Concordia University
Department of Computer Science and Software Engineering

COMP 479/6791

Instructor: Dr. Sabine Bergler. Contact the instructor through Moodle channels only for course related issues. Ask content questions on the *CourseContentQuestions* forum, so that all students can benefit from the answer. Contact instructor and TA for personal questions using the Moodle direct message system. Make sure you include the course number in all Moodle message communications.

Lectures: Content is delivered asynchronously through Moodle. If needed, lecture time (Tuesdays and Thursdays from 1:15-2:30) is used for Zoom meetings for questions regarding material. Due to the size of the course, using the *CourseContentQuestions* forum on Moodle for questions is strongly advised; I will give priority to this forum and attempt to answer all questions.

Labs: Lab instructors will answer questions regarding the projects in the *LabQuestions* forum on Moodle. Scheduled lab times will be used for Zoom meetings as needed.

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.

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

Environment for text pre-processing NLTK: Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit. online. Steven Bird, Ewan Klein, and Edward Loper. Formerly O'Reilly. Web publication at `https://www.nltk.org/book/`

Grading Scheme This course requires continuous preparation of course material. Preparedness and constructive participation in course and lab fora may influence the final grade.

There is a first midterm examination, worth 10%, a second midterm worth 15%, and a third 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. All three midterms will be delivered asynchronously as Moodle quizzes and consist mainly of multiple choice questions.

Four projects have to be developed by each student individually. Project 1 concerns preprocessing a text collection for IR using NLTK and counts 12%. Project 2 requires an implementation of an inverted index in the naive way and counts 8%. Project 3 addresses term weighting for ranking, including tf/idf and counts for 8%. Project 4 requires to put all previously developed components together and extend them for sentiment analysis of web pages and counts for 12%.

Midterm 1	10%	2.10.
Midterm 2	15%	30.10.
Midterm 3	35%	28.11.
Project 1	12%	30.9.
Project 2	8%	21.10.
Project 3	8%	11.11.
Project 4	12%	4.12.

Submission of deliverables via Moodle before the deadline. Deadlines mentioned here are suggestive and may change; consult project assignments on Moodle for updates.

Syllabus

Readings mandatory before the class for which they are assigned. This syllabus is suggestive and subject to change. Please consult the Moodle page for updates. Due dates, in particular, only illustrate course planning.

Week	Topic	Ch.	Due
1	Text pre-processing with NLTK, Boolean retrieval	1	
	Term vocabulary and postings lists, Porter Stemmer	2	
2	Dictionaries and tolerant retrieval	3	
3	Index construction	4	P1
4	Index compression	5	M1
5, 6	Scoring, term weighting, and the vector space model	6	P2
7	Web crawling, Computing scores in a complete search system	7	
8	Evaluation in information retrieval	8	M2
9	Relevance feedback and query expansion	9	P3
10	Vector space classification	14	
	Support vector machines	15	
11	Flat clustering	16	
	Hierarchical clustering	17	
12	Learning to rank	15	M3
13			P4

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.

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.

Remote teaching format for this course

This course is taught remotely. This means there is no need for students to be physically present in Concordia buildings to take and pass this course.

This course is taught asynchronously. This means that there is no Zoom lecture that every student has to follow at the same time in order to follow and pass this course.

The lecture material is covered perfectly by the textbook. The slides used are from a co-author, Hinrich Schütze. Both materials are linked to in the Moodle page.

A narrated version of the slides is available as a Yuja video. This narrated version is not as animated, as a lecture with direct contact and feedback can be.

Lecture time will not be used for lectures, but for questions regarding the course material assigned. Readings and viewings/listenings have to be done individually and in a proactive fashion, before the Q/A session. Due to the size of the class, not all students can ask all their questions during the Q/A session.

The instructor has to be contacted through Moodle, any email messages will be ignored. Personal and confidential communication has to be sent through direct Moodle messaging. Course questions have to be asked through the Moodle Forum CourseContentQuestions. Practical questions for the TAs have to be posted to the Forum LabQuestions. Please read all provided materials and attempt to answer your own question, possibly by discussing it with your classmates on the Student Exchanges Forum.

This outline, the syllabus and indeed the course design are subject to change during the class, it is therefore important to follow the progression of the class and refer to the documents on Moodle to get the latest updates.

General Conditions Specific to Remote Teaching and Assessment for Fall 2020

All students are expected to have access to a computer with following capabilities

1. reliable internet connection
2. camera and microphone (your computer and/or cellphone)
3. document scanning application such as Adobe Scan app (`https://play.google.com/store/apps/details?id=com.adobe.scan.android&hl=en_CA`)

All students should install VPN for remote desktop access to Concordia University computer labs

`https://www.concordia.ca/it/support/connect-from-home.html` Once you have VPN connection to Concordia University, you can access to all available software in Gina Cody School labs by following the process described in: `https://www.concordia.ca/ginacody/aits/support/faq/connect-from-home.html`

To download Microsoft Office 365: https://www.concordia.ca/it/services/office-365-education.html (optional)

All students are expected to do online, timed exams

1. midterm exams will be through Moodle Quiz
2. you will need a quiet place within which to take the exam
3. the course instructor reserves the right to conduct an individual oral examination to verify student?s response to online exam questions

Academic Integrity

Violation of the Academic Code of Conduct in any form will be severely dealt with. This includes copying (even with modifications) of program segments. *You* must demonstrate independent thought through your submitted work. The Academic Code of Conduct of Concordia University is available at: `https://www.concordia.ca/conduct/academic-integrity.html`

It is expected that during class discussions and in your written assignments you will communicate constructively and respectfully. Sexist, racist, homophobic, ageist, and ableist expressions will not be tolerated.

