Information Retrieval

CISC489/689-010, Lecture #1

Monday, Feb. 9

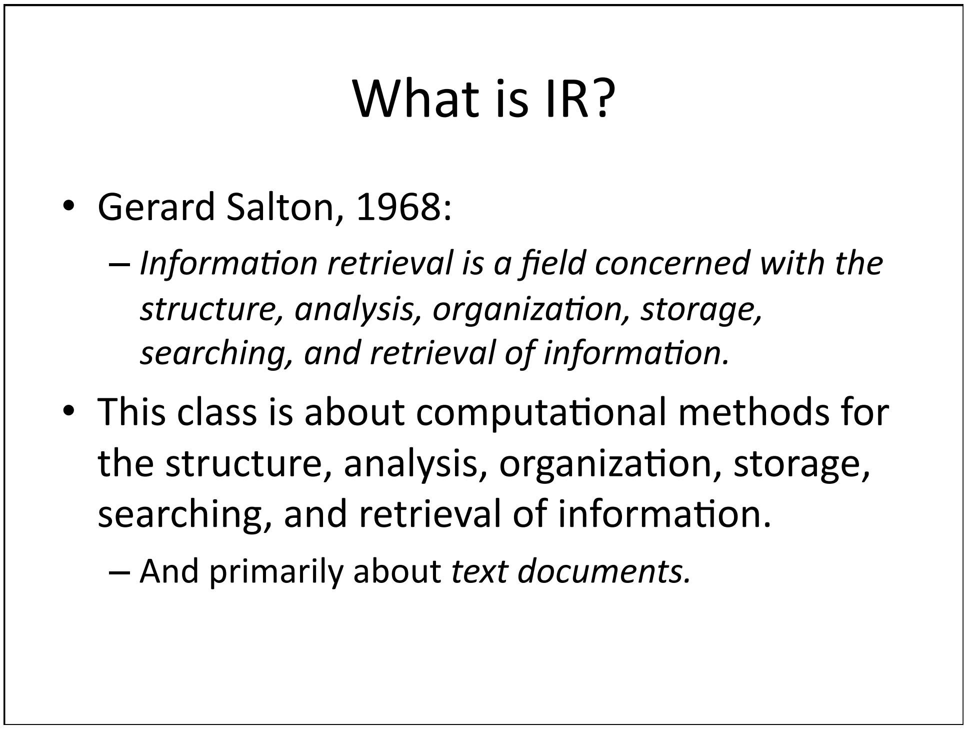
Ben Carterette

Information Retrieval

Information Retrieval

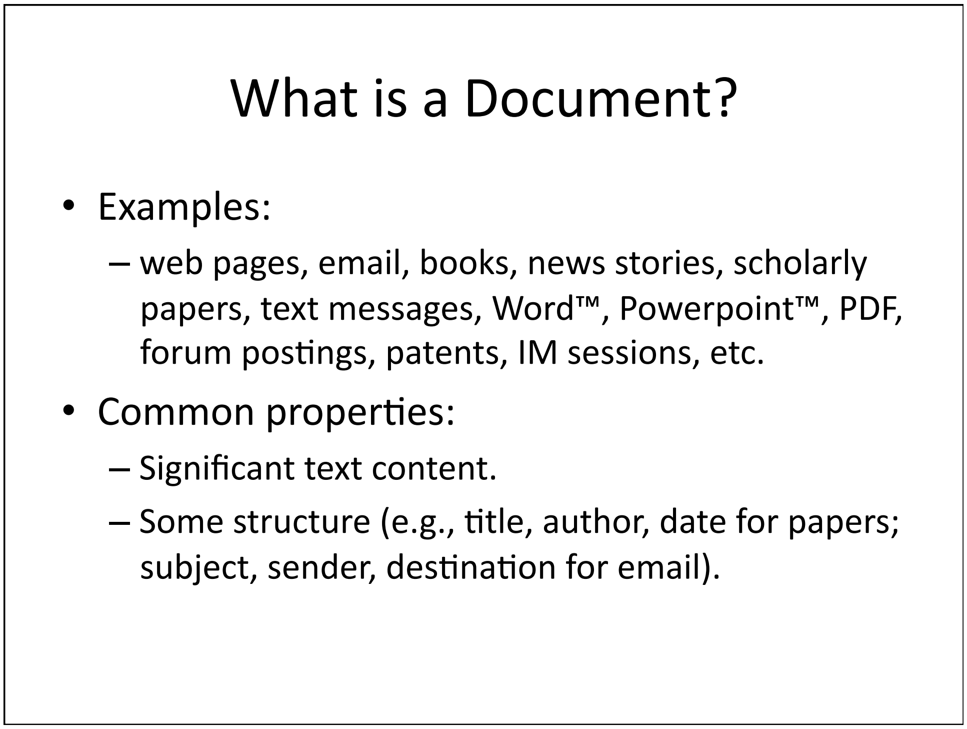
Domains, Applications, and Tasks

- 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?

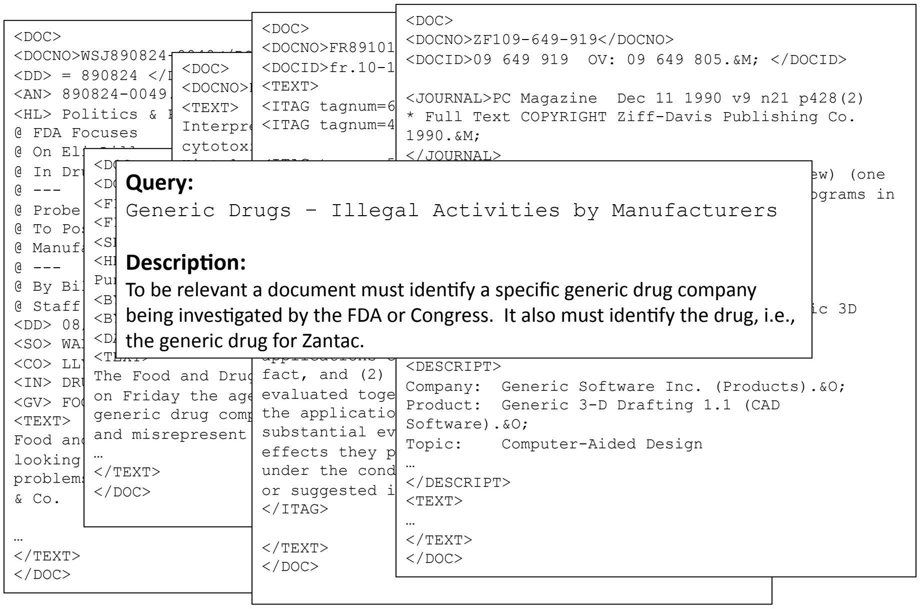
- Gerard Salton, 1968:
  - *Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.*
- This class is about computational methods for the structure, analysis, organization, storage, searching, and retrieval of information.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



What is a Document?

- Examples:
  - web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, 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 </I
<AN> 890824-0049
<HL> Politics & I
@ FDA Focuses
@ On El
@ In Dr
@ ---
@ Probe
@ To Po
@ Manufa
@ ---
@ By Bi
@ Staff
<DD> 08
<SO> WA
<CO> LL
<IN> DR
<GV> FO
<TEXT>
Food an
looking
problems
& Co.

…
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>
<TEXT>
Interpre
cytotox
```

