Concordia University
Department of Computer Science and Software Engineering
COMP 479/6791

Instructor: Dr. Sabine Bergler, bergler AT cse.concordia.ca, office: EV3.283,
office hours: Tuesdays 15:00-16:00, Wednesdays 15:00-16:00 or by appointment

Lectures: Tuesday, Thursday 13:15-14:30 in H 431 SGW

Labs: Tuesday, Wednesday, Thursday 14:45-16:35 in H 811

Calendar Description COMP 479 Information Retrieval and Web Search (4 credits) Prerequisite: COMP 233 or ENGR 371; COMP 352. Basics of information retrieval (IR): boolean, vector space and probabilistic models. Tokenization and creation of inverted files. Weighting schemes. Evaluation of IR systems: precision, recall, F-measure. Relevance feedback and query expansion. Application of IR to web search engines: XML, link analysis, PageRank algorithm. Text categorization and clustering techniques as used in spam filtering. Project. Lectures: three hours per week. Laboratory: two hours per week.

Graduate Attributes

Attribute 1: Knowledge-base for Engineering: Text cleaning. Tokenization of text. Information Retrieval principles: Indexing. Search. Map Reduce. Vector Space modelling. Flat and Hierarchical Clustering.

Attribute 4: Design: Design a complete web crawling and indexing system. Design a way to assess and compare predominant sentiment of web pages.

Attribute 5: Use of Engineering tools: Use of Linux, Java, and ancillary support tools such as Eclipse, as well as specific algorithms that have to be implemented or adapted from open source implementations.

Attribute 6: Individual and team work: Project 1 requires individual implementation of a given algorithm. Project 2 requires an individual experiment using Project 1 code. The final Project requires the design and implementation of a complex team project including web crawling, indexing web pages, sentiment analysis. The Final Project has to be presented in a 3min presentation. Project 1 and the Final Project have to be demonstrated to the lab instructor.

Textbook Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Cambridge University Press, 2008. Web publication at `http://informationretrieval.org`

Grading Scheme This course requires continuous preparation of course material. Preparedness and class participation, as well as lab participation influence the final grade. There is a first midterm examination, worth 25% of the final grade and a second midterm worth 35% of the grade. Exams test theoretical knowledge, such as understanding of the algorithms, their complexity, but also of application issues and design features for practical, large scale systems. For students who have better marks on the Final Exam than the Midterm (in percent), the Final counts for the combined 60%. Project 1, an individual project developed partly during lab time, counts 10% for undergraduates, 9% for graduates. Project 2, an individual assignment, counts 5% for all students. The final project, an individual project, counts 22% for undergraduates, 20% for graduates. Project presentations are graded. Graduate students have to complete a special project for 3%. Excellent assignment submissions may be awarded bonus points. Labs are mandatory, attendance will be taken randomly.

Midterm 1	25%	
Midterm 2	35%	
Project 1	9% (UG), 8% (G)	due tba
Project 1 Demo	1%	
Project 2	5%	due tba
Project 3	21% (UG), 19% (G)	due tba
Project 3 Demo	1%	
Graduate Project	3%	due tba

Submission of deliverables via Moodle before the deadline.

Requirements regarding expectations of originality Faculty Council approved the Expectations of Originality form. The purpose of the Expectations of Originality form is to remind students of the requirement not to plagiarize and of their obligation to submit original work.

Students have to fill out one copy of the form for each course, at the beginning of the semester, and submit it to Moodle. The student should write on the cover page of each assignment, lab report or project the statement:

> *I certify that this submission is my original work and meets the Faculty's Expectations of Originality*

and sign the statement, write the date and write his/her I.D. number, scan it in and submit through Moodle.

For group work, the statement is:

> *We certify that this submission is the original work of members of the group and meets the Faculty's Expectations of Originality.*

All the students in the group have to sign the statement with the date and their I.D. number, scan it in and submit with their project reports.

Syllabus

Readings mandatory before the class for which they are assigned.

Date	Topic	Readings Ch.	Lab work
3., 5.9.	Boolean retrieval	1	Unix. grep. Reuter's corpus RCV1
10.9.	Term vocabulary and postings lists	2	Porter Stemmer
12.,17.9.	Dictionaries and tolerant retrieval	3	
19.9.	Index construction	4	
24.,26.9.	Index compression	5	
1.,3.10.	Scoring, term weighting, and the vector space model	6	Web crawling
8.10.	Computing scores in a complete search system	7	
10.10.	Evaluation in information retrieval	8	
15.,17.10.	Relevance feedback and query expansion	9	
19.10.			Grad Review paper due
22.10.	Review	1-9	
24.10.	Midterm 1	1-9	
29.10.	Classification and Naive Bayes	13	
31.10.,5.11.	Vector space classification	14	
7.,12.,11.	Support vector machines	15	
14.,19.11.	Learning to rank (LTR)	15	
21.11.	Midterm 2	Chs.1-9, 13-17	
26.11.	Flat clustering	16	FP demos
28.11.	Hierarchical clustering	17	FP demos