ITI 471: Introduction to Information Retrieval Systems Development

Winter 2015 Online Course

Instructor: Dr. Chirag Shah
Email: chirags@rutgers.edu
Phone: (848) 932-8807
Office: Room 334 in SC&I
Office hours: By appointment
Website: http://comminfo.rutgers.edu/~chirags/teaching/2015_winter/ITI471/

Course Description

Tools for organizing and accessing information have become indispensable. It is critical, therefore, to understand their design and operational foundations. In this course students will have an opportunity to learn about search engines, web crawling, and search interface technologies based on hands-on experience and with a focus on techniques that can be used to access, retrieve, organize, and present information. Students will work with practical developmental tools and learn relevant concepts through experimentation. For instance, students will employ an open source search engine and learn about indexing, retrieving, and ranking techniques. This course is development-focused, but does not involve much programming. The student, however, is expected to have at least some prior programming experience.

Prerequisites

The students are expected to have a previous exposure to some programming (C, Java, Perl, Python, or PHP). Basic programming experience acquired in an introductory programming course is recommended.

Course Materials

The textbook for this course will be provided by the instructor as a free ebook. Various articles and book chapters will be assigned based on the content being covered. Students must have access to a computer with Internet connection (broadband recommended).

Course Goals

By the end of the course, students should be able to:

1. Explain various theoretical issues with IR systems, such as indexing and ranking.
2. Use and/or implement various search engine services such as stemming, indexing, retrieval, and ranking.
3. Build a set of Web crawlers to collect data from the Web.
4. Demonstrate how various components of online search service for structured and unstructured information could be integrated in one solution.

Instructional Methods

The structure of the course will be lab-like classes with power point presentations, video podcasts, online discussions, and hands-on exercises. The assignments and evaluations will be based on practical projects that the students will do by themselves or in small groups. Typically, a new assignment will be given with each class with 2 days to finish and submit.

Quizzes, Tests, and Final Examination

1. Quizzes will be given three times during the course on material covered from the lectures after the previous quiz. They will consist of short answer and identification, and given a day before the class. Since quiz material is immediately discussed following their administration, quizzes cannot be made up if missed.
2. Small assignments based on the material covered in the class are scheduled throughout the session.
3. Final project will be due at the end of the session.

Course Evaluation

- A (91-100%): **Outstanding and excellent work** of the highest standard, mastery of the topic, evidence of clear thinking, good writing, work submitted on time, well organized and polished.
- B+ (85-90%): **Very good work**, substantially better than the minimum standard, very good knowledge of the topic; error free.
- B (80-84%): **Good work**, better than the minimum standard, good knowledge of the topic.
- C+ (75-79%): **Minimum standard work**, adequate knowledge of the topic.
- C (70-74%): Work barely meeting the minimum standard, barely adequate knowledge of the topic; errors.
- D (65-69%): Writing not up to standard, disorganized, many errors
- F (< 65%): Unacceptable, inadequate work
- T: Temporary.

The final grade will be weighted based on the following: Assignments: 30%, Quizzes: 20%, Project: 50%.

Examinations

The quizzes will consist of multiple choice and short answer. The assignments will involve a little bit of programming and usually consist of extending something that we did in the class. The final project will require you to use several concepts learned during the course and put them in practice with some real-life problem. To prepare for the exams, you should carefully review the lesson objectives, handouts, and assigned readings.

Academic Integrity

Academic integrity means, among other things:

- Develop and write all of your own assignments.
- Show in detail where the materials you use in your papers come from. Create citations whether you are paraphrasing authors or quoting them directly. Be sure always to show source and page number within the assignment and include a bibliography in the back.
- Do not look over at the exams of others or use electronic equipment such as cell phones or MP3 players during exams.
- Do not fabricate information or citations in your work.
- Do not facilitate academic dishonesty for another student by allowing your own work to be submitted by others.