```
<D
<D
<F
<F
<S
<H
Pu
<B
<B
<D
<TEXT>
The Food and Drug
on Friday the age
generic drug comp
and misrepresent
…
</TEXT>
</DOC>
```

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

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>
                                           ew) (one
                                           ograms in

                                           ic 3D

<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

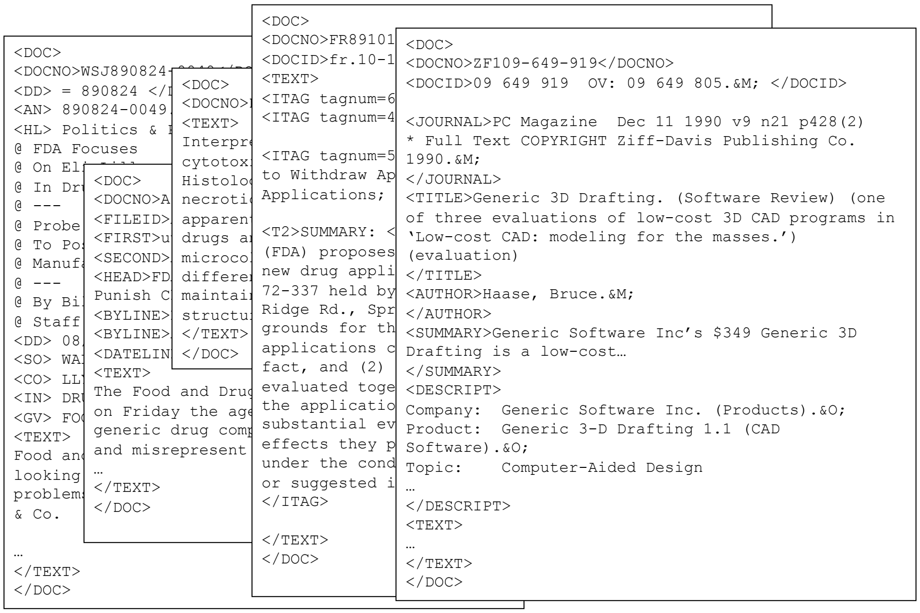
Information Retrieval

- 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



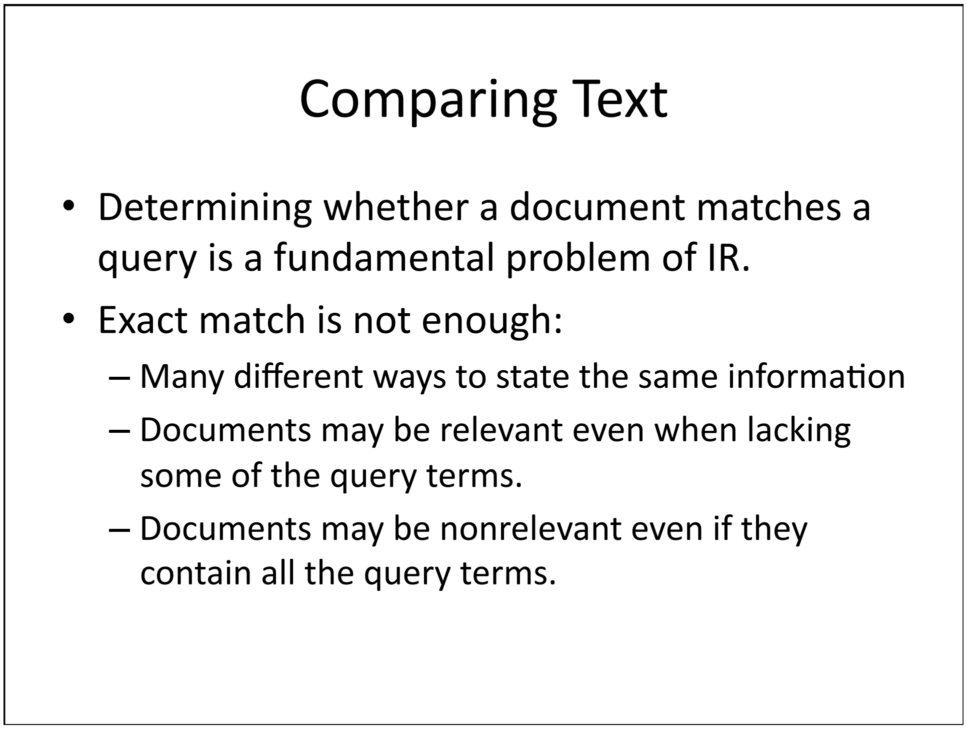
```
<DOC>
<DOCNO>WSJ890824-
<DD> = 890824 </
<AN> 890824-0049
<HL> Politics &
@ FDA Focuses
@ On El
@ In Dr
@ ---
@ Probe
@ To Po
@ Manuf
@ ---
@ By Bi
@ Staff
<DD> 08
<SO> WA
<CO> LL
<IN> DR
<GV> FO
<TEXT>
Food an
looking
problem
& Co.

