JOHNS HOPKINS
UNIVERSITY

EN.605. 744 section 81

Information Retrieval

Course Information

Information Retrieval

EN.605. 744 81 (3.0 Credits)
Fall 2022 [AE Fall 2022]

Description

A multibillion-dollar industry has grown to address the problem of finding information. Commercial search engines are based on information retrieval: the efficient storage, organization, and retrieval of text. This course covers both the theory and practice of text retrieval technology. Topics include automatic index construction, formal models of retrieval, Internet search, text classification, multilingual retrieval, question answering, and related topics in NLP and computational linguistics. A practical approach is emphasized and students will complete several programming projects to implement components of a retrieval engine. Students will also give a class presentation based on an independent project or a research topic from the IR literature. Prerequisite(s): 605.202 Data Structures or permission of the instructor

Department: PE Computer Science
College: Engineering and Applied Science Programs for Professionals

Expanded Course Description:

The only formal prerequisite for 605.744 is a course in Data Structures. However, as a graduate-level computer science course, the expectation is that you have acquired some foundational mathematical and computer programming skills, particularly algebra, discrete math, and the ability to write software in a modern programming language. Core concepts required to complete programming assignments in the course include: working with standard data structures (e.g., arrays, trees, hash tables), manipulating strings, sorting data, and performing binary file input/output operations. If you have a weak programming background, you will experience difficulty completing some assignments.

Instructor

Paul McNamee
✉ mcnamee@jhu.edu

Course Location:

Online

Communication Policy:

For clarifications about assignments or due days, the "Ask the Prof" section in the Discussion Forum is the preferred place to ask questions. For individual issues email is preferred. I am often able to respond quickly to short, focused messages; long, multi-topic messages take more time to read, research, and compose responses to. As I do receive a lot of daily email, it is helpful if you use include the course number (605.744) and a meaningful description in the subject line of the message

Office Hours:

This course will use Zoom to facilitate weekly, synchronous office hours. You are not required to participate in office hours; however, you may nd them useful for receiving more timely answers to questions related to the course content and assignments. During the first week of the semester I will conduct a student survey to determine the best day and time of the week to schedule office hours. Once the day and time have been determined, I will set up office hours links within the Calendar. Students can click that link to access Zoom and participate in the sessions.

Course Structure:

The course materials are divided into modules which can be accessed in Canvas by clicking **Modules** on the left menu. A module will have several sections including the overview, content, readings, discussions, and assignments. You are encouraged to preview all sections of the module before starting. Most modules run for a period of seven (7) days, exceptions are noted on the **Course Outline** document. You should regularly check the **Calendar** and **Announcements** for assignment due dates.

Modules will start on the Tuesday of each week, and finish on the following Monday. Thus assigned work will generally be due on Monday evenings.

Course Topics:

The topics covered in the course are also available in Canvas in the **Course Outline**.

- Course Overview; Boolean Retrieval
- Index Construction
- Efficiency Issues
- Vector Space Models
- Evaluation
- Relevance Feedback
- Advanced Retrieval Models
- Text Classification
- Multilingual IR
- Web Search
- Distributed Processing
- Multimedia Processing
- NLP & IR

Course Goals:

In this course we explore how computers represent and retrieve information from large collections of text documents. This is exactly the technology used by large Internet search and social media companies such as Amazon, Apple, Baidu, Google, Facebook, Microsoft, Twitter, Yelp, etc... You will learn about inverted file construction (i.e., algorithms for efficiently indexing text), and methods for accurate and efficient retrieval of relevant documents. Using this knowledge, you will write computer programs that illustrate true-to-practice components of retrieval engines. The second half of the course introduces specialized topics such as text classification, multilingual retrieval, and web search.

Course Learning Outcomes (CLOs):

No Course Learning Outcomes are available for this course.

Required Text and Other Materials

Textbooks:

There is one required textbook:

Manning, C.D., Raghavan, P., & Schütze, H. (2008), Introduction to Information Retrieval. New York, NY: Cambridge University Press.

