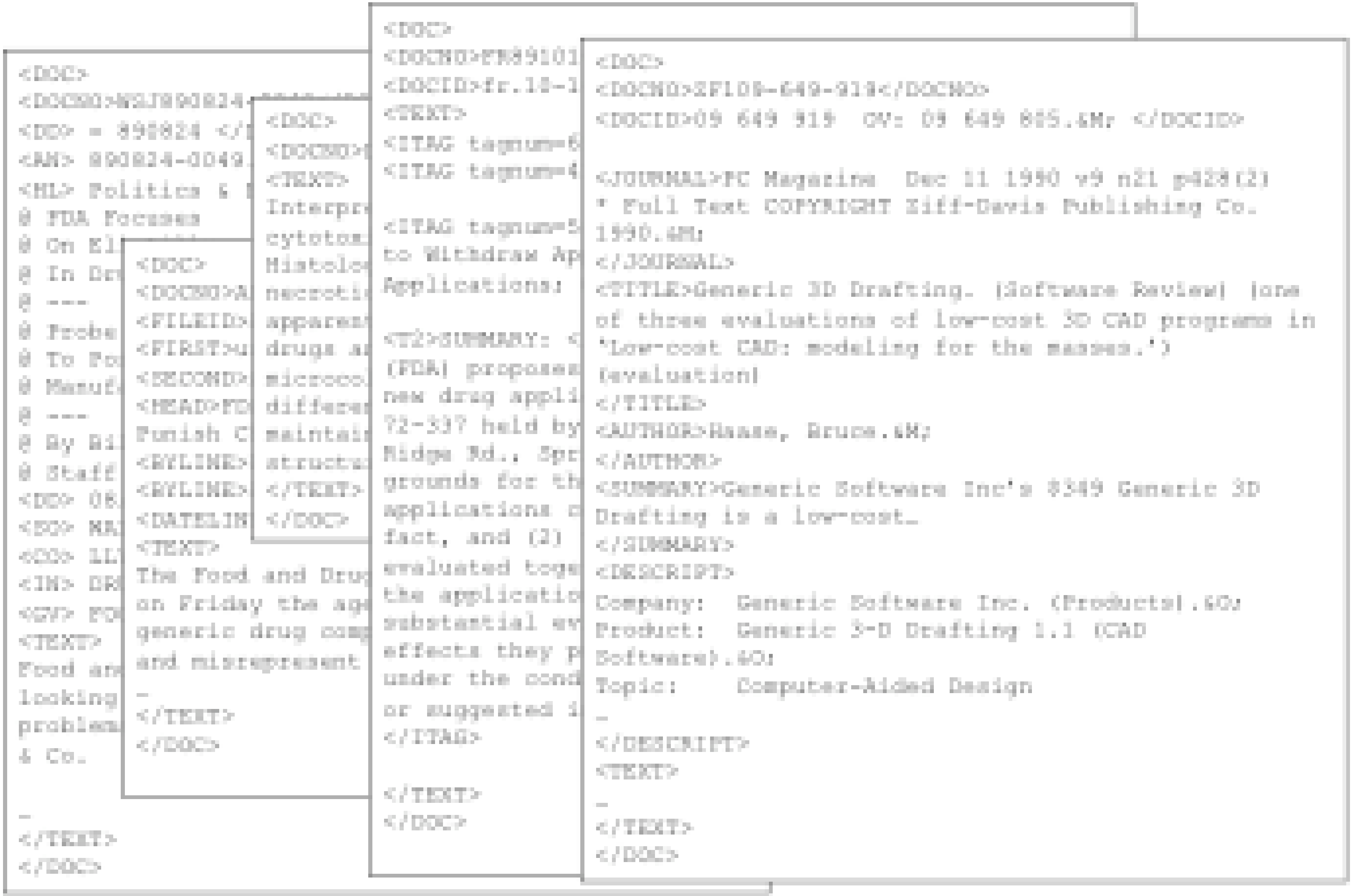
**Information Retrieval**

- Data:
  - Semi-structured.
  - Heterogeneous.
  - Noisy.
- Unstructured or semi-structured queries.
- Natural language semantics.
- Infrequent off-line index changes.

**Databases**

- Data:
  - Structured.
  - Homogeneous.
  - Clean.
- Structured queries.
- Well-defined field semantics.
- Frequent on-line index changes.

