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:	alexg@cs.njit.edu
Office:	GITC 4213, 4th floor	Tel:	(973)-596-3244
Office Hours:	Mon 4:00-5:30pm and Tue 4:00-5:30pm		
Office Hours:	By appointment Mon/Tue/Wed		
Class Hours:	Mon 10:00-12:55, FMH 110		
Web Page:	http://www.cs.njit.edu/~alexg/courses/cs345/index.html		

1.2 Course Administration

Prerequisites: CS 280 and one of CS 241/CS 252. 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(300) + Best-5-of-9(450).
HW1-HW5 are ordinary homeworks but HW6 is a paper presentation considered as the sixth homework (and handed out out of sequence with the rest). Each one is worth 90 points. You must work alone on them. Three programming assignments PA1-PA3. Each one is worth 90 points; a maximum of two students can work together on each one of them and each one would collect the assigned points. HW6 is a paper presentation; a 30-minute reservation slot needs to be booked in advance and a one-page summary needs submitted 3 working days in advance.

Exams: Both 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 21**, 90mins, 250 points. Exam2 is on **??? Dec ??**, 120mins, 300 points; date to be announced by the Registrar.

ExamConflicts: Per University regulations.

Due Dates: Paper submissions for HW1-HW5 by start of class; email submissions for HW1-HW5 **MUST be received by email before 10:00am** the day they are due. We acknowledge submissions promptly. It's up to you to properly form and submit an email. For HW6 the deadline is specified in it. For PA1-PA3 it is midgnight (23:59) the day they are due (Monday). Use an NJIT email address and include a Subject line as specified in Handout 0. Late submission penalty: 20% per 24-hours starting at one minute after deadline.

Topics: Tentatitive list of 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 2013				
Week**	Mon	HWout	HWin,PAout.in	Comments
W1	9/9			
W2	9/16	HW1 out	PA1 out	
W3	9/23	HW2 out	HW1in PA2 out	
W4	9/30	HW3 out	HW2in PA3 out	
W5	10/7			
W6	10/14		HW3in	
W7	10/21	**Exam1**		
W8	10/28	HW4 out		
W9	11/4	HW6 out	PA1 in	HW6 is paper presentation
W10	11/11	HW5 out		
W11	11/18		HW4in PA2 in	
W12	11/25			Thanksgiving week
W13	12/2		HW5in	
W14	12/9		HW6in PA3 in	HW6 presentation
W15	12/?	**Exam2**		12/13-12/19 is exam week

* Exam2 is scheduled by the Registrar ** 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. Programming problems will be graded based on test instances decided by the instructor on an AFS machine (afsconnect1,afsconnect2). 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 assignments PA1-PA3 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 we will filter other email address origins. Use the appropriate subject line as specified in Handbout 0. Include `CS 345-001` 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 Students.
Programs	Follow submission guidelines for PA1-PA3.

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!