IS392: Web Mining and Information Retrieval

Last updated Mar 12, 2019
(subject to change)

Faculty Instructor: Yang Liu
Office: GITC 5601
Office Hours: Tuesday and Thursday 11-12:45pm, and by appointments
E-mail: yl558 AT njit.edu

Classroom: CKB 315
Class Meets: Tuesday 1:00 pm-3:50 pm
Class Site: please go to moodle.njit.edu and login with your UCID. You will find IS 392, if you are enrolled in this class.

Overview

This course introduces the design, implementation and evaluation of search engines and web mining applications. Topics include: automatic indexing, natural language processing, retrieval algorithms, web page classification and clustering, information extraction, summarization, search engine optimization, and web analytics. Students will gain hands-on experience applying theories in case studies.

Prerequisites

- IS218 OR IT114 OR CS114

Learning Goals

1. Acquire a basic understanding of natural language processing.
2. Learn various automatic indexing techniques.
3. Obtain knowledge in retrieval models.
4. Learn web crawling.
5. Understand web usage, content and structure mining, with emphasis on the first two types.
6. Become familiar with web analytics.
7. Become familiar with applying web mining and analytics to search engine optimization.

NJIT University Code on Academic Integrity

https://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf is strictly enforced.

Textbook

Search Engines: Information Retrieval in Practice, by Croft, Metzler, and Strohman.
Publisher: Addision-Wesley
ISBN-13: 978—0-13-607224-9

Additional Materials

- Paper 1: What Do People from Information Retrieval?, W. Bruce Croft
- Paper 2: *Search Engine Optimization Starter Guide*, Google, http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en/us/webmasters/docs/search-engine-optimization-starter-guide.pdf

Assignments and Grading

• Participation and class activities			15%
○ Participation		4%	
○ In-class design activity		7%	
○ Alternative search engine presentation		4%	
• Assignments			45%
○ Assignment 1	Comparing Search Engines	6%	
○ Assignment 2	Developing a Web Crawler	12%	
○ Assignment 3	Developing an Indexer	12%	
○ Assignment 4	Mining a web collection	15%	
• Midterm			15%
• Final			25%
Total:			100%

Late Assignment Policy:
Last assignments will receive a penalty of 20% for each day. For example, suppose an assignment is worth 10 points and a student submits it one day late. The highest possible score for the student will be 8. After the assignment is discussed in class, no score will be given.

The **final letter grade** will be based on students' performance ranking, **approximately**: 15% of class will receive an A; 45% of class will receive a B+ or B; 30% of class will receive a C+ or C; and 10% of class will receive D or F.

Incompletes are only given to students with extenuating circumstance.

Weekly Coverage of Material

The following table shows approximately how much time may be devoted to each topic and the corresponding readings from the textbook and papers.

Week	Topics	Materials
1, Jan 22	Course Logistics and Introduction	What do people want from IR
2, Jan 29	Information Retrieval and Search Engines Assignment 1 out	Ch 1, 2
3, Feb 5	Crawls and Feeds Assignment 2 out	Ch 3
4, Feb 12	Crawls and Feeds (cont.) Processing text	Ch 3, 4
5, Feb 19	Processing Text (cont.) Assignment 2 Presentation	Ch 4
6, Feb 26	Ranking with Indexes	Ch 5
7, Mar 5	Assignment 3 out In-class design activity	
8, Mar 12	**Midterm**	
	Spring Break (No Class)	
9, Mar 26	Discussion on midterm exam results Ranking with Indexes (cont.) Assignment 3 Presentation	Ch 5
10, Apr 2	Web Mining Assignment 4 out	PPT on Moodle
11, Apr 9	Web Mining (cont.)	PPT on Moodle
12, Apr 16	Queries and Interfaces Assignment 4 Presentation	Ch 6
13, Apr 30	Retrieval Models, Evaluating Search Engines	Ch 7, 8
14, May 7	Social Search	Ch 10
15	**Final Exam (TBA)**	