CSE 435/535 Fall 2018
Information Retrieval

Lecture: Friday 3 – 5:30 pm
Farber 150 (South Campus)

Description:
This course will introduce students to text-based information retrieval (IR) techniques, i.e. search engines. Various IR models such as the Boolean model, vector space model, and probabilistic models will be studied. Efficient indexing techniques for (i) general document collections, (ii) specialized collections (e.g. Wikipedia, patents) and (iii) high velocity data such as social media will be discussed. Techniques for improving search efficiency, improving performance as well as evaluation methodology will be covered. The latter part of the course will focus on web search including link analysis techniques such as PageRank and HITS. Newer methods of representing content (neural embeddings, Word2vec) will be introduced. Students will work on programming projects to gain hands-on expertise in building IR systems.

Prerequisites: Programming expertise, Java, linear algebra

Textbook: Introduction to Information Retrieval by C. Manning, P. Raghavan, and H. Schütze, Cambridge University Press (2008)
Note: an online version of this book is available at http://informationretrieval.org

Instructor: Rohini K. Srihari, 338D Davis Hall
email: rohini@buffalo.edu
office hours: TBA

TA:

1. Archita Pathak architap@buffalo.edu
   Office hours: TBA
2. Lu Meng lumeng@buffalo.edu
   Office hours: TBA
3. Vicky Zheng vickyzhe@buffalo.edu
   Office hours: TBA
4. Kishlay Jha kishlayj@buffalo.edu
   Office hours: TBA

Course Details:

1. You are expected to attend all lectures and to complete all readings on time. In several weeks, there will be recitations in the latter part of class (see course schedule).

2. There will be 4 programming assignments in this course. The assignments cover the configuration of Solr for a particular search task, building of search indexes, evaluation of IR models, and a final assignment requiring the development of a complete IR solution based on a real-world problem. All programming assignments will require the use of an AWS account; more information on this will be provided in class.

3. We will use Piazza for course related discussion. The Piazza link is https://piazza.com/buffalo/fall2018/cse435535 Class notes will be posted there prior to class.

Projects and announcements will also be posted on this site. Piazza should be used for Q&A related to the course and particularly projects. **Note: Piazza site for this course will open on Monday, 27th Aug, 2018.**

4. Please read department policy on academic dishonesty; *this will be enforced strictly*.

IMPORTANT DATES

First day of class	Aug 31
Midterm- 1	Oct 5
Midterm- 2	Nov 9
Final Project Presentation & Last Lecture	Dec 7
Final Exam	Dec 13 11:45 – 2:45
Project 1 Due	Sept 20
Project 2 Due	Oct 18
Project 3 Due	Nov 15
Project 4 Due	Dec 9

GRADING

Midterms	30%
Final	20%
Projects	50%
Total	100%

COURSE SCHEDULE

Week and Date	Topics	Readings *	Important Activities
Week-1 Aug 31	Introduction to IR Conceptual Models of IR Boolean Model ***Project 1 release*** ***Recitation***	Chapter 1, 2	• Project 1 Release • Recitation - SOLR
Week-2 Sept 7	Tokenization Text analysis: stop lists, stemming Dictionaries, Tolerant Retrieval	Chapter 3 Supplements	
Week-3 Sept 14	Index Construction Distributed Indexing and Search Hadoop ***Recitation***	Chapter 4 Supplements	• Recitation – Project 1
Week-4 Sept 21	Text Properties: Heaps, Zipfs Laws Index Compression Vector-Space Model ***Project 2 release***	Chapter 5, 6	• **Project 1 Due on Sept 20** • Project 2 Release
Week-5 Sept 28	TF-IDF Weighting Scoring and Ranking in IR Systems ***Recitation***	Chapter 6, 7	• Recitation

Week-6 Oct 5	**Midterm** Evaluation Machine Learned Ranking	Chapter 8 Handouts	**Midterm 1**
Week-7 Oct 12	Relevance Feedback Query Expansion: Local and Global ***Recitation***	Chapter 9	• Recitation – Project 2, midterm 1 solutions
Week-8 Oct 19	Text Classification ***Project 3 release***	Chapter 13, 14 Handouts	• **Project 2 Due on Oct 18** • Project 3 Release
Week-9 Oct 26	Probabilistic IR: Okapi (BM 25), DFR Language Models for IR, Cross Lingual IR ***Recitation***	Chapter 11,12	• Recitation – Project 3
Week-10 Nov 2	Web Search Web Crawling ***Project 4 release*** ***Recitation***	Chapter 19, 20	• Project 4 Release • Recitation – Project 3
Week-11 Nov 9	***Midterm* 2** Social Network Analysis: Link Analysis PageRank, HITS	Chapter 21 Handouts	• **Midterm 2**
Week-12 Nov 16	Latent Semantic Indexing Word2Vec, Doc2Vec ***Recitation***	Chapter 18 Handouts	• Recitation – Project 4 • **Project – 3 Due on Nov 15**
Week-13 Nov 23	******THANKSGIVING BREAK******		
Week-14 Nov 30	Mobile Search, Personalized Search, Computational Advertising ***Recitation***	Handouts	• Recitation – Project 4
Week-15 Dec 7	***Student Project Presentation*** Summary		**Project – 4 Due on Dec 9**

*Chapters are from the *An Introduction to Information Retrieval* textbook unless specified.