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+cs345@njit.edu
Office:	GITC 4213, 4th floor	Tel:	(973)-596-3244
Office Hours:	Mon 4:00-5:30pm and Tue 4:30-5:30pm		
Office Hours:	By appointment Mon/Tue/Wed		
Class Hours:	Tue 13:00-15:55, CKB207		
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)**; **Homeworks**

Grading: 1000 points = Exam1(333) + Exam2(333) + Points-of-HW(334).
HW1-HW4 are ordinary homeworks, HW5 is a paper presentation, and HW6 is a programming project. HW5, 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. For HW6 ONLY, a maximum of three students can work together and each one would collect the assigned graded points. Each HW is at least 66 points.

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 **Tue Oct 25**, 105mins. Exam2 is on **Final Week** , 120mins on a date to be announced by the Registrar.

ExamConflicts: Per University regulations.

Due Dates: Homeworks HW1-HW4 are by email submission before or on 23:59:59 of the due date. They can be an attachment that is a SINGLE .txt file, or SINGLE Word (.doc, .docx) file, or even a SINGLE PDF file. 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 1 minute, 22 pts every 24hrs thereafter.

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

A1,B1,F1. Learn the fundamentals of Web searching and be able to communicate succintly in writing and/or orally concepts related to Web searching and the architecture of search engines and Web search engines.

A2,B2,I1. Learn how a Web search engine works and be able to identify and describe the components of its architecture and identify and explain the output results of search engines in the context of web searching.

A3,B3,J1. Learn the requirements and characteristics of web crawling, document, processing and indexing and be able to understand and enumerate and describe the stpes involves in each phase.

A4,B4,C1,I2,J2. Learn how to use fundamental data structures to organize, index and store information for processing web search requests and be able to design a search engine architecture based on input design requirements.

B5,C2,I3,J3. Learn the fundamentals of ranking, and indexing algorithms and be able to understand ranking and indexing algorithms and their limitations; be able to effectively apply ranking algorithms on sample problem instances.

A5,B6,I4,J4. Learn how high performance computing can benefit web searching and be able to apply its methods in the design of a Web search infrastructure.

I1, K1. Learn the fundamentals of web searching and be able to design and implement a desktop-based search engine architecture of varying complexity using available programming language and software tools of your choice.

2.2 Tentative Course Calendar

Fall 2016				
Week	Tue	HWout	HWin	Comments
W1	09/06	HW5 out,HW6out		Paper presentation, Mini-project
W2	09/13	HW1 out		
W3	09/20			
W4	09/27		HW1in	
W5	10/04	HW2 out		
W6	10/11			
W7	10/18		HW2in	
W8	10/25	**Exam1**		
W9	11/01	HW3 out		
W10	11/08			Mon Nov 7: Withdrawal Deadline
W11	11/15	HW4 out	HW3in	
W–	11/22			Thanksgiving week:Tue is a Thu
W12	11/29		HW4in	
W13	12/06		HW5in	HW5 presentation
W14	12/13		HW5in HW6in	HW5 presentation; HW6 mini-project
W15		**Exam2****	Fri Dec 16- Thu Dec 22	is Final Exam Week

* First day of classes is the Tuesday (9/6) after Labor Day (9/5) ** 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 homeworks HW1-HW4, HW5. An exception to this rule is HW6 that explicitly allows for 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: it is imperative then that you contact relevant parties even before the 3 working day period has expired if the accommodation period is shorter. You also need to present your case to the Dean of Student Services (DOSS). We will respond after receiving confirmation from DOSS or we can give a make up exam that would only count if DOSS approves and the makeup exam is taken within the accommodation period.
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 EX1 or EX2.
Programs	Follow submission guidelines for HW6, if you plan to do it.
Presentation	Follow submission guidelines for HW5, 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 Student Services (DOSS). Read this handout carefully!