# CSE 435/535 Fall 2022 – Draft Syllabus
# Information Retrieval

Reg # 18180/18181
Lecture: Monday, Wednesday, Friday 1:00 pm – 1:50 pm (Buffalo time)
Knox 109
Instructor: Sougata Saha

**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 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 (implemented on the AWS cloud computing platform) to gain hands-on expertise in building IR systems. This course provides the foundation for the follow-on course (CSE 635) which discusses natural language processing (NLP) and deeper text mining solutions.

**Prerequisites:** Programming expertise in Python, Linear Algebra

**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 made available on the piazza site during the semester.

**Instructor:** Sougata Saha, Dept. of Computer Science & Eng
email: sougatas@buffalo.edu
office hours: TBA

**TAs:**
Souvik Das. (souvikda@buffalo.edu)
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. The recordings are meant to serve as study aids, not as a substitute for attending class.

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 the Google Cloud Platform; 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/fall2022/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.

*** **You should not post class materials (notes, exams, projects) on public sites: this would be a violation of Intellectual Property rights** ****

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 29 |
| Midterm- 1 | October 5 |
| Midterm- 2 | Nov 14 |
| Final Project Presentation | Dec 7 & 9 |
| Last Lecture | Dec 9 |
| Project 1 Due | Sept 21 |
| Project 2 Due | Oct 12 |
| Project 3 Due | Nov 2 |
| Project 4 Due | Dec 7 |

**GRADING**

| | |
|---|---|
| Midterms | 40% |
| Projects | 60% |
| Total | 100% |

**COURSE SCHEDULE**

| Week and Date | Topics | Readings * | Key Activities |
|---|---|---|---|
| Week 1<br>Aug 29, 31, Sept 2 | Introduction to IR<br>Conceptual Models of IR<br>Boolean Model<br>***Project 1 release*** | Chapter 1, 2 | • Project 1 Release<br>• Create Reddit, GCP accounts |
| Week 2<br>Sept 5 (**holiday**), 7, 9 | Tokenization<br>Text analysis: stop lists, stemming<br>Dictionaries, Tolerant Retrieval | Chapter 3<br>Supplements | Recitation – SOLR, GCP setup (hands-on) |
| Week 3<br>Sept 12, 14, 16 | Index Construction<br>Distributed Indexing and Search<br>Hadoop | Chapter 4<br>Supplements | |
| Week 4<br>Sept 19, 21, 23 | Text Properties: Heaps, Zipfs Laws<br>Index Compression<br>Vector-Space Model<br>***Project 2 release*** | Chapter 5, 6 | • **Project 1 Due on Sept 21**<br>• Project 2 Release |
| Week 5<br>Sept 26, 28, 30 | TF-IDF Weighting<br>Scoring and Ranking in IR Systems | Chapter 6, 7 | |
| Week 6<br>Oct 3, 5, 7 | Evaluation<br>Machine Learned Ranking<br>**Midterm 1** | Chapter 8<br>Handouts | **Midterm 1** |
| Week 7<br>Oct 10, 12, 14 | Relevance Feedback<br>Query Expansion: Local and Global<br>***Project 3 release*** | Chapter 9 | • **Project 2 Due on Oct 12**<br>• Project 3 Release |
| Week 8<br>Oct 17, 19, 21 | Probabilistic IR: Okapi (BM 25), DFR, Language Models | Chapter 11, 12 | |
| Week 9<br>Oct 24, 26, 28 | Prob IR contd.<br>Text Classification | Chapter 13, 14 | |
| Week 10<br>Oct 31, Nov 2, 4 | Web Search<br>Web Crawling | Chapter 19, 20 | • **Project – 3 Due on Nov 2** |
| Week 11<br>Nov 7, 9, 11 | Social Network Analysis: Link Analysis, PageRank, HITS<br>***Project 4 release*** | Chapter 21<br>Handouts | • Project 4 Release |
| Week 12<br>Nov 14, 16, 18 | Word Vectors: Latent Semantic Indexing<br>Word2Vec, GloVe, Doc2Vec<br><br>Using word embeddings in Search<br>Computational Advertising | Chapter 18<br>Handouts | |
| Week 13<br>Nov 21 | ***Midterm* 2** | | **Midterm 2** |
| Nov 23, 25 | ******THANKSGIVING BREAK****** | | |
| Week 14<br>Nov 28, 30, Dec 2 | E-commerce, social media search<br>Knowledge Graphs | Handouts | |
| Week 15<br>Dec 5, 7, 9 | ***Student Project Presentations*** | | **Project – 4 Due on Dec 7** |

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