# Generic Drugs – Illegal Activities by Manufacturers



```
<DOC>
<DOCNO>WSJ890824-
<DD> = 890824 </I
<AN> 890824-0049.
<HL> Politics & I
@ FDA Focuses
@ On El
@ In Dr
@ ---
@ Proba
@ To Fo
@ Manuf
@ ---
@ By Bi
@ Staff
<DD> 08
<SO> WA
<CO> LL
<IN> DR
<GV> FO
<TEXT>
Food an
looking
problem
& Co.

–
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>A
<FILEID>
<FIRST>u
<SECOND>
<HEAD>FD
Punish C
<BYLINE>
<BYLINE>
<DATELIN
<TEXT>
The Food and Drug
on Friday the age
generic drug comp
and misrepresent
–
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>
<TEXT>
Interpre
cytotox
Histolo
necroti
apparen
drugs a
microco
differe
maintai
structu
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>FR89101
<DOCID>fr.10-1
<TEXT>
<ITAG tagnum=6
<ITAG tagnum=4

<ITAG tagnum=5
to Withdraw Ap
Applications;

<T2>SUMMARY: <
(FDA) proposes
new drug appli
72-537 held by
Ridge Rd., Spr
grounds for th
applications c
fact, and (2)
evaluated toge
the applicatio
substantial ev
effects they p
under the cond
or suggested i
</ITAG>

</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>ZF109-649-919</DOCNO>
<DOCID>09 649 919  OV: 09 649 885.&Mr </DOCID>

<JOURNAL>PC Magazine  Dec 11 1990 v9 n21 p428(2)
* Full Text COPYRIGHT Ziff-Davis Publishing Co.
1990.&M;
</JOURNAL>
<TITLE>Generic 3D Drafting. (Software Review) (one
of three evaluations of low-cost 3D CAD programs in
'Low-cost CAD: modeling for the masses.')
(evaluation)
</TITLE>
<AUTHOR>Haase, Bruce.&M;
</AUTHOR>
<SUMMARY>Generic Software Inc's $349 Generic 3D
Drafting is a low-cost…
</SUMMARY>
<DESCRIPT>
Company:   Generic Software Inc. (Products).&O;
Product:   Generic 3-D Drafting 1.1 (CAD
Software).&O;
Topic:     Computer-Aided Design
–
</DESCRIPT>
<TEXT>
–
</TEXT>
</DOC>
```

# Comparing Text

- Determining whether a document matches a query is a fundamental problem of IR.
- Exact match is not enough:
  - Many different ways to state the same information
  - Documents may be relevant even when lacking some of the query terms.
  - Documents may be nonrelevant even if they contain all the query terms.

# Relevance

- What does it mean for a document to be *relevant*?
  - Simple definition: A relevant document contains information that a person was looking for when they submitted a query to the search engine.
  - Many factors influence a person’s decision about what is relevant: e.g., task, context, novelty, style.
  - *Topical relevance* (same topic) vs. *user relevance* (everything else).

- How can we build an engine that retrieves relevant documents?

# Retrieval

- *Retrieval* models define a view of relevance.
- *Ranking algorithms* used in search engines are based on retrieval models.
- Most models describe statistical properties of text rather than linguistic properties.
  - i.e. counting simple text features such as words.
  - Statistical approach started with Luhn in the ‘50s.
  - Linguistic features can be part of a statistical model.

# Evaluation

- How do we know whether the engine is doing a good job of finding relevant documents?
  - *Evaluation* is experimental procedures and measures for comparing system output with user expectations.
  - IR evaluation methods now used in many fields.
  - *Recall* and *precision* are examples of effectiveness measures.

# Not Just Documents

- New applications increasingly involve media.
  - e.g. video, photos, music, speech
- Like text, contents is difficult to describe and compare.
  - text may be used to represent them (e.g. tags).
- IR approaches to search and evaluation are appropriate.

# Dimensions of IR

| Content | Applications | Tasks |
|---|---|---|
| Text | Web search | Ad hoc search |
| Images | Vertical search | Filtering |
| Video | Enterprise search | Classification |
| Scanned docs | Desktop search | Question answering |
| Audio | Forum search | |
| Music | P2P search | |
| | Literature search | |

