SI 650 / EECS 549 Information Retrieval
Course Syllabus

Wednesday 4:30-7:30 pm (EDT), Remote Only

Class Page https://umich.instructure.com/courses/395426

Instructor: David Jurgens (jurgens@umich.edu);
Office Hours: Monday 1:00-2:00 pm (EDT), Tues or Thurs TBD (am) where you reserve your time here.

GSIs: Shiyan Yan (shiyansi@umich.edu); Yuankun Zhu (jennyykz@umich.edu);
GSI Office Hours: TBD

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.

Textbooks and Readings

- Papers from SIGIR, WWW, and other venues to be assigned in class
Homework and Grading

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 workable, useful IR system. Students will be responsible for identifying a vertical domain, implementing a search engine or other types of information retrieval systems for the domain, evaluating it, and deploying it either on the Web or as an open-source tool. A general purpose search engine or Web-scale system is discouraged.
- 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 (in rare conditions) - 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: 20%
- Course Project: 40%
  - Proposal: 5% (due at the 7-th week)
  - Progress progress check: 10% (due at the 9-th week)
  - Blog Post: 5% (14-th week)
  - Final report: 20%
What does Participation Mean in the Zoom era?

Normally, participation is a bit easier to measure in class, but given that we're doing everything virtually and for some folks, asynchronous, we'll try measuring participation broadly to allow everyone to participate, even if they can't make it to class or don't want to speak up. Following are a few ideas on how to participate. We'll also have a running Piazza thread where we can pitch ideas for participation.

Participation ideas and suggestions

- Holding a multi-group work time for course projects
- Leading Break-time Calisthenics
- Sharing new class-related papers and resources on Piazza
- Asking questions in class
- Asking questions on Piazza
- Answering questions on Piazza (very helpful!)

Pandemic Time Considerations

The Fall 2020 semester is certainly going to be an unusual one. I will do all I can to help create a positive learning environment that works well regardless of location, bandwidth, and other factors in my control. However, the instructor fully recognizes that health-related and personal life events can (and often should) take priority over course work in some circumstances. If you find yourself in such circumstances, please reach out and we'll find a way for you to complete the course. I will typically not require extensive evidence of this but will ask for some evidence that is not of a sensitive nature (we will discuss this on a case by case basis). If something has become chronic or systemic, please do reach out early and we'll work together. If something is urgent, go take care of it now and we'll discuss later and work together on it; life is more than just this class!

I also recognize that this semester will likely be more stressful than past, so I encourage anyone experiencing such feelings to reach out to the CAPS folks (details below).

Grading Scale

Grading follows the traditional scale.
97-100: A+
93-97: A
90-93: A-
87-90: B+
83-87: B
80-83: B-
77-80: C+
73-77: C
70-73: C-
67-70: D+
63-67: D
60-63: D-
Below 60: F

There is no curve for this class. That said, usually the majority of students have received As. If you are concerned about your grade, please come see the instructors early (not the last week) while there is still time to improve your performance.

No extra credit or make-up assignments are allowed (e.g., no retroactive work after the end of the semester to bump up grade). However, if a numeric grade is just below the letter grade borderline, the instructor reserves the right to increase the letter grade if the student has done a truly exceptional project (but will never decrease the grade!).

Where and How We Meet (with Backup Plans)

The official Zoom link is on Canvas and linked here, but if for some reason this link doesn’t work, use whatever is on Canvas.

If his network connection drops or his video freezes, don’t give up on the meeting! Hang out for at least 5 minute while he reboots his computer or switches to his cell phone for Internet connectivity. Thanks!

If in the event that virtual class disaster ensues, your will record whatever remaining part of the lecture is left and upload this video to Canvas for asynchronous viewing.

How you are encouraged to use Zoom in class

- Have your camera on as much as possible during class meetings (but mute your microphone as much as possible, too)
- I strongly encourage you to keep your camera on for community-building purposes (but will not require it).
- Check out your background when you are on camera. Remove or cover items you don’t want others to see
- Set up a profile photo for your account (for when your camera is off)
- Pets are welcomed! 🐱
- Test your audio and video - use headphones if necessary
● Ask questions! Audio questions are preferred but chat questions are great too
  ○ If you see a question in the chat go by and I haven’t gotten to it, feel free to let me know (even if it’s not your question!). Sometimes I will miss these so your help is much appreciated.

Other general rules for Zoom use in this class:

● You must always use your UMich credentials/account to access this course. If you have another Zoom account, it will not work
  ○ See this page for information on how to log into your U-M Zoom account (Links to an external site.)
  ○ If you have already used your UM email to create a non-UM Zoom account, please see this page for information on how to migrate your account. (Links to an external site.) (Note that this can take several hours, so please do it well before the first class meeting.)
● Only connect using a Zoom client, preferably on your laptop/desktop (better than phone or tablet). NEVER on a web browser
● Please keep your Zoom client updated to get the newest features!
● Change your display name in Wolverine Access (“Preferred” name in “Campus Personal Information”)

What to do if you are on a low-bandwidth connection?

● Leave video off when you don’t need it
● Turn off HD video (in the Zoom app video settings)
● When you screen share, only do so for as long as necessary
● Use collaborative (e.g., Google) documents rather than screen sharing
● Mute your audio when not speaking (do this anyway!)

To improve your overall Zoom performance, consider asking others at your location to limit their high-bandwidth activities while you need to be online for a course or required meeting.

Avoid running other data-intensive programs during Zoom meetings, such as streaming video or music or other web sites with dynamic content. (Why are you watching streaming video during class?)

(These tips come from our colleagues at Cornell. See their doc for more details (Links to an external site.))
Tentative Schedule:

Week 1 (9/2): 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.)
- [Zhai and Massung] Chapter 2, “Background”

Week 2 (9/9): 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.)"

Week 3 (9/16): Building Text Retrieval Systems.

- System architecture;
- Boolean models;
- Inverted Indexes;
- Document ranking;
- IR Evaluation;

Reading:
Week 4 (9/23): 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"


Week 5 (9/30): 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

Week 6 (10/7): 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

Week 7 (10/14): 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”
• Dawei Yin, Yuening Hu, Jiliang Tang, Tim Daly, Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, Jean-Marc Langlois, Yi Chang, "Ranking Relevance in Yahoo Search", Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2016. (http://www.yichang-cs.com/yahoo/KDD16_yahoosearch.pdf (Links to an external site.))

• Optional: [Manning] Chapter 19-21;

Week 8 (10/21): Information filtering. User Interfaces.

- Adaptive filtering;
- Collaborative filtering;
- User Interfaces.

Reading:
• [Zhai and Massung]: Chapter 11, "Recommender Systems"

Week 9 (10/28): 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 (11/4): 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.)

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

Reading: TBA

Week 12 (11/18): Online experiments and causal factors
  · A/B tests, Online experiments
  · Analyzing experimental results
  · Offline experiments
  · Counterfactual Analysis

Reading: TBA
Week 13 (12/2): IR applications;

- Information extraction
- Question answering
- Sentiment and Opinions
- Computational advertising

Reading: TBA

<table>
<thead>
<tr>
<th>Week</th>
<th>Assigned</th>
<th>Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>HW1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>HW2</td>
<td>HW1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Project Proposal</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>HW2</td>
</tr>
<tr>
<td>8</td>
<td>HW3</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Project Update</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Midterm</td>
<td>HW3</td>
</tr>
</tbody>
</table>
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.) ) 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.) 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.) , 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.) .