If you are doubtful about any issue related to plagiarism or scholastic dishonesty, please discuss it with the instructor. At the instructor's discretion, work presented in this course is subject to verification of originality, using http://www.turnitin.com/. The consequences of scholastic dishonesty are very serious. Rutgers' academic integrity policy is at http://ctaar.rutgers.edu/integrity/policy.html. An overview of this policy may be found at http://academicintegrity.rutgers.edu/resources-for-students. Multimedia presentations about academic integrity may be found at http://www.scc.rutgers.edu/douglass/sal/plagiarism/intro.html

How to Succeed in this Course

Successful students will follow the class regularly by subscribing to the podcast and participating in online discussions. If you know you must miss a class, please contact the instructor in advance, either by phone or email. You can obtain assignments or notes from a fellow classmate or from the instructor. In the case of a prolonged absence from class, you should schedule an appointment with the instructor so we can discuss the course material and concepts that you missed.

Successful students will pay close attention to the course goals and objectives, because they will help you master the course material. If you have any questions about any of the objectives, please ask the instructor. Questions are encouraged on online discussion board for clarification. Remember that you're probably not the only one in the class with the same question. If you have questions about material from previous classes, please email me prior to the next class session, and I'll address your question at the beginning of the class session, prior to any quizzes.

Successful students will talk to their classmates about the course material. You will find that they can help you understand many complex issues.

Successful students will be prepared before the class with assigned readings for that class. This will help you comprehend the material for that class better. Regular assignments will also be given at the end of each class. Doing these assignments and turning them on time (typically before the next class), will help you obtain higher-order learning goals for this course.

Course Objectives

These are stated within the table for the class sessions and at the beginning of each class session. Use these as a study guide and as a checklist for your progress during the semester.

Professionalism

1. Access the class material promptly and on time.
2. Respect yourself, classmates, and the instructor.
3. Participate in online discussion.
4. Display preparedness for class through completing reading assignments.
5. Present content knowledgeably with supported reasoning.

Schedule

#	Date	Topics and Readings	Objectives	Quizzes & Assignments
1.	12/23/2014	Introduction • Introduction to the course • Setup your computer with necessary tools • Overview of some UNIX commands and utilities • Structured data access from a MySQL database Reading: UNIX Primer, A longer UNIX Primer	• Familiarize with basic terminology of a search system environment. • Practice accessing structured data.	**Assignment-1** Due on: 12/24/2014
2.	12/26/2014	IR with MySQL and Text Files • Structured data access and display in a webpage • Unstructured data access from (1) MySQL tables and (2) text files Reading: Getting Started with MySQL Reading: HTML Tutorial, HTML Forms and Input Reading: PHP introduction, installation, syntax, variables.	• Practice accessing and processing structured data using MySQL. • Demonstrate how textual data can be accessed from MySQL as well as flat-files.	**Assignment-2** Due on: 12/27/2014 **Quiz-1** Available for 12/27/2014
3.	12/29/2014	Indexing • Effective indexing of text documents • Basic understanding of information retrieval model • Work with Lemur Toolkit Reading: The Anatomy of a Large-Scale Hypertextual Web Search Engine Reading: Overview of Lemur	• Describe a general model of information retrieval. • Explain the importance of indexing, stemming, and stopwords removal. • Demonstrate how these processes are executed in a typical search environment. • Use Lemur to index a set of documents.	**Assignment-3** Due on: 12/30/2014
4.	12/31/2015	Query Processing and Retrieval • Represent query "by hand" and	• Process a text query for matching it	

		then using Lemur • Retrieve documents Reading: Google Basic Search Guidelines, Google Search Operators	with an indexed collection. • Retrieve a set of relevant documents matching the query using vector space model.	**Assignment-4** Due on: 01/01/2015 **Quiz-2** Available for 01/02/2015
5.	01/05/2015	Retrieval Models • Vector space, Boolean, and Language model Reading: Boolean retrieval [PDF] by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze Reading: A language modeling approach to information retrieval by Jay Ponte and W. Bruce Croft	• Demonstrate use of various retrieval models. • Describe the pros and cons of these models.	**Assignment-5** Due on: 01/06/2015
6.	01/07/2015	Building UI and Web Crawling • Develop a functional and user-friendly UI for search. • Web crawling with "wget" • Indexing the crawled data Reading: Focused crawling: a new approach to topic-specific Web resource discovery by Soumen Chakrabarti, Martin van den Berg, and Byron Dom Reading: Random Web Crawls [PDF] by Toufik Bennouas and Fabien de Montgolfier.	• Develop a basic web-based UI for search • Collect documents using Web crawling • Demonstrate how crawled data can be indexed and made retrieval-ready.	**Assignment-6** Due on: 01/08/2015 **Quiz-3** Available for 01/09/2015
7.	01/12/2015	Evaluation • Recall and precision measures in IR • TREC evaluation Reading: Evaluation of Evaluation in Information Retrieval [PDF] by Tefko Saracevic	• Demonstrate ways to evaluate retrieval performance. • Employ TREC measures to evaluate and report retrieval effectiveness of an IR system.	**Quiz-4** Available for 01/13/2015
8.	01/14/2015	Conclusion • Course summary • Project discussion		**Project** Due on: 01/16/2015

