Comp 8380: Information Retrieval Systems (2021 Winter)

The instructor is Professor Jianguo Lu from School of Computer Science, University of Windsor.

- Email: jlu@uwindsor.ca
- Course web site: Blackboard system and `http://cs.uwindsor.ca/~jlu/538`
- Instructor's web site: `http://cs.uwindsor.ca/~jlu`
- Office hours: Monday and Wednesday 10:00-11:00.

1 Course overview

Information retrieval (IR) is finding documents written in unstructured natural language text from within large collections. Text are usually interconnected, such as hyper links in the case of Web documents, and citations in the case of academic papers. Thus, the key issues in IR is to analyze text and graphs in large scale. This course will cover the basic techniques in text and graph analysis. The tentative topics covered in this course include the following:

Text analysis It will cover the following topics:
- Statistic properties of text, power laws, Zipf's law, Heaps' law.
- Language models, unigram model and bigram model. Smoothing techniques (Laplacian Smoothing and Good Turing amoothing)
- 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,

Graph Analysis it will cover
- Link analysis. PageRank algorithm. Markov chain.
- Graph representation. DeepWalk and Node2vec algorithms.

Search Engine Construction It includes:
- Searching, retrieving documents that contain a given query token from the inverted index; Use Lucene to construct a practical and large search engine.

2 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 and TF-IDF information retrieval model;
- Understand the link/graph analysis, in particular PageRank algorithm;
- Classify and cluster documents using various machine learning algorithms;
- Apply deep learning in text and graph analyses.

3 Grading scheme

There will be an exam and a project. The weight of the exam is 50%, the project is 50%. The project is to implement and compare text and graph analyses algorithms. For each topic, I will give starter code explained in Notebook. Your project is to expand some of the topics, and explain your work in Jupyter Notebook.

4 Text books

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.

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.

LA Michael McCandless, Erik Hatcher, and Otis Gospodnetic, Lucene in Action, Second Edition. 2010.

5 Feeling Overwhelmed?

From time to time, students face obstacles that can affect academic performance. If you experience difficulties and need help, it is important to reach out to someone.

For help addressing mental or physical health concerns on campus, contact (519) 253-3000:

- Student Health Services at ext. 7002 (http://www.uwindsor.ca/studenthealthservices/)
- Student Counselling Centre at ext. 4616 (http://www.uwindsor.ca/studentcounselling/)
- Peer Support Centre at ext. 4551

My Student Support Program (MySSP) is an immediate and fully confidential 24/7 mental health support that can be accessed for free through chat, online, and telephone. This service is available to all University of Windsor students and offered in over 30 languages. Call: 1-844-451-9700, visit https://keepmesafe.myissp.com/ or download the My SSP app: Apple App Store/Google Play.

A full list of on- and off-campus resources is available at http://www.uwindsor.ca/wellness. Should you need to request alternative accommodation contact your instructor, head or associate dean. For the revised bylaws, go to: www.uwindsor.ca/policies

6 Course Regulations

Can I record lectures? Course materials prepared by the instructor are considered by the University to be an instructor's intellectual property covered by the Copyright Act, RSC 1985, c C-42. These materials are made available to you for your own study purposes, and cannot be shared outside of the class or "published" in any way. Students who do not have the necessary accommodations are not permitted to record lectures in any format (audio, video, photograph, etc.). Posting course materials or any recordings you may make to other websites without the express permission of the instructor will constitute copyright infringement.

Bylaws 31 The following behaviour will be regarded as cheating (this list is not exhaustive. For more examples please refer to Senate Bylaws 31 (click here for the entire document):

- Copying lab assignments or presenting someone else's work as your own. Note that it is still considered as copying if you make trivial changes in programs, such as changing variable names and formatting. There are plagiarism detection programs that targeting software code.
- Allowing another student to copy an assignment from you and present it as their own work

Exam Content Confidentiality 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), 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.