CSE 435/535 Fall 2021 – Draft Syllabus
Information Retrieval

Reg # 18953/18954
Lecture: Monday, Wednesday 12:40 pm – 2 pm (Buffalo time)
Instructor: Rohini K. Srihari

Description:

This course will introduce students to text-based information retrieval (IR) techniques, i.e. search engines. The course begins with the fundamentals of processing large-scale, multilingual text document collections. 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, biomedical, patents) and (iii) high velocity data such as social media and e-commerce 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. The use of word vectors (Word2vec, GloVe) generated through neural models and their use in IR systems will be introduced. Students will work on programming projects to gain hands-on expertise in building IR systems. This course provides the foundation for the follow-on course (CSE 635) which discusses deeper text and web mining involving natural language processing (NLP).

Prerequisites: Programming expertise (Java, Python) 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
Other reference material will be made available on the piazza site during the semester.

Instructor: Rohini K. Srihari, Professor, Dept. of Computer Science & Eng
338D Davis Hall
email: rohini@buffalo.edu
office hours: TBD

TAs:

TBD

Course Details:

1. You are expected to attend all lectures and to complete all readings on time. Recordings will be made available shortly after live class concludes.

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 (group) project 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
piazza.com/buffalo/fall20201/cse435535/home. (tentative)

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.

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

UB Undergrad AI policy: https://catalog.buffalo.edu/policies/academic_integrity_2019-20.html

UB Graduate AI policy: https://grad.buffalo.edu/succeed/current-students/policy-library.academics.html#grievanceandintegrity

CSE AI policy: https://engineering.buffalo.edu/computer-science-engineering/information-for-students/policies/academic-integrity.html

IMPORTANT DATES

First day of class	Aug 30
Midterm- 1	October 7
Midterm- 2	Nov 11
Final Project Presentation & Last Lecture	Dec 9
Final Exam	TBD
Project 1 Due	Sept 19
Project 2 Due	Oct 14
Project 3 Due	Nov 5
Project 4 Due	Dec 11

GRADING

Midterms	30%
Final	15%
Projects	55%
Total	100%

COURSE SCHEDULE

Week and Date	Topics	Readings *	Key Activities
Week 1 Aug 30, Sept 1	Introduction to IR Conceptual Models of IR Boolean Model ***Project 1 release***	Chapter 1, 2	• Project 1 Release • Create twitter, AWS accounts
Week 2 Sept 7, 9	Tokenization Text analysis: stop lists, stemming Dictionaries, Tolerant Retrieval	Chapter 3 Supplements	Recitation – SOLR, AWS setup (hands-on)
Week 3 Sept 14, 16	Index Construction Distributed Indexing and Search Hadoop	Chapter 4 Supplements	
Week 4 Sept 21, 23	Text Properties: Heaps, Zipfs Laws Index Compression Vector-Space Model ***Project 2 release***	Chapter 5, 6	• **Project 1 Due on Sept 19** • Project 2 Release
Week 5 Sept 28, 30	TF-IDF Weighting Scoring and Ranking in IR Systems	Chapter 6, 7	
Week 6 Oct 5, Oct 7	Evaluation Machine Learned Ranking **Midterm 1**	Chapter 8 Handouts	**Midterm 1**
Week 7 Oct 12, 14	Relevance Feedback Query Expansion: Local and Global ***Project 3 release***	Chapter 9	• **Project 2 Due on Oct 14** • Project 3 Release
Week 8 Oct 19, 21	Probabilistic IR: Okapi (BM 25), DFR, Language Models	Chapter 11,12	
Week 9 Oct 26, 28	Prob IR contd. Text Classification	Chapter 13, 14	
Week 10 Nov 2, 4	Web Search Web Crawling	Chapter 19, 20	• **Project – 3 Due on Nov 5**
Week 11 Nov 9, 11	***Midterm* 2** Social Network Analysis: Link Analysis, PageRank, HITS ***Project 4 release***	Chapter 21 Handouts	• **Midterm 2** • Project 4 Release
Week 12 Nov 16, 18	Word Vectors: Latent Semantic Indexing Word2Vec, GloVe, Doc2Vec	Chapter 18 Handouts	
Week 13 Nov 23	Computational Advertising	Handouts	
Nov 25-29	******THANKSGIVING BREAK******		
Week 14 Nov 30, Dec 2	E-commerce, social media search Knowledge Graphs	Handouts	
Week 15 Dec 7, 9	***Student Project Presentations***		**Project – 4 Due on Dec 11**

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