SI 650 Course Syllabus

SI 650 / EECS 549 – Information Retrieval
Syllabus (Subject to Change)
Friday 1-4 pm, North Quad 1255
Instructor: Qiaozhu Mei (qmei@umich.edu);
Office Hours: Fridays 10-11am
GSI: Shiyan Yan (shiyansi@umich.edu);
Office Hours: Monday 1-2pm, Thursday 2-3pm, NQ 1270
Along with the explosive growth of online textual information (e.g., web pages, email, news articles, social media, and scientific literature), it is increasingly important to develop tools to help users access, manage, and exploit the huge amount of information. Web search engines, such as Google and Bing, are good examples of such tools, and they are now an essential part of everyone's life. In this course, you will learn the underlying technologies of these and other powerful tools for connecting people with information, for accessing and mining unstructured information, especially text. You will be able to learn the basic principles and algorithms for information retrieval as well as obtain hands-on experience with using existing information retrieval toolkits to set up your own search engines and improving their search accuracy.
Unlike *structured* data, which is typically managed with a relational database, textual information is *unstructured* and poses special challenges due to the difficulty in precisely understanding natural language and users' information needs. In this course, we will introduce a variety of techniques for accessing and mining text information. The course emphasizes basic principles and practically useful algorithms. Topics to be covered include, among others, *text processing*, *inverted indexes*, *retrieval models* (e.g., *vector space* and *probabilistic* models), *ir evaluation*, *text categorization*, *text filtering*, *clustering*, *topic modeling*, *deep learning, retrieval system design and implementation*, *web search engines*, and *applications of text retrieval and mining.*
This course is designed for graduate students and senior undergraduate students of the School of Information and Computer Science and Engineering. The course is lecture based. Grading is based on three individual assignments, a midterm exam, and a course project.

Learning Objectives:
- Understand how search engines work.
- Understand the limits of existing search technology.
- Learn to appreciate the sheer size of the Web.
- Learn to write code for text indexing and retrieval.
- Learn about the state of the art in IR research.
- Learn to analyze textual and semi-structured data sets.
- Learn to appreciate the diversity of texts on the Web.
- Learn to evaluate information retrieval systems.
- Learn about standardized document collections.
- Learn about text similarity measures.
- Learn about semantic dimensionality reduction.

- Learn about the idiosyncrasies of hyperlinked document collections.
- Learn about web crawling.
- Learn to use existing tools of information retrieval.
- Understand the dynamics of the Web by building appropriate mathematical models.
- Understand how text classification and clustering works
- Build working systems that assist users in finding useful information on the Web.

Text:

- Required: [Zhai and Massung] ChengXiang Zhai and Sean Massung, Text Data Management: A Practical Introduction to Information Retrieval and Text Mining. ACM and Morgan & Claypool Publishers, July 2016. http://dl.acm.org/citation.cfm?id=2915031 (Links to an external site.)Links to an external site. (Free U of M access)
- Required: [Manning] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. http://nlp.stanford.edu/IR-book/information-retrieval-book.html (Links to an external site.)Links to an external site.
- Papers from SIGIR, WWW, and other venues to be assigned in class

Sample assignments:

1. Getting familiar with text analysis (problem sets + programming);
2. Basic retrieval systems (coding and analysis);
3. Text classification (programming/data analysis)

Kaggle-in-class competition will be used in one or more of the assignments.

Final Project:

An open-ended project in one of three categories:

- IR system - develop a working, useful IR system (with either a user interface or an API). Students will be responsible for identifying a niche problem, implementing it and deploying it, either on the Web or as an open-source downloadable tool. A good example is a vertical search engine (a specialized search engine for a particular domain).
- Research project - Students will take on a research problem, conduct hypothesis formulation, experiments, and a technical report in the format of a conference submission.
- Survey paper - identify a cutting-edge topic of research in IR and summarize at least 10 recent papers on it, along with an synergy that compares and contrasts the papers involved.

Grading:

- Class Participation: 10%
- Homework assignments: 30%

- Midterm Exam: 25%
- Course Project: 35%
  - Proposal: 5% (due at the 7-th week)
  - Progress progress check: 5% (due at the 10-th week)
  - Presentation: 5% (13-th week)
  - Final report: 20%

Tentative Schedule:

Week 1 (9/7): Introduction to information retrieval; Mathematical basics
- Vector spaces and similarity;
- Probabilities and Statistics;

Reading:
- **Vannevar Bush** "As We May Think" Bush 45 (Links to an external site.)Links to an external site..
- [Zhai and Massung] Chapter 2, “Background”

Week 2 (9/14): Understanding Text Data.
- Document processing, stemming;
- String matching;
- Basic NLP tasks – POS tagging; shallow parsing.

Reading:
- [Zhai and Massung] Chapter 3: Text Data Understanding
- Thorsten Brant, "Natural Language Processing in Information Retrieval (Links to an external site.)Links to an external site."

