Teaching an Information Retrieval Course

David Kauchak
Middlebury College
dkauchak@cs.middlebury.edu

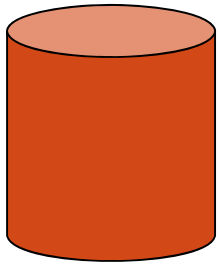
Information Retrieval (IR)

Study of searching and processing "*document*" collections

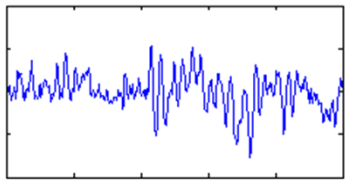
web pages

(micro) blogs

e-mail



databases



audio

image

video

Today...

Course content overview

Why teach an IR course?

Sample assignments: building an IR system from scratch

Designing an IR course

http://nlp.stanford.edu/IR-book/information-retrieval.html

- books
- other courses

McCown, F. (2010). Teaching Web Information Retrieval to Undergraduates. In *SIGCSE*.

Mizzaro, S. (2007). Teaching of Web Information Retrieval: Web First or IR First? In *TLIR*.

IR Course Content

The basics

- text processing
- IR system features and uses
- Index construction
- Evaluation
- IR and the web

IR extensions

alternate media (image, audio, video)

cross-lingual IR

parallel computing

machine learning

snippet generation

natural language processing

query expansion

relevance feedback

online advertising

question answering

algorithms/data structures

Date	Topic	Reading	Slides/Handouts
9/2	Admin. material, Introduction	Ch. 1 except 1.2	admin, slides, pdf
9/7	Text pre-processing	Ch. 2, 5.1	slides, pdf
9/9	Index construction	Ch 1.2, Ch. 4	slides, pdf
9/14	Index compression	Ch. 5	slides, pdf
9/16	TF-IDF	Ch. 6 except 6.4.4	slides, pdf
9/21	Faster TF-IDF	Ch. 7, article	slides, pdf
9/23	Evaluation	Ch. 8	slides, pdf
9/28	Spelling correction	Ch. 3.3, 3.4, article	slides, pdf
9/30	Relevance feedback/ query expansion	Ch. 9	slides, pdf
10/5	Web search basics	Ch. 19 (except 19.3), article	slides, pdf
10/7	Crawling	Ch. 20	slides, pdf
10/12	Link Analysis	Ch. 21	slides, pdf
10/14	**Midterm**		
10/19	**fall recess**		
10/21	Text segmentation	paper, article	slides, pdf
10/26	Audio processing basics	paper	slides, pdf
10/28	Audio search	paper	slides, pdf
11/2	Image processing basics	paper, article	slides, pdf
11/4	Project proposal discussion		
11/9	Document Image search	paper	slides, pdf
11/11	Information Extraction	paper, article	slides, pdf
11/12 4:15 Rose Hills	Document modeling (*substitute lecture*)	paper	
11/16	Text classification	Ch. 13 (except 13.5), 14.intro, 14.1, 14.3-6, 15-15.3	slides, pdf
11/18	Text classification2	article	slides, pdf
11/23	Text clustering	Ch. 16	slides, pdf
11/25	*No class*, substituted on 11/12		
11/30	Hierarchical clustering	Ch. 17, paper, article	slides, pdf
12/2	Online Advertising	Ch. 3.9.2, 19.3	slides, pdf
12/7	Ethics in IR		
12/9	Review.(cross-lingual IR?)		
12/14	Final time 9am - project presentations		

Course variability

basics

extensions

http://www.cs.pomona.edu/~dkauchak/classes/f09/cs160-f09/

Course variability

The first part of this course covers the foundations on Information Retrieval, constructed and functions -- in particular, the material needed to carry out the determined in part by student interest and the projects that are selected.

