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 $561 billion (1/2017) with other search providers (e.g., Baidu) at roughly 10 to 15% of that. The core technology underlying these 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, formal models for comparing documents and queries, 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 given several programming assignments to implement components of a retrieval system.

605.744 has no official prerequisites. However, as a graduate-level computer science course, the expectation is that you have attained the typical mathematical and computer programming skills, particularly algebra, discrete math, and the ability to write software in a modern programming language. If you have a weak programming background, you will experience difficulty on some assignments.

Instructor

Paul McNamee

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

Email:	paulmac@jhu.edu	(best)
Telephone:	+1 410 516-4831	(slower - often gets voicemail)

Time and Location

The class will meet in classroom K7 on Monday's 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.

In my opinion, the text is excellent, and two of the authors (Manning and Schütze) co-authored one of the best available books on statistical natural language processing. Chapters in the text are relatively short and average only about 20 pages in length. However, some sections may require

careful attention or re-reading to fully comprehend. There is a companion web site for the text and I believe the complete text is available online:

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

Some students prefer not to purchase a hardcopy version of the text, and it is quite possible to complete the course using the freely available digital versions (i.e., PDF, HTML) from that website. Others prefer to have a tangible, printed copy. Either choice is fine.

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. It also appears to be very good (*Information Retrieval: Implementing and Evaluating Search Engines*).

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

Communication

Usually the best way to contact me is by electronic mail. I can often provide 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://pmcnamee.net/ir.html

Grading Policy

Course grades will be assigned using letter grades with plus/minus modifiers as below:

Assuming project submitted	
Grade assigned	average >=
A+	97
A	93
A-	90
B+	87
B	83
B-	80
C+	77
C	73
and similarly	

If no project is submitted	
Grade assigned	average >=
B+	97
B+	93
B+	90
B+	87
B	83
B-	80
C+	77
C	73
and similarly	

Overall grades are based on:

(40%) Six homework assignments. These include programming assignments, and questions about concepts from the textbook and lecture material. Seven assignments will be given. The first three are compulsory, and you can submit any 3 of the final 4 assignments.
(30%) Two in-class exams
(25%) Independent project (** see paragraph below)
(5%) In-class participation determined by, participating in discussions, quizzes (if any), written summaries of papers (if any), oral presentations

To be eligible to receive a course grade of A-, or higher, students must complete an independent project. However, students may opt out of submitting a project. In this case, grades will instead be computed based on the average of the other components of the course work (*i.e.,* HWs, exams, participation), however, no grade higher than a B+ will then be assigned.

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 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) MS Office.

I prefer assignments that are legible, printed 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 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 for a specific assignment; other times you can choose test cases of your own design.

Programming Languages

In the past I have had students successfully use a variety of programming languages including Java, C++, Perl, Python, Lisp, among others. HWs 1-3 will require use of language features including: string manipulation, 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.

Academic Integrity

All students are required to read, know, and comply with the Johns Hopkins University Krieger School of Arts and Sciences (KSAS) / Whiting School of Engineering (WSE) Procedures for Handling Allegations of Misconduct by Full-Time and Part-Time Graduate Students available at: https://ep.jhu.edu/wseacademicmisconductpolicy

This policy prohibits academic misconduct, including but not limited to the following: cheating or facilitating cheating; plagiarism; reuse of assignments; unauthorized collaboration; alteration of graded assignments; and unfair competition. You may request a paper copy of this policy at this by contacting Mark Tuminello (Phone 410-516-2306; email mtumine2@jhu.edu).

Additional Remarks on Citations and Collaboration

I consider collaborations and discussions between students to be key ingredients to success in a graduate course. Thus in 605.744 it is permissible, and often even desireable for you to discuss the general nature of course content and assignments with your peers. However, the line between collaboration and cheating needs to be carefully delineated. When you submit work with your name on it for evaluation it must represent an individual effort by you alone. You should not discuss or reveal solutions to assigned problems with others, or share any unpublished source code.

This course requires you to write computer programs, and unless explicitly prohibited on an assignment, it is perfectly acceptable to make use of published examples and source code from the literature or public domain, but only if attribution is given. You must provide a citation for source code that you do not write yourself (e.g., URLs to websites, pointers to GitHub repos, Numerical Recipes in C, etc…).

