Spring 2020

IS 392-002: Web Mining and Information Retrieval (Revised for Remote Learning)

Y.F. Brook Wu

IS392: Web Mining and Information Retrieval

Last updated March 24, 2020
(subject to change)

Faculty Instructor: Y.F. Brook Wu, Ph.D.
E-mail: wu AT njit.edu
Office: GITC 5607
Office Hours:

- before Spring break, in person Wednesday 3-5pm.
- **after Spring break, virtual office hours for questions related to recorded lectures: Wednesday 3-5pm, via WebEx https://njit.webex.com/meet/wu**

TA: Ms. Ye Xiong
Email: yx98 AT njit.edu
Office: GITC 5601
Office Hours: regularly Wednesday 10-11:15am and 2:30-5pm. **By appointment only:** please email directly.

- before Spring break, in person Wednesday 10-11:15am and 2:30-5pm.
- **after Spring break, virtual office hours for questions related to assignments: Wednesday 2-5pm, via WebEx https://njit.webex.com/meet/yx98**

Classroom: CKB 219
Class Meets: Wednesday 11:30 am-2:20 pm
Class Site: please go to canvas.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:
After an assignment is due, it will remain open for two additional days to accept late submissions. **There will be a 10% penalty applied to submissions that are 1 day late, and a 30% penalty for those that are 2 days late.** The assignment will then be closed and no more late submission will be accepted anymore.

Final Letter Grades:
The final letter grades 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, and documented proof for such circumstances must be provided to and verified by the Dean of Students office.

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	Ch 1, 2
3, Feb 5	Crawls and Feeds	Ch 3
4, Feb 12	Crawls and Feeds (cont.) Processing text	Ch 3, 4
5, Feb 19	Processing Text (cont.)	Ch 4
6, Feb 26	Ranking with Indexes	Ch 5
7, Mar 4	Ranking with Indexes (cont) In-class design activity	Ch 5
8, Mar 11	**Midterm**	
	Spring Break (No Class)	
9, Mar 25	Ranking with Indexes (cont.)	Ch 5
10, Apr 1	Web Mining	PPT on Canvas
11, Apr 8	Web Mining (cont.)	PPT on Canvas
12, Apr 15	Queries and Interfaces	Ch 6
13, Apr 22	Retrieval Models, Evaluating Search Engines	Ch 7, 8
14, April 29	Social Search	Ch 10
15	**Final Exam (Date/Time/Room TBD by Registrar's office)**	