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.  
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  
Publisher: Addision-Wesley  
Additional Materials

- Paper 1: What Do People from Information Retrieval?, W. Bruce Croft

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.

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