…
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>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
apparent
drugs a
microco
differe
maintai
structu
</TEXT>
</DOC>
```

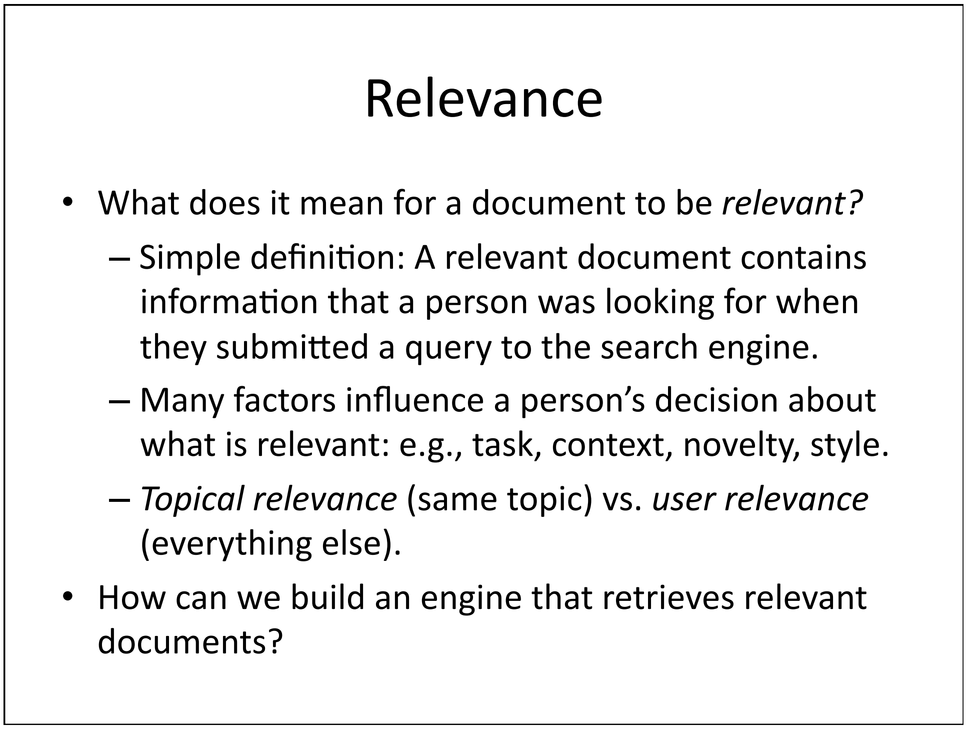
```
<DOC>
<DOCNO>FR89101
<DOCID>fr.10-1
<TEXT>
<ITAG tagnum=6
<ITAG tagnum=4

<ITAG tagnum=5
to Withdraw Ap
Applications;

<T2>SUMMARY: <
(FDA) proposes
new drug appli
72-337 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: 09 649 805.&M; </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 new media.
  - e.g. video, photos, music, speech
- Like text, content 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	

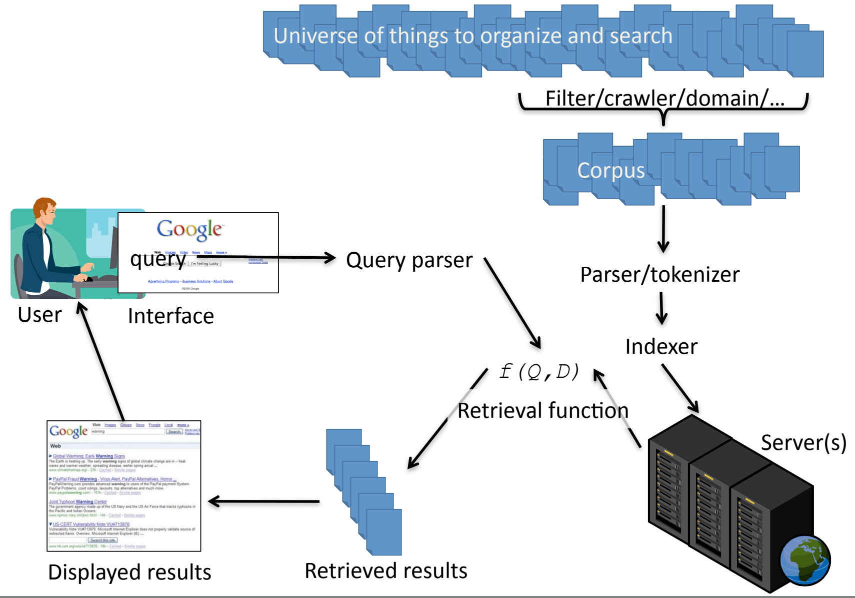
IR Tasks

- 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 a 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