1. Evaluation
2. Retrieval models
   - Language modeling, Boolean
   - Vector space, Latent Semantic Indexing
   - Probabilistic IR
3. Statistics of text
4. Indexing models (storing and accessing)
5. File organization
6. Efficiency, possibly including compression
7. Clustering
8. Relevance feedback

basics

The second part of the course covers advanced or more recent topics. The ac selected. Much of this information will be taken from recent research papers

1. Document filtering
2. Distributed retrieval
3. Web search
4. Question answering
5. Multimedia retrieval
6. Cross-language retrieval
7. Advanced evaluation issues
8. Interactive retrieval
9. Interation with Natural Language Processing
10. ...

extensions

http://ciir.cs.umass.edu/cmpsci646/syllabus.html

Course variability

Session	Date	Topic
1	Jan 26	Cancelled!
2	Feb 2	Structure of IR Systems Evidence from Terms
3	Feb 9	Vector Space Models
4	Feb 16	Language Models
5	Feb 23	Evidence from Behavior
6	Mar 1	Evidence from Metadata
7	Mar 8	User Interaction
8	Mar 15	Evaluation
	Mar 22	Spring Break
9	Mar 29	Indexing
10	Apr 5	Cross-Language Retrieval
11	Apr 12	Document Image Retrieval
12	Apr 19	Speech and Music retrieval
13	Apr 26	Photograph and Video Retrieval
14	May 3	Project Reports
15	May 10	Final Exam

basics

extensions

http://terpconnect.umd.edu/~oard/teaching/796/spring04/syllabus.html

Why IR?

Relevant application

- IR systems show up everywhere!
- Good skills for finding a job

Low barrier to entry

- Technically I only require CS2 (though some further maturity helps)

Good combination of NLP and data structures

Flexible course design

- Bias towards NLP
- Bias towards alternate media
- Bias towards cluster computing
- Bias towards theory
- Bias towards software development
- …

Sample IR assignments

Develop a working IR system from the ground up

High-level goals

- Reinforce the classroom concepts
- Improve programming skills and experience more realistic software project environment
- Introduce students to the research process

Project overview

4 structured assignments

- cover the basic IR concepts
- assignments are cumulative
- build a barebones IR system

basics

Final project

- extend the basic system
- student selected topics
- research-oriented
- open-ended

extensions

Assignment 1: Text processing

Implement text processing methods

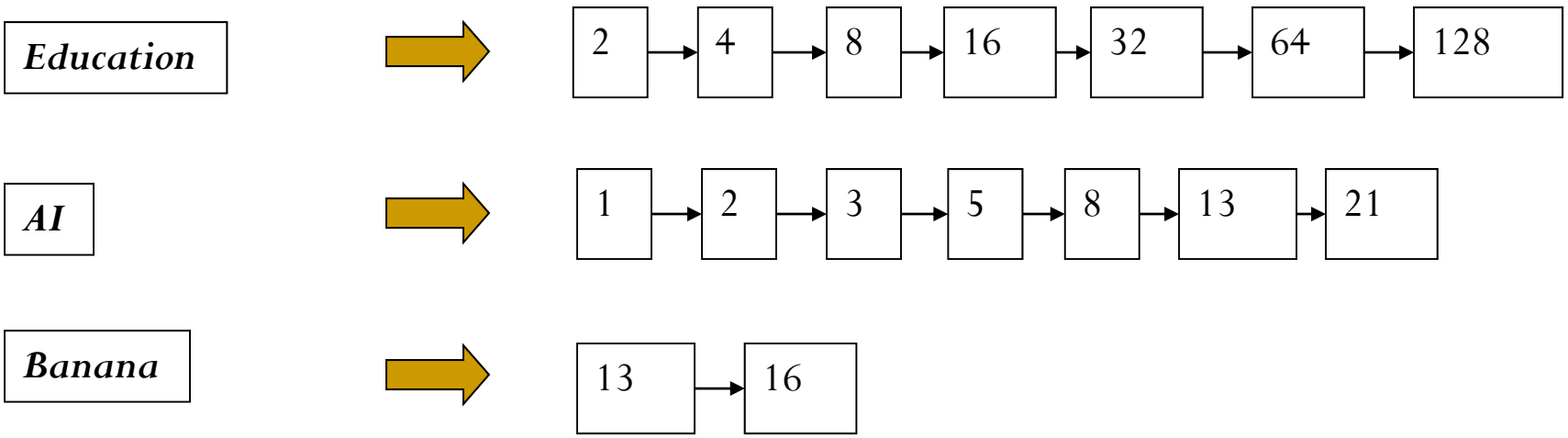
- tokenization
- casing
- normalization
- stemming
- lemmatization
- stop-word removal
- …

Analyze the impact

Modifier	Vocab size
simple tokenization	198K
improved tokenization	114K
number folding	108K
lowercasing	95K
stemming	91K
stop list	114K
num fold+lower+top	89K
all	68K

