An introductory course on web-searching. Information vs data retrieval. The architecture of a search engine. Web crawling. Processing text (tokenization, stemming, stopwords, link analysis and markup). Ranking algorithms based on indexes and links (e.g. Kleinberg’s HITS, Google’s PAGERANK). Retrieval Models. Search engine evaluation. Case studies (e.g. Google cluster architecture).

1.1 Contact Information

Instructor: Alex Gerbessiotis  
E-mail: alexg+cs345@njit.edu  
Office: GITC 4213, 4th floor  
Tel: (973)-596-3244  
Office Hours: Tue 4:00-5:30pm and Thu 4:00-5:30pm  
Office Hours: By appointment Mon/Tue/Thu  
Class Hours: Mon 10:00-12:55 (GITC 1205)  

1.2 Course Administration

Prerequisites: CS 280 and one of CS 241/CS 252. Last 4 digits of your NJIT id.


CourseWork: 2 exams (including the final); Assignments

Grading: 1000 points = Exam1(335) + Exam2(335) + Best-5-of-7(330). HW1-HW4 are ordinary homeworks, HW5-HW6 are programming projects, and HW7 is a paper presentation; HW5-HW7 handed-out out of sequence. Each one is worth 66 points. For HW5-HW6 ONLY, a maximum of three students can work together and each one would collect the assigned graded points. HW7, the paper presentation requires a 20-minute reservation slot to be booked in advance, a one-page summary advance submission (see homework for details) and presentation.

Exams: All exams are open-textbook only. You may bring a hard-copy of the textbook but you are not allowed to borrow one during the exam or bring in class other material. Exam1 is on Mon Oct 26, 90mins. Exam2 is on Final Week, 120mins on a date to be announced by the Registrar.

ExamConflicts: Per University regulations.

Due Dates: Paper (aka Hard-copy) submissions for HW1-HW4 before class; email submissions (txt or pdf or MSWord) by midnight the day they are due. We acknowledge email submissions promptly. It’s up to you to properly form and submit an email. Use an NJIT email address and include a Subject line as specified in Handout 0. 11 pts deducted from grade at deadline plus 2 minutes, 22 pts every 24hrs thereafter.

Tentative list of topics

T1: WebSearching: Introduction  
T3: The retrieval process: Crawlers and crawling.  
T4: Search Engine Architecture, Duplicate Handling  
T5: Document Processing: Parsing and Tokenization  
T6: Document Processing: Indexing  
T7: Modeling retrieval and ranking  
T8: Queries, Query processing, and Interfaces  
T9: Search engine evaluation  
T10: Classification and categorization  
T11: Google MAPREDUCE model  
T12: Case Studies: GFS  
T13: Other Topics: Social Search
2.1 Course Objectives and Outcomes

Objective 1  Learn the fundamentals of Web searching.
Objective 2  Learn how a search engine works and identify the components of its architecture.
Objective 3  Learn the requirements and characteristics of web crawling, document fetching and processing.
Objective 4  Learn how to use fundamental data structures to index and store information for processing web search requests.
Objective 5  Learn the fundamentals of ranking and ranking algorithms.
Objective 6  Learn how high performance computing can benefit web searching.
Outcome 1   Be able to explain fundamental concepts related to Web searching and the architecture of search engines.
Outcome 2   Be able to identify and explain the output of search engines in the context of web searching.
Outcome 3   Be able to understand ranking and indexing algorithms and their limitations.
Outcome 4   Be able to design a search engine architecture based on input design requirements.
Outcome 5   Be able to effectively use high performance computing in the design of a Web search infrastructure.
Outcome 6   Be able to effectively apply ranking algorithms.