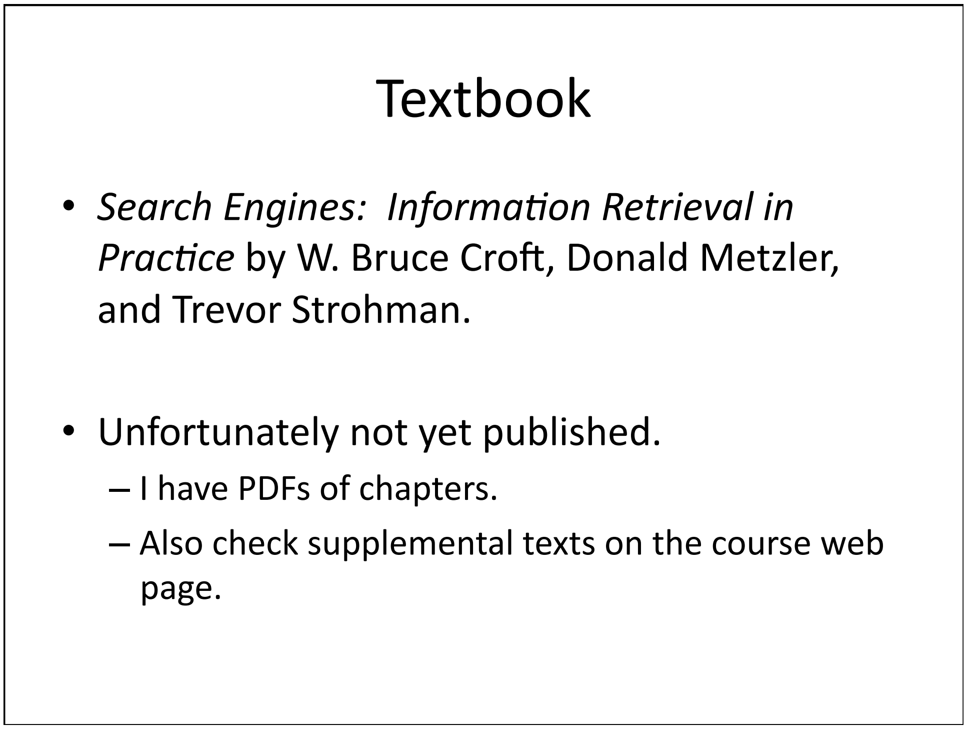
- Text processing and indexing.
  - Parsing; tokenizing; stopping and stemming; inverted indexes; scalability; index updates.
- Query processing and ranking.
  - Query languages; index look-up; retrieval models; features; relevance feedback; user interaction.
- Evaluation.
  - Effectiveness at performing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

