University of Windsor

SCHOOL OF COMPUTER SCIENCE

Comp-8380
Information Retrieval Systems
Fall 2023
COURSE SYLLABUS

LAND ACKNOWLEDGEMENT

The School of Computer Science at the University of Windsor *sits on the Traditional Territory of the Three Fires Confederacy of First Nations*. We acknowledge that this is the beginning of our journey to understanding the Significance *of the history of the Peoples of the Ojibway, the Odawa, and the Pottawatomie.*

INSTRUCTOR:

Jianguo Lu

E-mail: jlu@uwindsor.ca
Office Location: *Lambton Tower 5112*
Office Hours: **9:00-11:00, Thursday**

COURSE DESCRIPTION:

Information retrieval (IR) is finding documents, i.e., unstructured natural language text, from within large collections. Text documents are usually interconnected, such as hyperlinks in the case of Web documents, and citations in the case of academic papers. Thus, the key issues in IR are analyzing text and graphs in large scale. This course will cover the basic techniques in text and graph analysis.

LEARNING OUTCOMES:

After the course you will be able to
– Understand text statistics and language models, the power law in natural language, Zipf's law, Heaps' law;
– Understand some fast algorithms in search engine constructing, indexing, vector space model and TF-IDF information retrieval model;
– Understand the link/graph analysis, in particular the PageRank algorithm;
– Classify and cluster documents using various machine learning algorithms;
– Apply deep learning in text and graph analyses.

REQUIRED TEXTBOOK:

I will mainly follow the IIR book listed below. This is an excellent book and also available online. Some algorithms for large data processing are described in more detail in the MMD book, which is also available online. The third book (LA) is a practical introduction to Lucene search engine. SLP book has more details on n-gram language models and neural language models.
IIR Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze, Introduction to Information Retrieval, Cambridge University Press. 2008.
MMD Anand Rajaraman and Jeff Ullman, Mining of massive datasets , 2020.
SLP Dan Jurafsky and James H. Martin, Speech and Language Processing (3rd ed. draft), 2023.
LA Michael McCandless, Erik Hatcher, and Otis Gospodnetic, Lucene in Action, Second Edition. 2010.

COURSE EVALUATION:

The grading will be based on mainly on exam and project. The weight of the final exam is 50%, the project is 50%. The project is about constructing a real search engine for academic papers. It is divided into two components. You need to accomplish each component in time.

COURSE SCHEDULE:

Topics*

(The instructor reserves the right to change the outline to accommodate student pace and understanding of the subject matter.)
– Text operations before indexing, such as stop word removal, stemming;
– Statistic properties of text, power laws, Zipf's law, Heaps' law.

– Language models, unigram model and bigram model. Smoothing techniques.
– Indexing, constructing an inverted index of word to document pointers;
– Vector space model. TF-IDF and their variants;
– Document classification. Naive Bayes classification (multinomial and Bernoulli) , Feature selection. Mutual information. Feature transformation.
– Distributional representation of words. .Word co-occurrence. Neural Network based text processing, word2vec, and doc2vec.
– Latent semantic indexing. Singular value decomposition,
– Ranking, scoring retrieved documents according to relevance or importance metrics. PageRank algorithm. Markov chain.
– Evaluation criteria, precision and recall, F1.
– Document clustering, K-means, HAC.
– Graph representation. DeepWalk and Node2vec.
– Searching, retrieving documents that contain a given query token from the inverted index; Use Lucene to construct a practical and large search engine.

**__Note:__ Students are advised that the schedule and topics described above are tentative and that the material and/or depth and order of presentation are subject to change at the discretion of the instructor and student pace.*
This course assumes the student will allocate a significant amount of independent study and time spent on reading and researching materials as needed. You are strongly encouraged to ensure sufficient time needed to succeed in this course.