Students using published material without citation, or who copy the work of another individual and represent it as their own (*i.e.*, including source code) may face consequences such as receiving a zero on the assignment, and having the matter referred to the associate dean. Contact me if you have any questions about course policies, or if you have questions pertaining to a particular assignment.

References

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

Feedback

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

Policy on Disability Services

Johns Hopkins University (JHU) is committed to creating a welcoming and inclusive environment for students, faculty, staff and visitors with disabilities. The University does not discriminate on the basis of race, color, sex, religion, sexual orientation, national or ethnic origin, age, disability or veteran status in any student program or activity, or with regard to admission or employment. JHU works to ensure that students, employees and visitors with disabilities have equal access to university programs, facilities, technology and websites.

Under Section 504 of the Rehabilitation Act of 1973, the Americans with Disabilities Act (ADA) of 1990 and the ADA Amendments Act of 2008, a person is considered to have a disability if c (1) he or she has a physical or mental impairment that substantially limits one or more major life activities (such as hearing, seeing, speaking, breathing, performing manual tasks, walking, caring for oneself, learning, or concentrating); (2) has a record of having such an impairment; or (3) is regarded as having such an impairment class. The University provides reasonable and appropriate

accommodations to students and employees with disabilities. In most cases, JHU will require documentation of the disability and the need for the specific requested accommodation.

The Disability Services program within the Office of Institutional Equity oversees the coordination of reasonable accommodations for students and employees with disabilities, and serves as the central point of contact for information on physical and programmatic access at the University. More information on this policy may be found at:
http://web.jhu.edu/administration/jhuoie/disability/index.html or by contacting (410) 516-8075.

Disability Services

Johns Hopkins Engineering for Professionals is committed to providing reasonable and appropriate accommodations to students with disabilities.

Students requiring accommodations are encouraged to contact Disability Services at least four weeks before the start of the academic term or as soon as possible. Although requests can be made at any time, students should understand that there may be a delay of up to two weeks for implementation depending on the nature of the accommodations requested.

Requesting Accommodation

New students must submit a Student Request for Accommodation form along with supporting documentation from a qualified diagnostician that:

- Identifies the type of disability
- Describes the current level of functioning in an academic setting
- Lists recommended accommodations

Questions about disability resources and requests for accommodation at Johns Hopkins Engineering for Professionals should be directed to: Mark Tuminello, Disability Services Coordinator (Phone: 410-516-2306; Fax: 410-579-8049; Email: mtumine2@jhu.edu or ep-disability-svcs@jhu.edu).

Tentative Outline

About the first two-thirds of the course focuses on foundational topics, generally following the course text. The last third covers more specialized topics. I do not plan to give lectures on the topics in Chapter 10 (XML Retrieval), or Chapters 16-18 (Clustering).

The dates listed in the table below could change due to weather, a cancelled class, or other factors. Consider dates mentioned in class and posted to the website as authoritative rather than this table.

Date	Topic	Anticipated Readings	Work assigned	Work due
1/23	Class Introduction, Unstructured Information Access, Tokenization	Chap 1-2 Paper by Lesk	HW 1:Term statistics	
1/30	Building Inverted Files, Dictionaries,	Chap 3-4	HW 2: Inverted files	HW 1
2/6	Efficiency Techniques	Chap 5		
2/13	Vector Space Model; Term Frequency	Chap 6-7 Salton/Buckley paper	HW 3: Ranked retrieval	HW 2
2/20	Test Collections & Evaluation Relevance Feedback, Query Expansion	Chap 8-9, Economics of TREC	Project assigned	
2/27	Implementation Issues, Term Similarity Additional Ranking Methods	Chap 11-12		
3/6	Exam #1			HW 3
3/13	Text Classification, Collaborative Filtering	Chap 13-15 Joachims paper	HW4: Classification	Proj. Proposal
3/20	***NO CLASS – SPRING BREAK***			
3/27	The Web, Web Search Engines Mar 30: Last day to Withdraw / Switch to Audit	Chap 19-21	HW5: Commercial search	HW 4
4/3	Multilingual Retrieval, Translation Resources	Kishida paper	HW 6: Multilingual	
4/10	Distributed Retrieval, Speech Retrieval			HW 5
4/17	NLP and IR	Sanderson paper	HW 7: NLP&IR	HW 6
4/24	Exam #2 (cumulative)			HW 7
5/1	Project Presentations			Project Report