Assignment 2: Boolean queries

Building an inverted index

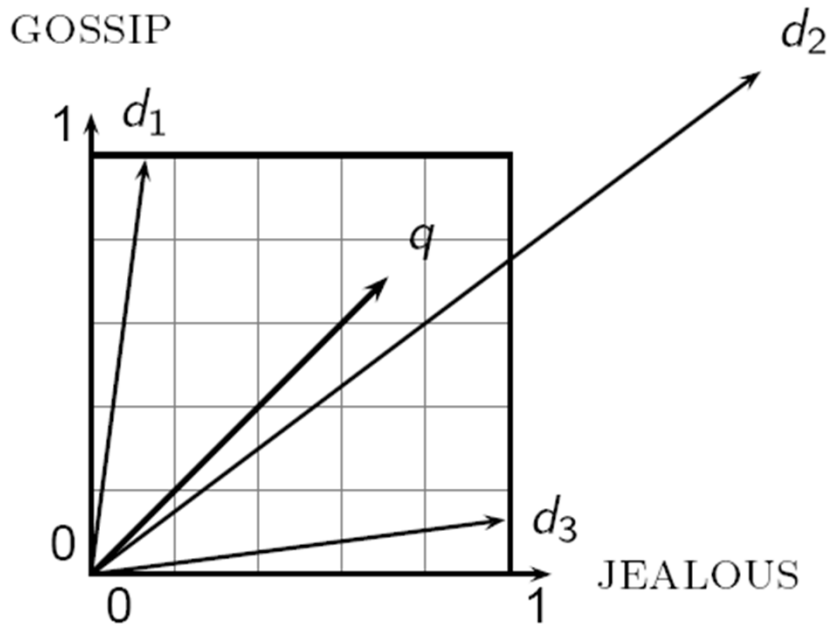


Support boolean query operations

- AND: intersection of two entries
- OR: union of two entries
- NOT: complement of an entry

Assignment 3: Ranked IR

- Query and document are represented as word count vectors
- Documents are ranked by the similarity between vectors
- Implement and explore possible variations on rankings

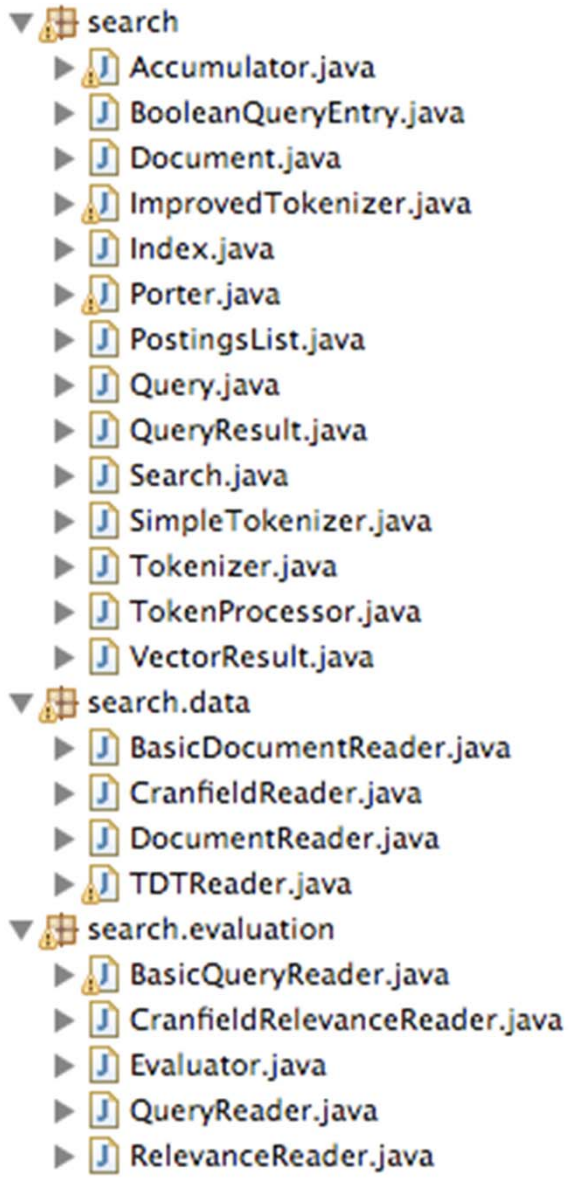


Assignment 4: Evaluation

- Implement different evaluation measures
  - precision
  - recall
  - mean average precision (MAP)
  - normalized discounted cumulative gain (NDCG)
  - …
- Analyze the results along many dimensions (ranking, text processing)

