605.744 Information Retrieval

Course Description

A multi-billion dollar industry has grown to address the problem of finding information on the Web. For instance, in January 1999, the Excite search engine was purchased for more than $6 billion (approximately the same purchase price as Ford Motor's acquisition of automaker Volvo in the same month). In 2004 Google's co-founders Sergey Brin and Larry Page received about $3 billion apiece due to the company's IPO; current valuations put Google's market capitalization at around $192 billion (1/2011). The technology underlying such enterprises is based on information retrieval – the field concerned with the efficient storage, organization, and retrieval of text. This course will cover both the theory and practice of text retrieval. Topics to be covered include automatic index construction, models of retrieval, textual representations, evaluation, web search, text classification, retrieval from noisy documents (speech/OCR), and multilingual retrieval. A practical approach will be emphasized and students will be assigned several programming projects to implement components of a retrieval system. Students will also be expected to give a class presentation based on an independent project.

Instructor

Paul McNamee

JHU Human Language Technology Center of Excellence
810 Wyman Park Drive (Stieff Building)
Baltimore MD 21211-2840 USA

Email:	paulmac@apl.jhu.edu	(best method)
Telephone:	+1 410 516-4831	(probably will get my voicemail)
Fax (shared):	+1 410 516-6700	(emailing PDF better)

Time and Location

The class will meet in room K-2, Mondays from 19:20 to 22:00.
I have no fixed office hours, but I can meet with students by appointment.

Textbook

1. C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

The text is excellent and new, and two of the authors (Manning and Schütze) co-authored one of the best available books on statistical natural language processing. There is a companion web site for the text and I believe the complete text is available online in PDF and HTML:

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

In the past I have used the book *Modern Information Retrieval* by Baeza-Yates and Ribeiro-Neto. While that book gave a decent treatment, individual chapters were written by different contributors and there is a lack of organization. Witten, Moffat, and Bell's *Managing Gigabytes* and Grossman and Frieder's *Information Retrieval: Algorithms and Heuristics* are other reasonable IR texts. In mid-2010 a book by Büttcher, Clarke, and Cormack was released by MIT Press. I have only

skimmed it, but it appears to be excellent (*Information Retrieval: Implementing and Evaluating Search Engines*).

Additional readings will supplement chapters from the text, as needed.

Communication

The best way to contact me is by electronic mail. When I am online I can often give you a quick response, especially for straightforward or short questions. Outside the classroom I will convey information to the class chiefly through email and with updates to the course web page.

Course web page

I will maintain a course web page that will be updated with homework assignments, handouts, and other useful information throughout the semester. The course page also contains a number of useful IR-related links. I recommend bookmarking the page and reviewing it occasionally.

http://apl.jhu.edu/~paulmac/ir.html

Grading Policy

(40%) Roughly six homework assignments. These are exercises focused on the lecture topics and programming is often involved.
(25%) In-class midterm exam
(25%) Independent project
(10%) In-class participation determined by, participating in discussions, quizzes (if any), short written summaries of papers, oral presentations (if any)

Late work may not be accepted, or may be accepted with penalty at my discretion; however, students who contact me about extenuating circumstances (prior to the date the work is due) will be given consideration. I find it helpful to return submitted assignments to the class promptly and I sometimes provide solutions or review problems in class – this is harder to do when not everyone has turned in work. I try to accommodate *forced* absences due to business travel, birth of a child, illness, or other reasons consistent with university policy. In case of an absence you can send completed work to me at my postal address or electronically in *a surpassingly-easy-to-print* format. My preferred order would be (1) a single PDF file; (2) multiple PDF files; (3) plain text files (especially for code); (4) PostScript; (5) MS Word.

I prefer assignments that are legible, on 8.5" x 11" paper, stapled, and if feasible, printed on two sides. Source code should be correct, readable (useful variable names, indentation, consistent style, straightforward logic), meaningfully organized, and containing suitable comments that explain what the code does or intends to do (which differs from how something is being implemented in the syntax of a given language). Code quality and readability is an important component of graded programs, though correctness is just a bit more important. Programs should be tested, and demonstrative working test cases or other evidence of correctness should be supplied in addition to source code. Sometimes I ask for these explicitly; other times you can choose test cases of your own design.

Programming Languages

I have had students successfully use any number of programming languages including Java, C++, Perl, Python, and others. HWs 1-3 will require use of language features including: data structures such as hash tables or binary trees; sorting (alphabetically or numerically); and use of binary file I/O. No graphical / GUI programming will be required for any assignment.

Integrity

Work for this class is expected to be the result of individual effort; however, unless explicitly prohibited, it is perfectly acceptable to make use of published examples and even source code from the literature or public domain – but only if attribution is given. Furthermore, while it is permissible to discuss the general nature of lecture material and assignments with your peers, this does not extend to discussing or revealing solutions to assigned problems or sharing source code. Students are expected to uphold the academic integrity of the university. Students using without reference, published material or copying the work (*i.e.,* particularly source code) of another individual will face consequences such as receiving a zero on the assignment and having the matter referred to the dean. Contact me if you have any questions about this policy, or if you have questions about a particular assignment.

References

A number of useful resources are listed on the course web page. These include leading periodicals and conferences, tools for NLP software, links to several search engines, and more.

Feedback

I welcome any feedback from students on how this course can be improved.

Tentative Outline

About the first 60% of the course focuses on foundational topics. Then for the next few weeks we will explore specialized topics. Finally, the last two weeks will be reserved for oral presentations of student projects. Chapters in the text tend to be short but dense (*i.e.,* denser means takes more time and effort to comprehend), averaging only about 20 pages in length. I do not plan to give lectures on the topics in Chapter 10 (XML Retrieval), or Chapters 16-18 (Clustering).

Date	Topic	Readings	Work assigned	Work due
1/31	Class Introduction, Unstructured Information Access, Tokenization	Chap 1-2 Paper by Lesk	HW 1:Term statistics	
2/7	Building Inverted Files, Dictionaries, Efficiency Techniques	Chap 3-5	HW 2: Inverted files	HW 1
2/14	Vector Space Model; Term Frequency	Chap 6-7 Salton and Buckley on term weighting		
2/21	Test Collections & Evaluation Relevance Feedback, Query Expansion	Chap 8-9, Economics of TREC	HW 3: Ranked retrieval	HW 2
2/28	Implementation Issues, Term Similarity Additional Ranking Methods	Chap 11-12	Project assigned	
3/7	Text Classification, Collaborative Filtering	Chap 13-15 Joachims paper	HW4: Classification (spam?)	HW 3
3/14	The Web, Web Search Engines	Chap 19-21	HW5: Commercial search	
3/21	NO CLASS - SPRING BREAK			Proj. Proposal
3/28	Midterm Exam			HW 4
4/4	Multilingual Retrieval, Translation Resources	Kishida paper	HW 6: CLIR	HW 5
4/11	Distributed Retrieval, Speech Retrieval NOTE: Last day to Withdraw / Switch to Audit			
4/18	Linguistic Approaches (POS / WSD)	Sanderson paper	HW 7: NLP&IR	HW 6
4/25	Information Extraction & Question-Answering			
5/2	Project Presentations			HW 7
5/9	Project Presentations			Project Report