ISBN-10: 0521865719
ISBN-13: 978-0521865715

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. Supplementary materials and **an online version of the text is available in digital form at the companion web site provided by the authors at**: 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 version (i.e., PDF, HTML) from that website. Others prefer to have a tangible, printed copy. Either choice is fine.

Access to textbooks via the JHU Libraries:

EP students may access electronic versions of textbooks through the Sheridan Libraries. Instructions on how to search for available textbooks are accessible through this link: Browse Electronic Textbook Instructions

Technical Requirements:

You should refer to General Technical Requirements for guidance on system requirements. Access support resources from the **Help** menu if you encounter any technical issues.

Evaluation and Grading

Student Coursework Requirements:

Student Coursework Requirements:

605.744 is an upper-level graduate computer science course, and completing the work for each module can take over 10 hours a week to complete, depending on the module and your background. An approximate breakdown of the main components would be: (a) reading the assigned materials (2 – 2.5 hours per week); (b) listening to the audio annotated slide presentations (2 – 3 hours per week); (c) online interactions (< 30 minutes); (d) completing a short problem set (2 hours); (e) working on programming assignments or the class project (4+ hours per week).

Readings and Reviewing Video Lectures

The assigned readings and prepared video lectures are carefully chosen to help you succeed in the course. They are not directly evaluated, but reading and reviewing this content is crucial for mastering the course material.

Programming Assignments (30%)

There will be five (5) programming assignments. The first three build upon each other towards the creation of a true-to-practice basic information retrieval system. You can use any programming language you want on the assignments (see the section on Programming Languages below).

Code style and readability is an important component of graded software programs. Nonetheless, the major emphasis will be on correctness.

Generally programs should be validated using demonstrative test cases or some other evidence of correctness should be supplied. On some assignments I ask for specific tests cases; otherwise you are free to use examples of your own choosing. If a program is not working 100% correctly, you can still provide examples or an explanation of what works correctly and what does not.

Short Problem Sets (20%)

There will be approximately 10 short problem sets. These will consist of questions that facilitate learning of terminology, basic concepts, and the ability to perform calculations. When calculating course grades your lowest problem set score will be dropped. While I think there is educational benefit in doing all of the problem sets, this means that you could choose to not submit one problem set with only minimal effect on your grade. That score would be a zero, which would be ignored in the average.

Exam (15%)

There will be one open-book exam during the next-to-last module of the course. The exam will be focused on short response questions and problem computation. It will require analysis and calculation, and is not solely a regurgitation of facts.

Scholarly Engagement / Online Discussion (15%)

There are three components to "Scholarly Engagement": discussion forum participation, one research paper summary, and reviewing other student's class projects.

You are expected to participate in the course discussion forum. While I will sometimes pose questions and participate in the forum, the class will actually learn the most from student-initiated questions and responses to one another’s posts. At three (3) roughly evenly spaced times throughout the course (i.e., covering approximately 4 modules), I will assess your contributions to the online forum using the following criteria, and provide feedback.

	Excellent (2 pts)	Satisfactory (1 pt)	Below Expectations (0 pts)