IMPORTANT DATES:

Fall 2023

Thursday, September 7: First day of classes
Wednesday, September 20: Last day for late registration for Fall classes (to add classes)
Wednesday, October 4: Fall financial drop date.
Saturday, October 7 – Sunday, October 15: Fall Reading Week
Monday, October 9: Thanksgiving Day (Statutory Holiday – University closed
Wednesday, November 15: Last day to voluntarily withdraw from Fall classes (to drop classes)
Wednesday, December 6: Last day of classes
Saturday, December 9 – Wednesday, December 20: Fall Final Exams
Thursday, December 21: Alternate Exam Day
Saturday, December 23 – Tuesday, January 2: University offices closed for December Holiday recess.

RESOURCES:

The course website is on ***https://brightspace.uwindsor.ca/***
Please check it frequently for announcements and other useful info.

GRADING:

A numeric grade on a scale of 0 to 100 will be assigned (rounded integer).

Passing grade:
A minimum grade of 50% is required to pass this course (70% for grad courses). Your individual program may have higher requirements to maintain good standing; please consult your program requirements and plan accordingly. If you are registered in a course and do not attend or participate or write any evaluations will be assigned a grade of NR (No report). You must withdraw from the course if you do not wish to attend it; not showing up does not constitute withdrawal and will impact your academic record.

Voluntary withdrawal (dropping the course):
You may drop a course within the first 2 weeks add/drop period (1 week in case of 6-week courses) without it showing up on your academic record. Please check with the Registrar's office calendar on the important dates for withdrawing voluntarily from a course after the add/drop period should you feel you need to withdraw. It is strongly recommended that you seek academic advice from your instructor or an academic advisor prior to withdrawing from courses.

Absences due to medical or other extenuating circumstances:
Medical leaves, illness, death (in the family), and other difficult circumstances as determined in bylaw 54 are at times unavoidable and would interrupt your academic career. You must report any issues to the instructor as soon as possible prior to considering any academic accommodations. The instructor reserves the right to determine if an accommodation

is merited and the nature of the accommodation related to the course evaluation. All requests for alternate considerations on medical grounds or other difficult matters must be made in writing (email) to the instructor along with supporting documents prior to the end of the course. No alternate accommodations will be considered after the end of the course.

Makeup and missed assessment policy:
If you miss a test, assignment, or other assessment in the course you will receive a zero mark for the missed work. If you wish to have alternate considerations due to a valid reason (as per senate bylaw 54) you must inform the instructor in writing (email) as soon as possible, preferably before the assessment, and not later than seven calendar days. Considerations for any make-up or late submissions will be done on a case-by-case basis on compassionate grounds while maintaining fairness as much as possible. No alternate considerations will be given to any missed assessment if the instructor is not informed within seven calendar days after its due date. The instructor will refuse any unsubstantiated and late requests.

Grade appeal:
Informal reviews and appeals of the marks for assignments, midterm, exams and/or projects will be considered only if requested within 10 days after the release of the corresponding grades. After the 10-day period students will have to submit a formal appeal if they wish within 6 weeks. See Senate Bylaws 54 (Undergraduate Students) and Senate Bylaws 55 (Graduate Students) for more details on appealing about grades.

Other Notes:
1.A. Undergraduate Students: (Please review Bylaw 54) The last seven calendar days prior to, and including, the last day of classes are free from any procedures for which a mark will be assigned. (Extensions on compassionate grounds are excluded). (In the case six weeks courses, the last three calendar days before the start of the examination period are free from any assessment procedures).
1.B. Unannounced quizzes/graded activities will not exceed 5% of the final grade.
1.C. Participation marks in online courses will not exceed 20% of the final grade.

2. The final exam schedule is announced by the Registrar's office, normally after the add/drop period, and students are expected to be available for the entire exam period and not make any prior travel plans, vacations, or other commitments until after the exam dates are announced. No alternate exams accommodations will be made on those grounds.

3. No forms of assessment shall be scheduled or made-due on days identified as break days such as reading weeks, holidays, or days that the University is officially closed.

