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 (eg. 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:	alg485@cs.njit.edu
Office:	GITC 4213, 4th floor	Tel:	(973)-596-3244
Office Hours:	Mon 4:00-5:30pm and Thu 4:00-5:30pm		
Class Hours:	Mon 6:00-9:00pm, FMH 305		
Web Page:	http://www.cs.njit.edu/~alexg/courses/cs485/index.html		

1.2 Course Administration

Prerequisites	No course prerequisites. Knowledge of last 4 digits of your NJIT id.
Textbook	Search Engines: Information Retrieval in Practice by B. Croft et al., Addison-Wesley, ISBN-10: 0136072240, 2010.
CourseWork:	**2 exams (including the final)**; **Assignments**
Grading:	1000 points = Exam1(250) + Exam2(250) + Best-4-of-8(500). (Although 8 or more assignments are to be handed out, you are expected to submit for grading no more than 7 for grading; the best 4 of those 7 submitted will count towards the final grade.)
HW	Each assignment is worth 12.5%. (In the remainder, 1% is 10 points.) Assignment A8 will be a paper presentation. A 30-minute reservation slot needs to be booked in advance.
Exams	Both exams are open-textbook only. You may bring a copy of the textbook but you are not allowed to borrow one during the exam. Exam1 is on **Mon Oct 24**, 120mins, 250 points. Exam2 is on **Mon Dec 19**, 120mins, 250 points
ExamConflicts	Per University regulations. (This is a higher numbered course.)
Due Dates	Email submissions **MUST be received by email before noon of the last day** they are due. We acknowledge submissions promptly. It's up to you to properly form and submit an email (see Handout 2). Use an NJIT email address. Late submission penalty: 20% per 24-hours. Written submissions are due by the beginning of a class at the classroom.

Tentatitive list of topics

Topics

```
T1 : WebSearching : Introduction
T2 : Fundamentals of Information Retrieval.
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

Fall 2011				
Week**	Mon	Out	In	Comments
W1	9/12			
W2	9/19	A1 out		
W3	9/26	A2 out	A1 in	
W4	10/3	A3 out	A2 in	
W5	10/10	A4 out		
W6	10/17		A3 in	
W7	10/24	**Exam1**		
W8	10/31	A8 out	A4 in	A8 is paper presentation
W9	11/7	A5 out		
W10	11/14	A6 out	A5 in	
W11	11/21		A6 in	
W12	11/28	A7 out		
W13	12/5		A7 in	
W14	12/12		A8 in	A8 is paper presentation
W15	12/19	**Exam2**		12/14-12/20 is exam week

* Exam2 is predetermined ** In this calendar, a week ends on a Monday

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.
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, 90% or more usually guarantees an A.
Collaboration	**Collaboration of any kind is NOT allowed in the in-class exams and the assignments. An exception to this rule is assignments that explicitly allow collaboration (teams of two); in such a case collaboration is allowed between members of the team only for the specific assignment component. 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 us will filter other email address origins. Do not send course email to the instructor's email address unless there is a good reason (e.g. you don't want the grader to read the email or it's urgent and you believe the instructor will respond faster). Include `CS 485` in the subject line then.
Missing class	If you miss a class and there is no Exam 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.
Programs	If an assignment requires programming work, submission guidelines will be provided for email submission of the assignment. It is imperative that you follow the guidelines in such a case. Submitted code must conform to the requirements of Handout 2.

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!