Project Guidelines

As a part of this course, you are required to do a project demonstrating several of the concepts covered during the course. Before starting the actual project, you will be providing a proposal (details below), and upon the approval of the instructor, you can commence the project. This project will be due on **Fri, January 16**. Early submissions are encouraged. The project submission should be in form of (1) source files, (2) online working site, and (3) a report. The project will be tested with Firefox browser. Remember, the project carries 50% of the grade for this class. Any document (proposal or final report) submitted should have the following this format: PDF, 12pt Times family fonts, single-spacing, 1" margins.

Project proposal (Due: Fri, January 2)

You need to submit a brief article (1-2 pages) proposing the project that you want to do. This article should have

- Problem description (what is it that you're trying to address - e.g., provide search functionality for a set of websites)
- Your approach/design (how would you collect documents, indexing parameters, user interface details)
- What is unique about your approach? In other words, if you're trying to sell this, why would anyone bother investing in this instead of using an off-the-shelf product?

- Mockup and/or a brief description of what the outcome will be.

Final project (Due: Fri, January 16)

You are allowed to pick your application/domain. You can even use some existing project, but you need to specify how much is already done. Your final project MUST have the following components.

- A text collection (unstructured data) of a "reasonable" size (preferably collected by a crawl).
- A user interface with proper navigational tools so that even a naive user can utilize it with ease.
- A search facility that allows one to perform full-text search in the collection.
- A way to visualize the information (rank list, clusters, tag-cloud, etc.).
- A mechanism to log all the user interactions.

For text processing (mostly search-related), you are required to use Lemur/Indri. For structured databases, MySQL is recommended. For building UI, PHP is recommended.
You may *optionally* have

- Advance search.
- Dynamic UI to enhance user experience.
- Interactive session support.
- Meshing with a live Web application.
- Clever use of CSS and Javascript for validation and site configuration.
- Evaluation of processes/results with appropriate measures.

Finally, this should be the kind of work that you feel comfortable (and proud) demonstrating and listing on your portfolio. At the least, your project will be showcased (at your discretion) on the course website. Your final submission should include all the source files (including the parameter files), a link to the online working site, and a brief article documenting the project. This document should have

- Introduction - what is this project about and what it does/serves. (1/2 to 1 page)
- Design details (may include figures). Explain your decisions behind certain design choices. (1-2 pages)
- Usage scenario (may include screenshots). (1-3 pages)
- Known issues and future work. (1-2 pages)
- License. An appropriate Creative Commons License is recommended.

Grade rubric

- Proposal (15 points)
- Source and parameter files (10 points)
- Working site
    - Proper use of tools (10 points)
    - Ease of use of the interface (10 points)
    - Functionality (does it serve what it says in the proposal?) (10 points)
    - Features (10 points)
    - Explanation of processes and results (10 points)

- Final report (25 points)

Project ideas

- A semi-structured information retrieval system, which crawls webpages from a narrow domain, identifies some of the facets (title of the page, author, etc.) and lets one search in those fields or the whole webpage.
- A search system for a focused crawl. For instance, crawl the blogosphere and collect blog postings on a specific topic, and present it to the user with a search interface.
- An IR system that crawls webpages from a narrow domain, creates and presents clusters of them, along with a search interface.

- A faceted browsing and searching system that allows one to browse through information content (text documents) based on various facets (category, author, date, etc.) as well as do full-text search.