SPTs:

The Student Perceptions of Teaching (SPTs) forms will be administered in the last two weeks of classes for courses 12-24 weeks in duration, in the last week of classes for courses 6-11 weeks in duration, or in the last two days of classes for courses of 5 or fewer weeks in duration. Students should be provided with up to 15 minutes at the beginning of a class to complete the SPTs online. Senate Policy

SUPPORT CONTACTS:

The School of Computer Science has a team of support staff and access to student academic advisors to assist you through any inquiries you may have about our courses and programs. Please use one of the following emails:
For CompSci undergraduate programs and advising, including IT certificate: csinfo@uwindsor.ca
For CS Tutors (free tutoring support for all CS undergrad courses): http://tutor.cs.uwindsor.ca/
For Computer Science Society: https://css.uwindsor.ca/
For CompSci graduate programs (MSc, MSc-AI stream, and PhD): csgradinfo@uwindsor.ca
For CompSci professional graduate programs (MAC/MAC-AI stream): macprogram@uwindsor.ca
For the office of the Director of the School of Computer Science: csdir@uwindsor.ca
For CompSci technical support: https://help.cs.uwindsor.ca/
For International Student Centre: https://www.uwindsor.ca/international-student-centre/
For Student Accessibility Services: https://www.uwindsor.ca/studentaccessibility/
For other general inquiries: https://ask.uwindsor.ca/
For Student counselling services (ext. 4616): https://www.uwindsor.ca/studentcounselling/
For Student health services (ext. 7002): https://www.uwindsor.ca/studenthealthservices/
For Student Peer Support Centre (ext. 4551): https://www.uwindsor.ca/studentexperience/wellness/
For USci Faculty of Science student support network: https://www.uwindsor.ca/science/usci/

Good2Talk provides free, 24/7 single-session professional counselling and referral by phone to post-secondary students in Ontario. Services are provided in English and French, with translation services available in 100+ languages.

- Call: 1-866-925-5454 (reach professional counsellors)

- Text: GOOD2TALKON to 686868 (reach trained volunteers)

Wellness Together Canada provides free, 24/7 professional mental health and substance use counselling by phone to anyone in Canada and Canadians abroad. Service is provided in English and French, with translation services available by request.

- Call: 1-866-585-0445 (reach professional counsellors)
- Text: WELLNESS to 686868 (reach trained volunteers)

STUDENT ACCOMMODATIONS:

Students with disability:
Students who require academic accommodations in this course due to a documented disability must contact an Advisor in Student Accessibility Services (SAS) to complete SAS Registration and receive the necessary Letters of Accommodation. After registering with SAS, you must present your Letter of Accommodation and discuss your needs with the course instructor as early in the term as possible. Please note that deadlines for the submission of documentation and completed forms to SAS are available on their website:

- http://www.uwindsor.ca/studentaccessibility/

Exam conflicts:
If you have a conflict with two exams at the same time, you will need to talk to both instructors and ask which one is willing to move your exam to a different day or time.

If you have a conflict with examinations due to the following reasons, view the Office of Registrar Alternative Final Exam Policy:

- Conflict with religious conviction during the regularly scheduled time slot.
- Three or more final examinations in a 24-hour period.

Religious Observances:
Requests for accommodation of specific religious or spiritual observance must be presented to the instructor no later than 2 weeks prior to the conflict in question (in the case of final examinations within two weeks of the release of the examination schedule). In extenuating circumstances, this deadline may be extended. If the dates are not known well in advance because they are linked to other conditions, requests should be submitted as soon as possible in advance of the required observance. Timely requests will prevent difficulties in arranging constructive accommodations.
religious_accommodation_for_students.01mar2013.web_ver.pdf (uwindsor.ca)

PRIVACY AND COPYRIGHTS:

Content confidentiality:
Lectures, examinations, quizzes, assignments, and projects given in this course are protected by copyright. Reproduction or dissemination of examinations or the contents or format of examinations/quizzes in any manner whatsoever (e.g., sharing content with other students or websites), without the express permission of the instructor, is strictly prohibited. Students who violate this rule or engage in any other form of academic dishonesty will be subject to disciplinary action under Senate Bylaw 31: Student Affairs and Integrity.

Recording of lectures:
Lectures and discussions can be recorded by requesting explicit permission from the instructor. Students planning to do so shall send a request (via email is sufficient) before the lecture is delivered. Students, however, are not allowed to post or share any recorded material to any other individual or party outside of this course.
See Senate Policy on recording lectures.

