CSE 5/7337 - Spring 2012
Information Retrieval and Web Search

Introduction

Introduces the field of information retrieval with an emphasis on its application in Web search. Introduces the basic concepts of web crawling, stemming, tokenizing, inverted indices, text similarity metrics and the vector-space model. Popular Web search engines are studied and the concepts are applied in several Java-based projects. We will also survey state-of-the-art techniques and frameworks like MapReduce and Apache Lucene.

Prerequisites: CSE 3353 or departmental permission.

Instructor Contact Information

Dr. Michael Hahsler
Caruth 451
(214) 768-8878
mhahsler@lyle.smu.edu
Office hours: TBD

Course Web Site

http://michael.hahsler.net/SMU/7337/

Textbook

(IIR) Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

Online version at http://nlp.stanford.edu/IR-book/information-retrieval-book.html

Course Work and Grading

The final course grade will be based upon performance on various assignments and class participation. The percentage break-down is as follows:

Assignment	Percentage
Homework	15.00%
Project 1	25.00%
Project 2	25.00%
Project 3	25.00%
Participation	10.00%

Homework assignments and project instructions will be posted on the course web site.

It is expected that each student will keep up with the reading as outlined in the Covered Topics section below. Additional materials may be referenced in class as needed. Assignments and due dates will be posted on the web site. **Assignments need to be turned in per email on or before the due date.**

Final grades in this course are determined as follows:

93 - 100 A	80 - 82 B-	67 - 69 D+
90 - 92 A-	77 - 79 C+	63 - 66 D
87 - 89 B+	73 - 76 C	60 - 62 D-
83 - 86 B	70 - 72 C-	00 - 59 F

Covered Topics

Session	Date	Lecture	Assignment	Reading (IIR)
Week 1	1/18	Course overview, Introduction to IR,		
Week 2	1/23	Boolean Retrieval, Terms, Posting Lists	**HW 1**	1, 2
Week 3	1/30	The Web and crawling		19, 20
Week 4	2/6	Building a Web Crawler	**Project 1**	
Week 5	2/13	Wildcards and spelling errors		3
Week 6	2/20	Indexing	**HW 2**	4
Week 7	2/27	Hadoop, MapReduce		
Week 8	3/5	Walk through of Project 1	**M – Project 1 is due, Project 2**	
Week 9	3/12	**Spring break**		
Week 10	3/19	Scoring (Vector space model), Evaluation of IR systems	**HW 3**	6, 8
Week 11	3/26	Text Classification		13
Week 12	4/2	Walk through of Project 2	**M – Project 2 is due, Project 3**	
Week 13	4/9	Document Clustering		16, 17
Week 14	4/16	Link Analysis		21
Week 15	4/23	Walk through Project 3	**W – Project 3 is due**	
Week 16	4/30	**M – last day of class,** Master's Presentations (Nutch, Lucene, Solr)		

Learning Outcomes

After successful completion of this course, you should be able to:

1.0 – DEMONSTRATE COMPETENCY IN BASIC INFORMATION RETRIEVAL TECHNIQUES

1.1 Understand the concept of index terms and their use in an inverted index.

1.2 Understand how to score and rank query results.

1.3. Understand the criteria to evaluate the results of a IR system.

1.4 Understand the basic ideas of text classification and clustering.

2.0 – DEMONSTRATE COMPETENCY IN AUTOMATED WEB SEARCH

2.1 Understand how the web is organized and its fundamental properties.

2.2 Understand how search engines collect and index web content.

2.3 Understand how web search engines present the most relevant results for a given query.

2.4 Understand how to design and construct software that implements significant web IR concepts.

Attendance Policy

Because of the nature of this class, attendance of and participation in class is very important and students are expected to attend regularly. If a student is absent from class, it is that student's responsibility to make arrangements with the professor to make up any work missed or to ensure that assignments are submitted on time or early. Late assignments will not be accepted except in extreme instances. Any assignments that will be missed (including those due to university-sanctioned events) **must be completed before the due date.** This includes exams, homework and other assignments. Note that five percent of the semester grade is based upon class attendance and active participation.

Academic Ethics and Collaboration

Studying together is highly encouraged. However, you are expected to do your homework and projects independently unless stated otherwise in the instructions. All submitted work is expected to be your own. In particular:

- On-line sources can be only used in your work when properly stated where it came from and what adaptations you made. Also identify under what license you are using the code.
- You cannot copy text, even if you cite the source. You have to describe the information found in the literature in your own words and show how it fits into the context of your paper.

You will receive an automatic 0 on an assignment where you have been found copying. In severe cases, you will receive an F in the course and may be brought in front of the SMU Honor Council. It is your responsibility to know and understand the University's Honor Code and the expectations for collaboration in this course. The instructor of this course reserves the right to impose less severe penalties as seen fit.

Additional Information

Disability Accommodations: Students needing academic accommodations for a disability must first be registered with Disability Accommodations & Success Strategies (DASS) to verify the disability and to establish eligibility for accommodations. Students may call 214-768-1470 or visit http://www.smu.edu/alec/dass to begin the process. Once registered, students should then schedule an appointment with the professor to make appropriate arrangements.

Religious Observance: Religiously observant students wishing to be absent on holidays that require missing class should notify their professors in writing at the beginning of the semester, and should discuss with them, in advance, acceptable ways of making up any work missed because of the absence. (See University Policy No. 1.9.)

Excused Absences for University Extracurricular Activities: Students participating in an officially sanctioned, scheduled University extracurricular activity should be given the opportunity to make up class assignments or other graded assignments missed as a result of their participation. It is the responsibility of the student to make arrangements with the instructor prior to any missed scheduled examination or other missed assignment for making up the work. (University Undergraduate Catalogue)