Demonstrates knowledge of content	Posts and responses consistently demonstrate strong knowledge of assigned readings and video lectures. Responses may clarify points to aid learning of others.	Posts and responses demonstrate knowledge of assigned readings and video lectures. Reponses do not confuse others.	Posts and responses suggest an incomplete understanding of assigned readings and video lectures. Responses may confuse others.
Critical thinking; Insightfulness	Student actively stimulates and sustains intellectual inquiry by asking thoughtful questions. The student recognizes the accuracy, logic, relevance, or clarity of statements. Responses reflect original thinking.	Student relies on the momentum of the group to motivate inquiry. Posts may be repetitive with previous comments. Positions taken are not strongly justified. Responses show a mixture of original thinking and contributions from others.	Student accepts ides of others without much thought. Little or no original thinking is present. Few ideas are contributed to discussions.
Frequency	Averages 1.5 to 2+ contentful posts per week. Activity is spread across multiple modules.	Averages approximately 1 contentful post per week.	Averages 0 to 1 contentful post per week.
Clarity & Conciseness	Posts and responses are focused. Points are consistently clear, and can be understood by others. Distracting or extraneous comments are minimal. Grammar and spelling are excellent.	Posts and responses are usually focused. Posts show evidence of attempting to be clearly understood by others.	Posts and responses are not generally focused. Posts may not be understood by others, without significant effort to comprehend meaning. Tangential material is present, or the quality of the writing may be poor.
Utility / Fosters learning and engagement among the class	Posts elicit responses and reflections from other learners. Responses build upon or synthesize multiple opinions to improve discussions. Posts are likely to contain content of interest to many participants.	Posts attempt to elicit responses and reflections from other learners. Responses attempt to improve the discussion. Posts are likely to contain content of interest to some participants.	Posts do not appear to elicit responses or learning from others. Responses do not take the discussion deeper. Posts are unlikely to be very interesting to others.

Unlike some online courses, in 605.744 there are no “required” discussion topics. While I will sometimes provide questions or comments at the start of a module to suggest a topic or encourage discussion, it is not intended for these to constrain the online discussion. You are strongly encouraged to ask your own questions. Some uses for the discussion forum are to ask for clarification of difficult material, to challenge presented material, to discuss a previous problem set, or to relate the current module to other areas (e.g., topics from previous course modules, contemporary topics in computer science; activities by large search companies; topics in related JHU EP courses, etc…).

Separate discussion areas are provided for:

- Ask the Prof / Administrative Issues: Discussion about course logistics (e.g,. when assignments are due; clarifications about a homework assignment) that multiple students might benefit from. You are also welcome to contact me by email with individual concerns.
- Main Discussion Area: Most posts will happen here. These threads are focused on the current module's technical content. But it is also okay to post to previous module threads, especially when relating new knowledge to previous topics.
- Research Paper Summaries: This area is reserved for you to post your reviews of assigned scientific articles (see below), and for other students to pose questions about that article.
- Class Projects: Later in the semester a forum for class projects will be added. This is where students submitting projects will post an abstract and video of their project presentation.

You will be assigned one research paper from the scientific literature to review. A 1 to 2 page **summary and critique of the paper** will be shared with the class. The student providing the summary is expected to respond to basic questions about the paper in the "Reviews" area of the discussion forum. Details about the papers we will review and the expectations for the summaries will be provided several weeks into the course.

The final module is devoted to class project (details below). The final component of scholarly engagement is **brief review of other student's class projects**.

Individual Class Project (20%)

This is an opportunity for you to investigate a topic of your own choosing. Projects typically involve working with a dataset and conducting an experiment or building a proof-of-concept system. Listed below are a few examples of past student projects:

- Analyzing police crime reports and classifying narratives by type of criminal activity
- Exploring methods to compress indexes using document-identifier reassignment
- Extraction of fields from Craigslist apartment rentals (i.e., automatic identication of the number of rooms, monthly rent, location, if smoking is allowed, etc...)
- Attempting the Netflix Challenge
- Predicting attributes of document authorship (e.g., author gender, century of authorship, or who authored a particular document) Making phrasal querying fast using nextword indexing
- Predicting stock price movement using open source data (e.g., Twitter streams, SEC lings)

More detailed information about the project will be communicated about 4 weeks into the course. The primary deliverables are a written report (approx. 5-8 pages) and a short video presentation (approx. 10 minutes of voice-annotated slides or video). The last week of the course will be devoted to class projects.