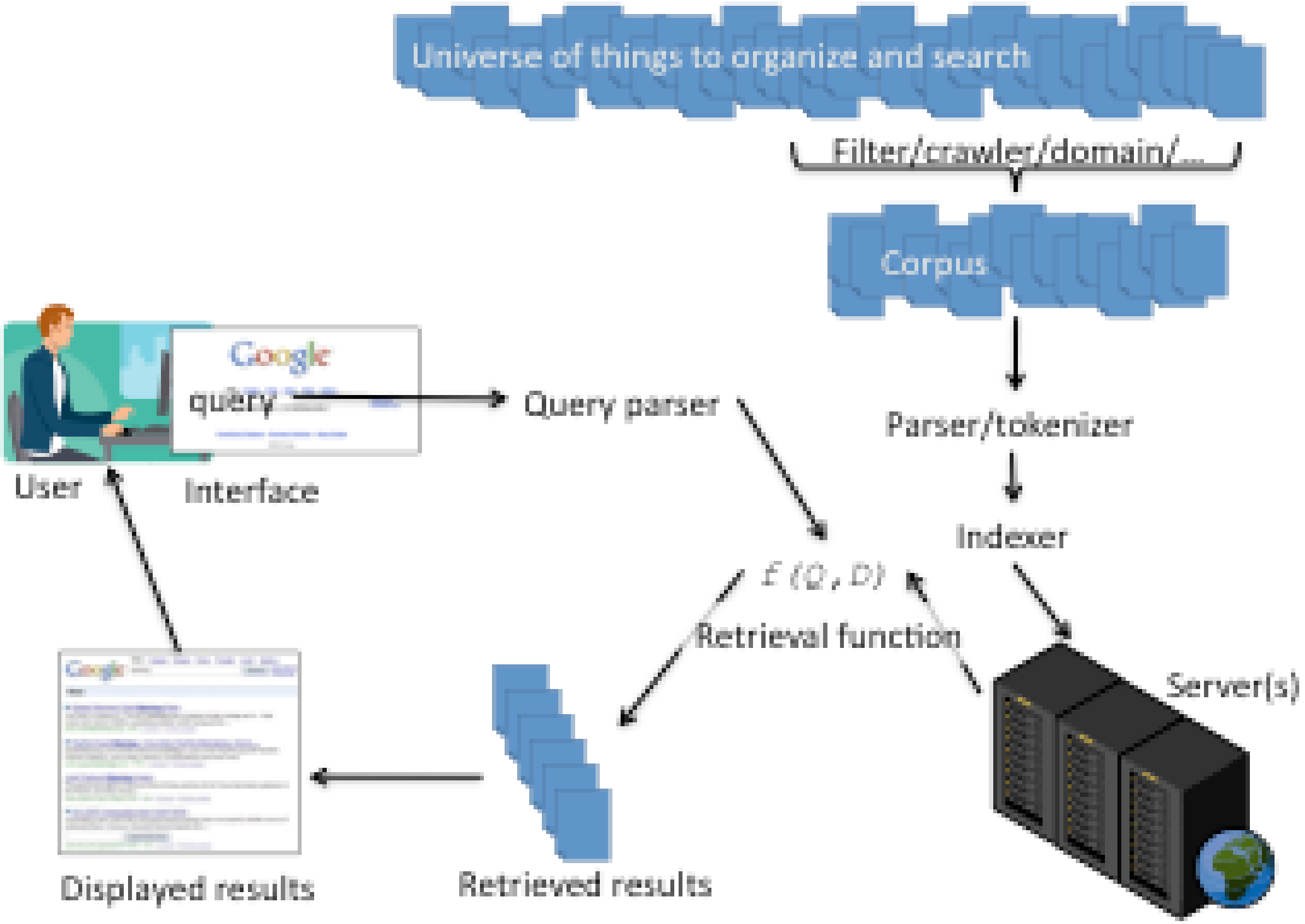
# IR Tasks

- Ad-hoc search:
  - Find relevant documents for an arbitrary text query.
- Filtering:
  - Identify relevant user profiles for a new document.
- Classification:
  - Identify relevant labels for documents.
- Question answering:
  - Give specific answer to a question.

# IR and Search Engines

- A search engine is the practical application of information retrieval techniques to large scale text collections.
- Relevance, retrieval, evaluation are issues.
- So are users and information needs, performance, coverage, updating, scalability, adaptability, and ability to handle specific problems (like spam).

# Components of a Search Engine



# Building a Search Engine

- Text processing and indexing.
  - Parsing; tokenizing; stopping and stemming; inverting indexes; scalability; index updates.
- Query processing and ranking.
  - Query languages; index look-up; retrieval models; features; relevance feedback; user interaction.
- Evaluation.
  - Effectiveness at performing task; querying speed; user satisfaction.

# Course Overview

- This course is about information retrieval in practice: the application of IR to search engine design and implementation.
- Course project:
  - Design and implement a small search engine capable of indexing and searching Wikipedia pages.
  - Evaluate its performance over provided queries.
  - Add something interesting to it.