CSE 435/535 Fall 2023 – Syllabus

Information Retrieval

Lecture: Monday, Wednesday 5:00 pm – 6:20 pm (ET) NSC 216
Instructor: Sayantan Pal

Description:
This course delves deep into text-based information retrieval (IR) techniques, offering a comprehensive look at search engines. It begins with the fundamentals of conceptual models of IR, including the Boolean model, and will delve into tokenization and text analysis. Throughout the course, students will be exposed to various IR models. These include the Boolean, vector space, and probabilistic models. It explores efficient indexing techniques and methods for indexing general document collections. The course will also touch upon vital components like scoring and ranking in IR systems and introduce students to embeddings and Latent Semantic Indexing (LSI). It touches upon the basics of deep learning for IR, the role of transformers in IR, and the evaluation methodologies in the field. The power of web search and crawling will be studied, along with link analysis techniques such as PageRank and HITS.

The course will deeply explore word vectors such as Word2Vec, GloVe, and Doc2Vec alongside cutting-edge topics like large language models and using RAG(Retrieval-Augmented Generation) for question-answering. Students will also learn about recommendation systems and collaborative filtering. Emphasis will be on hands-on projects using the GCP to bolster practical IR skills. Concluding the course, students will present their projects, bridging theory with practice. This course is foundational for understanding IR and paves the way for advanced classes like CSE 635, focusing on NLP and text mining.

Prerequisites: Programming expertise (Java, Python), Linear Algebra, Basic probability and statistics

Textbook: Introduction to Information Retrieval by C. Manning, P. Raghavan, and H. Schütze, Cambridge University Press (2008, online version 2012)
Note: An online version of this book is available at http://informationretrieval.org
Other, more recent reference material will be available on the piazza site during the semester.

Instructor: Sayantan Pal, Ph.D. student, Dept. of Computer Science & Eng 338Z Davis Hall
email: spal5@buffalo.edu
Office hours: TBA

Teaching Assistant: Xixian Yang, Ph.D. student, Dept. of Computer Science & Eng 338Z Davis Hall

email: spal5@buffalo.edu
Office hours: TBA

Graders:
Debasmit Roy, Shalini Agarwal (M.S. student, Dept. of Computer Science & Eng)

Course Details:

1. You are expected to attend all lectures and to complete all readings on time. Recordings will be made available shortly after the live class concludes. The recordings are meant to serve as study aids, not as a substitute for attending class.
2. There will be 3 projects in this course. The projects cover Solr configuration for a particular search task, building search indexes, evaluating IR models, and a final (group) project requiring developing a complete IR solution based on a real-world problem. All projects require a GCP account; more information will be provided in class.
3. We will use Piazza for course-related discussion. The Piazza link is: https://piazza.com/buffalo/fall2023/cse4535
4. Class notes will be posted there before class. Projects and announcements will also be posted on this site alongside the website. Piazza should be used for Q&A related to the course and particularly projects. **Students should not post class materials (notes, exams, projects) on public sites: this would be a violation of Intellectual Property rights**
5. Please read the department policy on academic dishonesty; this will be enforced strictly.
    a. UB Undergrad AI policy: https://catalogs.buffalo.edu/content.php?catoid=1&navoid=19#academic-integrity
    b. UB Graduate AI policy: https://www.buffalo.edu/grad/succeed/current-students/policy-library.html#academic-integrity
    c. CSE AI policy: https://engineering.buffalo.edu/computer-science-engineering/information-for-students/graduate-program/cse-graduate-academic-policies/cse-academic-integrity-policy.html

IMPORTANT DATES

First day of class	Aug 28
Midterm 1	Oct 2
Midterm 2	Nov 6
Final Project Presentation & Last Lecture	Dec 6
Project 1 Due	Sept 24

Project 2 Due	Oct 16
Project 3 Due	Dec 10

Syllabus

Week and Date	Number of Classes (Lec)	Topics	Readings	Key Activities
Week 1 Aug 28 Aug 30	2	Introduction to IR Conceptual Models of IR Boolean Model Project 1 release	Chapters 1, 2	
Week 2 Sept 4 (holiday) Sept 6	1	Tokenization Text analysis: stop lists, stemming Dictionaries, Tolerant Retrieval	Chapter 3 Supplements	Project 1 release Create accounts - GCP, Social Media (TBD) Tutorial – SOLR
Week 3 Sept 11, 13	2	Index Construction and Compression	Chapter 4 Supplements	
Week 4 Sept 18, 20	2	Text Properties (Heaps, Zipfs Laws) Index Compression, TF-IDF Weighting, Vector-Space Model	Chapters 5, 6	Sept 24 P1 Due
Week 5 Sept 25, 27	2	Scoring and Ranking in IR Systems Introduction to Embeddings & Latent Semantic Indexing (LSI)	Ch 6, 7, 18 Notes	Sept 25 P2 Rel
Week 6 Oct 2, 4	1	Midterm 1 Basics of Deep Learning for IR, Introduction to Transformers &	Notes	Midterm 1 - Oct 2

		their Role in IR		
Week 7 Oct 9 (holiday), 11	1	Evaluation in IR Relevance Feedback, Query Expansion	Ch 8,9	
Week 8 Oct 16, 18	2	Machine Learned Ranking, Probabilistic IR, Okapi BM25	Ch 11, 12	Oct 16 P2 Due Oct 18 A1 release
Week 9 Oct 23, 25	2	Web Search, Web Crawling	Ch 19,20	
Week 10 Oct 30, Nov 1	2	Social Network Analysis (Link Analysis, PageRank, HITS),	Ch 21	
Week 11 Nov 6, 8	1	Midterm 2 Word Vectors in-depth (Word2Vec, GloVe, Doc2Vec)	Notes	Midterm 2 - Nov 6 Nov 8 A1 Due, P3 rel
Week 12 Nov 13, 15	2	Intro to Large Language Models & Retrieval Techniques Practical Uses of Word Embeddings RAG for Question Answering	Notes	
Week 13 Nov 20, 22 (holiday)	1	Introduction to Recommendation Systems, Basics & Collaborative Filtering	Notes	
Week 14 Nov 27. 29	2	Projects - Applications of IR, Fact-Checking Past Project Demonstrations		
Week 15 Dec 4, 6	0	Student Project Presentations		P3 due Dec 10