SAFETY, ACADEMIC INTEGRITY, AND NON-ACADEMIC MISCONDUCT:

Equity, Diversity, and Inclusiveness (EDI)
This course, along with all its components such as lab sections are, without question, safe places for students of all races, genders, sexes, ages, sexual orientations, religions, disabilities, and socioeconomic statuses. Disrespectful attitude, sarcastic comments, offensive language, or language that could be translated as offensive and/or marginalize anyone are absolutely unacceptable. Immediate actions will be taken by the instructor to protect the safety and comfort of the students. An ethnically rich and diverse multi-cultural world should be celebrated in the classroom. The instructor, too, must treat every student equally and with the respect and compassion that all students deserve. Furthermore, UWindsor is committed to combatting sexual misconduct. All members are required to report any instances of sexual misconduct, including harassment and sexual violence, to the Sexual Misconduct Response & Prevention Office so that the victim may be provided appropriate resources and support options.

- https://www.uwindsor.ca/sexual-assault/

- For police/ambulance emergency call 911 (in Canada)
- For campus police call 519-253-3000 ext. 4444 for emergency, and 1234 for non-emergency issues.

Academic Integrity
Please refer to: https://www.uwindsor.ca/academic-integrity/

As defined in the University of Windsor's **Student Code of Conduct**, plagiarism is the act of copying, reproducing or paraphrasing significant portions of one's own work, or someone else's published or unpublished material (from any source, including the internet), without proper acknowledgement, representing these as new or as one's own.

Tips and resources to help you prevent plagiarism:
https://www.uwindsor.ca/academic-integrity/sites/uwindsor.ca.academic-integrity/files/tips_for_preventing_plagiarism.pdf

The instructor will put a great deal of effort into helping students to understand and learn the material in the course. However, the instructor will not tolerate any form of cheating. The instructor will report any suspicion of academic integrity to the Director of the School of Computer Science. If sufficient evidence is available, the Director will begin a formal process according to the University Senate Bylaws which will lead to more review, a strict punishment if convicted, and a note on your permanent student record.

The following behaviours will be regarded as cheating:
- *Copying assignments or quizzes or presenting someone else's work as your own.*
- *Allowing another student to copy an assignment/project from you and present it as their own work; protect your own work and never share it with anyone!*
- *Copying from another student or any other unauthorized source during a test or exam.*
- *Falsifying your identity during the exam or having someone else assist or complete your assessment.*
- *Referring to notes, textbooks, and any unauthorized sources during a test or exam (unless otherwise stated).*
- *Speaking or communicating without permission during a test or exam.*
- *Not sitting at the pre-assigned seat during a test or exam.*
- *Communicating with another student in any way during a test or exam.*
- *Having unauthorized access to the exam/test paper prior to the exam/test.*
- *Explicitly asking a proctor for the answer to a question during an exam/test.*
- *Modifying answers after they have been marked.*
- *Any other behaviour which attempts unfairly to give you some advantage over other students during the grade-assessment process.*
- *Refusing to obey the instructions of the officer in charge of an examination.*

The list given above is not exhaustive. More examples are given in Appendix A, Senate Bylaws 31 – Complete guidelines and procedures on the sanctions imposed by the university are also listed in Table A.1 of the Senate Bylaws 31

In this course any assessment that is deemed plagiarized or in violation of the academic integrity policy will NOT BE GRADED and receive a grade of ZERO unless a different ruling is provided by the adjudication committee formally reviewing the case.

Examples of sanctioning include: *(from Table A.1 in Appendix A of Bylaw 31)*
For first offence: mark reduction up to zero, censure 6-12 months; and for subsequent offence: suspension 4-24 months, censure up until graduation.

Plagiarism detection software:
Plagiarism-detection software *SafeAssign* will be used for all student assignments in this course. You will be advised how to submit your assignments. Note that students' assignments that are submitted to the plagiarism-detection software become part of the institutional database. This assists in protecting your intellectual property. However, you also have the right to request that your assignment(s) not be run through the student assignments database. If you choose to do so, that request must be communicated to the course instructor in writing at the beginning of the course. The instructor reserves the right to choose another plagiarism detection software and students would be notified of this once it is put in use.