2.2 Tentative Course Calendar

<table>
<thead>
<tr>
<th>Week</th>
<th>Mon</th>
<th>HW out</th>
<th>HW in</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>Tue* 9/8</td>
<td>HW5 out</td>
<td>HW6 out</td>
<td>HW5, HW6 are mini-projects</td>
</tr>
<tr>
<td>W2</td>
<td>9/14</td>
<td>HW1 out</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W3</td>
<td>9/21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W4</td>
<td>9/28</td>
<td>HW2 out</td>
<td>HW1in</td>
<td></td>
</tr>
<tr>
<td>W5</td>
<td>10/5</td>
<td>HW3 out</td>
<td>HW2in</td>
<td></td>
</tr>
<tr>
<td>W6</td>
<td>10/12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W7</td>
<td>10/19</td>
<td></td>
<td>HW3in</td>
<td></td>
</tr>
<tr>
<td>W8</td>
<td>10/26</td>
<td><strong>Exam1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W9</td>
<td>11/02</td>
<td>HW4 out</td>
<td></td>
<td>Mon Nov 2: Withdrawal Deadline</td>
</tr>
<tr>
<td>W10</td>
<td>11/09</td>
<td></td>
<td>HW5 in</td>
<td></td>
</tr>
<tr>
<td>W11</td>
<td>11/16</td>
<td></td>
<td>HW4 in</td>
<td></td>
</tr>
<tr>
<td>W12</td>
<td>11/23</td>
<td></td>
<td></td>
<td>Thanksgiving week:Tue is a Thu</td>
</tr>
<tr>
<td>W13</td>
<td>11/30</td>
<td></td>
<td>HW7?</td>
<td>HW7 presentation?</td>
</tr>
<tr>
<td>W14</td>
<td>12/07</td>
<td>HW6in, HW7</td>
<td></td>
<td>HW7 presentation</td>
</tr>
<tr>
<td>W15</td>
<td></td>
<td><strong>Exam2</strong></td>
<td>Tue Dec 15- Mon Dec 21</td>
<td>is Final Exam Week</td>
</tr>
</tbody>
</table>

* First day of classes is the Tuesday after Labor Day (9/8) that is ”Monday” for NJIT  
** Check with the Registrar

Any modifications or deviations from these dates, will be done in consultation with the attending students and will be posted on the course Web-page. It is imperative that students check the Course Web-page regularly and frequently.
Grading

Written work will be graded for conciseness and correctness. Be brief and to the point and write clearly. Programming problems will be graded based on test instances decided by the instructor on an AFS machine (afsconnect1, afsconnect2, or osl11). Do not expect partial credit if your code fails to run on all test instances, and you do not provide a bug report.

Grades

Check the marks in written work and report errors promptly. Resolve any issue no later than the Reading Day. For students who submit programming work or have a paper presentation, an email with your grade will be sent back to you. The final grade is decided based on a 0 to 1000 point performance. A 50% or more is C or better, 85-90% or more usually guarantees an A.

Collaboration

Collaboration of any kind is NOT allowed in the in-class exams and the homeworks. An exception to this rule is HW5-HW6 that explicitly allow collaboration (teams of no more than 3); in such a case collaboration is allowed between members of the team only for the specific homework only. Students who turn in work/answers to questions sourced through the Internet or otherwise, or is product of another person’s/student’s work, risk severe punishment, as outlined by the University. The work you submit must be the result of your own effort.

Mobile Devices

Mobile phones/devices and/or laptops/notebooks MUST BE SWITCHED OFF (NOT JUST SILENCED) before the class exams. Switch off noisy devices before class.

Email/SPAM

Send email from an NJIT email address. NJIT spam filters or we will filter other email address origins. Use the appropriate subject line as specified in Handbout 0. Include cs345 in the subject line then.

Missing class

If you miss a class and there is no Exam or Homework due it’s up to you to make up for lost time.

Missing Exam

If you miss an exam and there is a valid documentation for your absence, such documentation must be presented within 3 working days from the day the reason for the absence is lifted. The maximum accommodation will be the number of missing days to the exam date. You also need to present your case to the Dean of Student Services (DOSS); we will respond after receiving confirmation from DOSS.

Missing HW

If you are sick (see Missing Exam for the procedure) there is no notion of a make-up homework or delayed submission of a homework other than the penalties specified on page one of this document. Per DOSS and Instructor approvals, a homework grade might get extrapolated from the final exam grade (EX2).

Programs

Follow submission guidelines for HW5-HW6, if you plan to do it/them.

Presentation

Follow submission guidelines for HW7, if you plan to do it.

The NJIT Honor Code will be upheld; any violations will be brought to the immediate attention of the Dean of Students. Read this handout carefully!