Count norm.	Term weighting	Length norm.		Precision@ 20	Recall@20	RPrecision	MAP
none	none	none		0.006	0.052	0.002	0.015
log	none	none		0.029	0.224	0.060	0.076
none	IDF	none		0.046	0.390	0.109	0.158
none	none	cosine		0.037	0.308	0.100	0.158
log	IDF	none		0.055	0.455	0.132	0.189
none	IDF	cosine		0.065	0.539	0.148	0.218
log	IDF	cosine		0.068	0.547	0.166	0.229

Code base after the assignments



- 27 classes
- ~2400 lines of code
- The students have been working on this code base throughout the entire course

Final project

Student requirements

- Implement an addition to the system related to class
- Work in teams
- Integrate with current codebase
- Evaluate
- Give write-up and presentation

Goals

- Explore advanced topics in IR that *interest* them
- "Real" software development
- An introduction to the research process

Final project goals: explore advanced topics

- GUI development
- Query optimization
- Query suggestion/rewriting
- Index compression
- Faster/approximate ranked retrieval
- Relevance feedback
- Result clustering
- Result classification
- Web crawling
- Position index/phrasal queries
- Wildcard/regex support
- Document segmentation
- Snippet generation
- Parallelized indexing
- Parallelized querying
- Multimedia search
- Document importance (e.g. PageRank)
- Advertising

Final project goals: real software development

Must work in teams

Must integrate with other teams

Use code repository (Git, SVN)

Final project size was over 40 classes and over 6000 lines of code

Final project goals: explore research process

Evaluate the impact of the addition

Write-up results in a research format

Peer review research write-ups

Give short, oral presentation summarizing project and results

Experiences

Pomona College (Fall 2009)

- 13 students
- Initial four assignments required roughly 2 weeks each
  - which were done in tandem with some written work
- Implemented in Java

Middlebury College (Fall 2012)

- 11 students
- Minor tweaks, but overall same structure

Experiences

Assignments

- Reinforce good coding practices
- Regular expressions
- Actual use for implementing your own linked list!
- Experiment design and data analysis

Final projects

- 6 different groups
- Integration wasn't too painful (though I did much of the coordination)
- Groups working on same topic (e.g. snippet generation) were able to compare results
- Evaluation was the most challenging part for most groups

- Supports full text queries with ranking
- Supports boolean queries
- Snippet generation
- 15K data set
- Works very efficiently

stupid Search

Clinton to American youth: don't inhale
... A **stupid** thing to do ... It is a **stupid** thing to ...

Treasury official blasts anti-crime budget cuts
... proposed cutbacks in his budget as **stupid**. ... This is **stupid** ...

Half Way Mark For GOP Scorecard Shows Points Drop
... **stupid** test to regulations. And if it is **stupid**, ... they are substitution appeals to fear for a discussion of the facts and the ...

Czech editor gets suspended sentence for racism
... Jews and half-breeds in December 1992 and started a campaign to brand the ... the **stupid** Czech nation merely sits back and watches this Jewish looting going ...

Monty Python's Flying Circus Celebrates 25 Years
... of Monty Python's Flying Circus would have to calm down a bit to be called ... weren't- Usually when they were doing something **stupid**, they also were being incredibly intelligent ...

Flexibility

Four assignments cover the basics of building a search engine

Many components can be easily altered based on the particular needs of the course

Final projects can be adapted to course particular course material

For larger classes

- interfaces well defined: grading scripts are very useful
- may have students to do less written work
- split final projects into different IR systems

EAAI

Educational Advances in Artificial Intelligence

http://eaai.cs.mtu.edu/

Resources

http://www.cs.middlebury.edu/~dkauchak/ir_project/

- Two variants of the course:
  - http://www.cs.middlebury.edu/~dkauchak/classes/f12/cs458/
  - http://www.cs.middlebury.edu/~dkauchak/classes/f09/cs160-f09/
- Course syllabus with slides
- 4 assignment descriptions (pdf and .tex)
- Final project description (pdf and .tex)
- Starter AND solution code in Java
  - Contact me for these
- Sample final project
  - demo and student final papers
- Sample grading scripts (though these should be considered in beta ☺)