Grading Policy:

Course grades are based on the following components:

- (30%) Programming Assignments. There will be five (5) assignments. The first three build upon each other towards the creation of a true-to-practice basic information retrieval system.
- (20%) Short Problem Sets. There will be approximately 10 problem sets. Questions will be in a short answer format to assist learning of terminology, basic concepts, and the ability to perform calculations.
- (15%) Scholarly Engagement. Students are expected to participate in course discussion fora. Participation will be measured by clarity, insightfulness, utility, appropriateness, frequency, and etiquette of posts. There is also a research paper summary that will be shared with the rest of the class and a review of other student's class projects.
- (15%) Final Exam. There will be one exam near the end of the course, which will be focused on problem computation, and higher-level material.
- (20%) Individual Class Project. This is an opportunity for students to investigate a topic of their own choosing. Projects typically involve working with a dataset and testing a theory or building a proof-of-concept system. Students may use computer programs they have written for this class, or open-source software and tools. Submitting a project is required to attain a grade of A- or higher, but the project is optional if not aiming for a grade above B+. Students not submitting a project will have their grades based on the average of the other course assessments, and will not be eligible for a grade above B+.

Course grades will be assigned using letter grades with plus/minus modifiers as specified in the table below. To be eligible to receive a course grade of A-, or higher, you must complete a class project. However, you may choose to opt out of submitting a project. In this case, grades will instead be computed based on the average of the other components of course work, and no grade higher than a B+ will be assigned.

With a project		Without a project	
Grade Assigned	average >=	Grade Assigned	average >=
A+	97	B+	97
A	93	B+	93
A-	90	B+	90
B+	87	B+	87
B	83	B	83
B-	90	B-	80
C	70	C	70
F	otherwise	F	otherwise

A grade of A indicates achievement of consistent excellence and distinction throughout the course—that is, conspicuous excellence in all aspects of assigned work.

A grade of B indicates work that meets all course requirements on a level appropriate for graduate academic work.

Policies

Course Policies:

Submitting Individual Assignments

Your name, the course number, and a title (e.g., “Problem Set #4”) should be present on the first page of each submission.

Work for the class, such as Programming Assignments (including source code) and the Short Problem Sets **should be submitted as a single PDF file**. Contact me if this stipulation creates a significant difficulty for you.

Policy on Late Work

Assignments are due according to the dates posted on the Canvas course site. You may check these due dates in the Course Calendar or by viewing the Assignments in the corresponding modules. Grades are generally available in Canvas approximately one week after due dates.

I find it helpful to return graded work to the class promptly and I sometimes provide solutions or review problems in the discussion forum or during office hours. This is much harder to do when not everyone has turned in their work. Thus, submitting assignments late is discouraged.

All work must be submitted by the last day of the term. However during the semester I will generally accept late work for up to one week. Assignments that are submitted 1 to 2 days late will be penalized 10%. Submissions 3 to 4 days late will be penalized 25%. Submissions up to one week late will be penalized 40%. No work will be accepted beyond a week after the due date -- a grade of zero will be assigned instead. Generally speaking, it is better to submit something incomplete or imperfect on time or a day late than to let it go longer.

In extraordinary circumstances you should contact the instructor. Reasonable accommodation will be made for an extended hospitalization or other serious situations. However documentation is expected (e.g., signed note on letterhead with printed contact information of the physician, etc...).

In some situations withdrawing from the course (no permission needed) or taking an incomplete (permission required) are appropriate. You are encouraged to speak with the instructor and/or your advisor if you are considering pursuing either course of action.

Programming Languages

In the past I have had students successfully use a variety of programming languages including Java, C++, Perl, Python, Lisp, and many others. The first three programming assignments require use of language features including: data structures such as arrays, binary trees, and hash tables; string manipulation, sorting (alphabetically or numerically); and use of binary file I/O. No graphical / GUI programming will be required on any assignment.

Web Resources

I maintain a course-related web page with a number of resources that you may nd useful for the class:

http://pmcnamee.net/ir.html

Resources on this webpage include links to other IR text books, archives of scientific papers, links to open source software, and text retrieval datasets. You should browse the page a couple of times during the semester, particularly to see if there is a resource that would help you on a homework assignment or the class project. If you know of a software or data resource that would be helpful to include there, please write to me about it.

Additional Comments About Academic Integrity for 605.744

Collaborations and discussions between students are key ingredients to success in a graduate course. It is permissible, and often even desirable 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. You should not discuss or reveal solutions to assigned problems with others, or share any unpublished source code. When you submit work with your name on it for evaluation it must represent an original, individual effort by you alone.