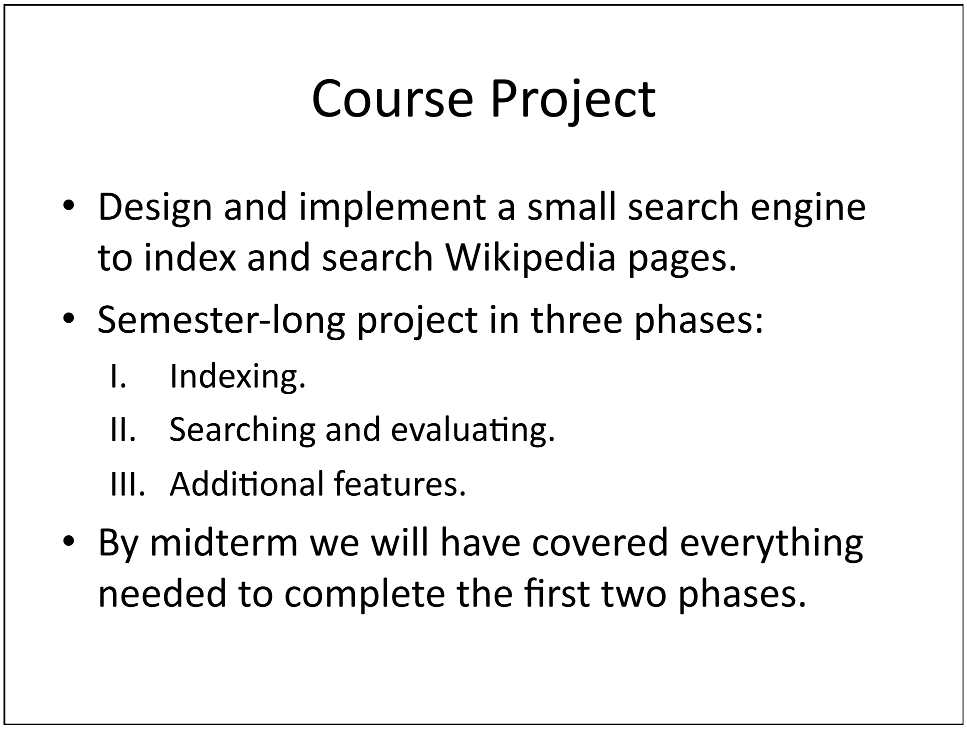
Course Structure

- First half:
  - – Fundamentals of indexing, retrieval, and evaluation.
  - – By the midterm we will have covered all aspects of designing a basic search engine.
- Second half:
  - – Additional topics in search engine functionality.
  - – Fielded search, user interaction, clustering, link-graph features, crawling, etc.



Textbook

- *Search Engines: Information Retrieval in Practice* by W. Bruce Croft, Donald Metzler, and Trevor Strohman.
- Unfortunately not yet published.
  - – I have PDFs of chapters.
  - – Also check supplemental texts on the course web page.



Course Project

- Design and implement a small search engine to index and search Wikipedia pages.
- Semester-long project in three phases:
  - I. Indexing.
  - II. Searching and evaluating.
  - III. Additional features.
- By midterm we will have covered everything needed to complete the first two phases.

Course Project: Phases

- For phases I and II, you will produce:
  - A written report of your design decisions and implementation details, including problems you encountered and how you resolved them.
  - Code.
  - Milestone worksheet responses.
- Timeline:
  - Phase I: about 1.5 months.
  - Phase II: about 1 month.
  - Phase III: about 1 month.

Course Project: Milestones

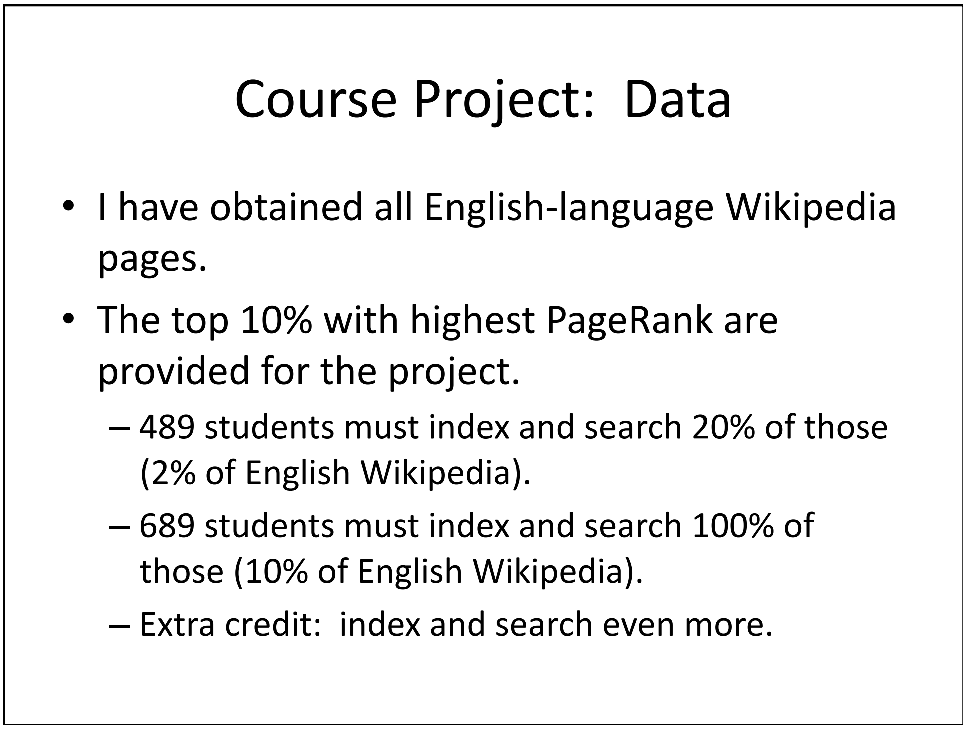
- Each phase has milestones to make sure you are not running into trouble.
  - Worksheets with questions you can answer using your code.
  - Milestones and worksheets will be available in advance so you may work ahead if desired (we recommend it).
  - If you are having trouble, we will be able to help you early.