Homework 1 released

Week 3 (9/21): Building Text Retrieval Systems.
- System architecture;
- Boolean models;
- Inverted Indexes;
- Document ranking;
- IR Evaluation;

Reading:
- [Zhai and Massung]: Chapter 5 “Overview of text data access”; Chapter 9 “Search Engine Evaluation”.
- [Manning] Chapter 1 “Boolean retrieval”; Chapter 8 “Evaluation in information retrieval”

Week 4 (9/28): Retrieval Models: Vector Space Models
- Vector space models;
- TF-IDF weighting;
- Retrieval axioms;

· Implementation issues;

Reading:
* [Zhai and Massung] Chapter 6: “Retrieval Models”, up to 6.3
* [Manning] Chapter 6 "Scoring, term weighting & the vector space model"
* A. Singhal. Modern information retrieval: a brief overview. IEEE Data Engineering Bulletin, Special Issue on Text and Databases, 24(4), Dec. 2001. (http://singhal.info/ieee2001.pdf (Links to an external site.)Links to an external site.).
* Hui Fang, Tao Tao and ChengXiang Zhai. "A Formal Study of Information Retrieval Heuristics". In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 49-56, Sheffield, United Kingdom, 2004. (http://www.ece.udel.edu/~hfang/pubs/sigir04-formal.pdf (Links to an external site.)Links to an external site.)

Homework 1 due; Homework 2 (Part I) released.

Week 5 (10/5): Retrieval Models: Probabilistic models
· Okapi/BM25;
· Language models;
· KL-divergence;
· Smoothing;

Reading:
- [Zhai and Massung] Chapter 6: “Retrieval Models”, 6.4
- [Manning] Chapter 11 & 12

Homework 2 (Part I) due; Homework 2 (Part II) released.

Week 6 (10/12): Query expansion and feedback
· Query reformulation;
· Relevance feedback;
· Pseudo-relevance feedback;
· Language model based feedback;

Reading:
- [Zhai and Massung] Chapter 7, “Feedback”
- [Manning] Chapter 9

Homework 2 (Part II) due.

Week 7 (10/19): Web Search Engines;
· Models of the Web;
· Web crawling;
· Static ranking: PageRank; HITS;
· Query log analysis;
· Adversarial IR;

Reading:
- [Zhai and Massung] Chapter 10, “Web Search”
- [Manning] Chapter 19-21;

Project Proposal Due

Week 8 (10/26): Information filtering. User Interfaces.
· Adaptive filtering;
· Collaborative filtering;
· User Interfaces.

Reading:
- [Zhai and Massung]: Chapter 11, “Recommender Systems”

Week 9 (11/2): Text classification;
· Naïve Bayes; k-nearest neighbors;
· Feature selection;
· Supervised learning;
· Semi-supervised learning;

Reading:
- [Zhai and Massung] Chapter 15, “Text Categorization”
- [Manning] Chapter 13,14, and 15 up to SVM.

Week 10 (11/9): Text clustering;
· Vector-space clustering;
· K-means; EM algorithm;
· Text shingling;

Reading:
- [Zhai and Massung] Chapter 14, “Text Clustering”, Chapter 17, “Topic Analysis”
- [Manning] Chapter 16, 17, 18
- PLSA note (Links to an external site.)Links to an external site.

Project Progress Check; Midterm Released (Due in one week).

Week 11 (11/16): Deep Learning for Text;
· Deep Neural Networks
· Word representation learning
· Deep learning for text classification
· Other applications of deep learning

Reading: TBA

Week 12 (11/30): IR applications;
· Information extraction

- Question answering
- Sentiment and Opinions
- Computational advertising

Reading: TBA

Week 13 (12/7): Project Presentations
Week 14 (12/14): No Class. **Project Report Due**

—-

Academic Integrity and Misconduct:
Collaboration
UMSI strongly encourages collaboration while working on some assignments, such as homework problems and interpreting reading assignments as a general practice. Active learning is effective. Collaboration with other students in the course will be especially valuable in summarizing the reading materials and picking out the key concepts. You must, however, write your homework submission on your own, in your own words, before turning it in. If you worked with someone on the homework before writing it, you must list any and all collaborators on your written submission. Each course and each instructor may place restrictions on collaboration for any or all assignments. Read the instructions careful and request clarification about collaboration when in doubt. Collaboration is almost always forbidden for take-home and in class exams.
Plagiarism
All written submissions must be your own, original work. Original work for narrative questions is not mere paraphrasing of someone else's completed answer: you must not share written answers with each other at all. At most, you should be working from notes you took while participating in a study session. Largely duplicate copies of the same assignment will receive an equal division of the total point score from the one piece of work. You may incorporate selected excerpts, statements or phrases from publications by other authors, but they must be clearly marked as quotations and must be attributed. If you build on the ideas of prior authors, you must cite their work. You may obtain copy editing assistance, and you may discuss your ideas with others, but all substantive writing and ideas must be your own, or be explicitly attributed to another. See the (Doctoral, MSI, MHI, or BSI) student handbooks available on the UMSI intranet for the definition of plagiarism, resources to help you avoid it, and the consequences for intentional or unintentional plagiarism.

Accommodations for Students with Disabilities:
If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way the course is usually taught may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Services for Students with Disabilities (SSD) office to help us determine appropriate academic accommodations. SSD (734-763-3000; http://ssd.umich.edu (Links to an external site.)Links to an external site. *) typically*

recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. Any information you provide is private and confidential and will be treated as such.

Student Mental Health and Wellbeing

The University of Michigan is committed to advancing the mental health and wellbeing of its students, while acknowledging that a variety of issues, such as strained relationships, increased anxiety, alcohol/drug problems, and depression, directly impacts students' academic performance. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764-8312 and https://caps.umich.edu/ (Links to an external site.)Links to an external site. during and after hours, on weekends and holidays or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (732) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs (Links to an external site.)Links to an external site. , or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources. For a more comprehensive listing of the broad range of mental health services available on campus, please visit: http://umich.edu/~mhealth/ (Links to an external site.)Links to an external site. .