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 or text that you do not write yourself (e.g., URLs to websites, pointers to GitHub repos, Numerical Recipes in C, Stack Overflow, etc...). Contact the instructor if you have any questions about this policy.

Additional Resources:

Personal Wellbeing
If you are struggling with anxiety, stress, depression or other mental health related concerns, please consider connecting with the Johns Hopkins Student Assistance Program (JHSAP). If you are concerned about a friend, please encourage that person to seek out our services. JHSAP can be reached at 443-287-7000 or https://jhsap.org/

Tutoring Website

Johns Hopkins Engineering for Professionals offers a tutoring connection network that allows students to connect with other Johns Hopkins Engineering students or alumni for tutoring services. This service allows students to search a list of courses to “Find a Tutor” or complete a profile to “Become a Tutor.” More information about this service can be found on the tutoring website (https://tutor.ep.jhu.edu/).

Deadlines for Adding, Dropping and Withdrawing from Courses

Students may add a course up to one week after the start of the term for that particular course. Students may drop courses according to the drop deadlines outlined in the EP academic calendar (https://ep.jhu.edu/student-services/academic-calendar/). Between the 6th week of the class and prior to the final withdrawal deadline, a student may withdraw from a course with a W on their academic record. A record of the course will remain on the academic record with a W appearing in the grade column to indicate that the student registered and withdrew from the course.

Academic Misconduct Policy

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.

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. Course materials (old assignments, texts, or examinations, etc.) should not be shared unless authorized by the course instructor. Any questions related to this policy should be directed to EP's academic integrity officer at ep-academic-integrity@jhu.edu.

Students with Disabilities - Accommodations and Accessibility

Johns Hopkins University values diversity and inclusion. We are committed to providing welcoming, equitable, and accessible educational experiences for all students. Students with disabilities (including those with psychological conditions, medical conditions and temporary disabilities) can request accommodations for this course by providing an Accommodation Letter issued by Student Disability Services (SDS). Please request accommodations for this course as early as possible to provide time for effective communication and arrangements.

For further information or to start the process of requesting accommodations, please contact Student Disability Services at Engineering for Professionals, ep-disability-svcs@jhu.edu.

Student Conduct Code

The fundamental purpose of the JHU regulation of student conduct is to promote and to protect the health, safety, welfare, property, and rights of all members of the University community as well as to promote the orderly operation of the University and to safeguard its property and facilities. As members of the University community, students accept certain responsibilities which support the educational mission and create an environment in which all students are afforded the same opportunity to succeed academically.

For a full description of the code please visit the following website: https://studentaffairs.jhu.edu/policies-guidelines/student-code/

Classroom Climate

JHU is committed to creating a classroom environment that values the diversity of experiences and perspectives that all students bring. Everyone has the right to be treated with dignity and respect. Fostering an inclusive climate is important. Research and experience show that students who interact with peers who are different from themselves learn new things and experience tangible educational outcomes. At no time in this learning process should someone be singled out or treated unequally on the basis of any seen or unseen part of their identity.

If you have concerns in this course about harassment, discrimination, or any unequal treatment, or if you seek accommodations or resources, please reach out to the course instructor directly. Reporting will never impact your course grade. You may also share concerns with your program chair, the Assistant Dean for Diversity and Inclusion, or the Office of Institutional Equity. In handling reports, people will protect your privacy as much as possible, but faculty and staff are required to officially report information for some cases (e.g. sexual harassment).

Course Auditing

When a student enrolls in an EP course with “audit” status, the student must reach an understanding with the instructor as to what is required to earn the “audit.” If the student does not meet those expectations, the instructor must notify the EP Registration Team [EP-Registration@exchange.johnshopkins.edu] in order for the student to be retroactively dropped or withdrawn from the course (depending on when the "audit" was requested and in accordance with EP registration deadlines). All lecture content will remain accessible to auditing students, but access to all other course material is left to the discretion of the instructor.