Course Project: Phase III

- Phase III involves adding extra features to your engine.
- Anything we cover in the second half of the course, or anything in the book but not covered, or something else.
- You will write a 2-4 page proposal explaining how you would add the feature to your current code base.
- At the end of the semester you will give a short presentation on your engine.

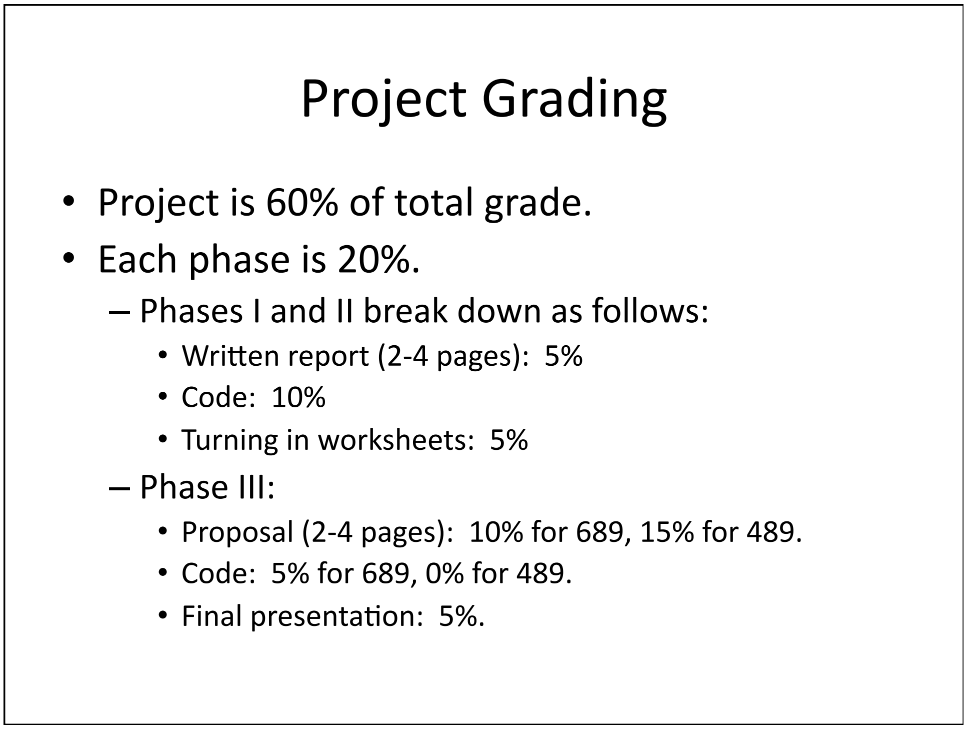
Course Project: Implementation

- This is a programming project!
- You may use any programming language the professor and/or TA understand.
  - We highly recommend C, C++, or Java.
- You will have accounts on my lab cluster.
  - No disk quotas; 16Gb RAM per node; 8 cores per node.
  - Do not use for file sharing or other illicit activity!



Course Project: Data

- I have obtained all English-language Wikipedia pages.
- The top 10% with highest PageRank are provided for the project.
  - 489 students must index and search 20% of those (2% of English Wikipedia).
  - 689 students must index and search 100% of those (10% of English Wikipedia).
  - Extra credit: index and search even more.



Project Grading

- Project is 60% of total grade.
- Each phase is 20%.
  - Phases I and II break down as follows:
    - Written report (2-4 pages): 5%
    - Code: 10%
    - Turning in worksheets: 5%
  - Phase III:
    - Proposal (2-4 pages): 10% for 689, 15% for 489.
    - Code: 5% for 689, 0% for 489.
    - Final presentation: 5%.

Homeworks and Exams

- In addition to the project, there will be 5 homeworks and 2 exams (midterm and final).
- Each homework is 4% of total grade.
- Each exam is 10% of total grade.
- Exams will cover implementation details of project.

Books and Resources

- *Information Retrieval,* Keith van Rijsbergen.
  - http://www.dcs.gla.uc.uk/Keith/Preface.html
- *Introduction to Information Retrieval,* Manning et al.
  - http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html
- Check the course web page often!
  - http://www.cis.udel.edu/~carteret/CISC689