Third-party software/website and personal information

Note that, as a part of this course, some or all of the lectures and/or other activities in this course may be recorded. Recordings will be focused on the instructor and will normally exclude students. It is possible, however, that your participation may be recorded. If you wish to ensure that your image is not recorded, speak to your instructor as soon as possible. Also, please note that you may not share recordings of your classes and that the instructor will only share class recordings for the purpose of course delivery and development. Any other sharing may be in violation of the law and applicable University policies, and may be subject to penalties.

Third-party software/website usage for work submission

Students are advised that external software and/or websites will be used in the course and students may be asked to submit or consent to the submission of their work to an online service. Students are responsible for reading and deciding whether or not to agree to any applicable terms of use. Use of this software and service is voluntary. Students who do not consent to the use the software or service should identify themselves to the course instructor as soon as possible to discuss alternate modes of participation that do not require them to give copyright or the right to use their work to a third party. By using the external software or websites, students agree to provide and share their work and certain personal information (where applicable) with the website/software provider. Students are advised that the University cannot guarantee the protection of intellectual property rights or personal information provided to any website or software company. Intellectual property and personal information held in foreign jurisdictions are subject to the laws of such jurisdictions.

On Campus Resources

`https://www.concordia.ca/content/dam/ginacody/miae/docs/undergraduate/CampusResources-offices.pdf`

ADDENDUM (general text, adapt yourself for this course)

ACADEMIC CONDUCT ISSUES THAT APPLY IN GENERAL
The basic ten rules that make you a good engineer

The B. Eng. program is set to satisfy most of the requirements for your education and prepares you for a professional engineering career that requires dedication and knowledge. What you learn, and how you learn, will be used extensively in your engineering profession for the next 30 to 40 years. Therefore, the four years spent in the engineering program are crucial towards your professional formation. The first step is for you to learn to “think like an engineer” which means:

- accept responsibility for your own learning
- follow up on lecture material and homework
- learn problem-solving skills, not just how to solve each specific homework problem
- build a body of knowledge integrated throughout your program
- behave responsibly, ethically and professionally

One of the mainstays of being a professional engineer is a professional code of conduct and as an engineering student this starts with the Academic Code of Conduct (Article 16.3.14 of the undergraduate calendar). However, you may encounter situations that fall outside the norm and in such cases, you use your common sense.

Further, the following issues should be given serious consideration:

1) Attendance at lectures and tutorials are major learning opportunities and should not be missed. The labs represent a unique opportunity for you to acquire practical knowledge that you will need in your career. Class and tutorial attendance is important for you to comprehend the discipline and make the connections between engineering skills. You are strongly encouraged to participate in the class, ask questions and answer the instructor’s questions. Tutorials are just extensions of the classes in which application of the concepts presented during the lectures are presented and problems are practically solved.

2) The decision to write tests that are not mandatory is entirely yours. For example, midterm test are often stated in many courses as optional. However, one the objectives of midterms is to check on your comprehension of the material and allow time for whatever action is necessary (from more study time to discontinuing a course). Plan to attend the class tests even if they are not mandatory. If you pay attention in the lectures, it will take you significantly shorter time to comprehend the material. Note also that if you are in the unfortunate position of being unable to write a final exam due to medical reasons and seek a deferral, this may not be possible if the instructor has no information indicating that you have been attending the course and assimilating the material (ie through midterms, quizzes, assignments etc).

3) Homework is usually mandatory and it has some weight in the final grade (such information is given in the course outline). Homework may also be conceived as training material for the class tests. Under all circumstances, it is highly recommended to carry out the home work on time and submit it on the prescribed date. Late submissions are not granted to individual cases regardless of the reason. This is part of the training for being in the workforce where deadlines have to be met. Please, plan your work such that you submit all the assignments and lab reports on time and in the correct place (not in the corridor or on the street!).

4) Office hours with tutors, lab instructors or class instructors are listed in the course outline/website/office doors. Please respect these office hours and in case you have a serious conflict, contact the instructor asking for a special time arrangement.

5) Class tests (midterms, quizzes) are returned to the student. The final exams are not. If you wish to see your exam paper, be aware that most instructors allow only a narrow window of time for that purpose. For the fall term, exams may usually be reviewed in January and May for the spring term.

6) When you see your marked work (assignments, midterms, final exam etc), be aware that you are supposed to review your material and see the type of errors you made and if marks have been added incorrectly. This is not an opportunity to try and “negotiate” a higher grade with the instructor. If you believe that your grade is not right, you may apply for a formal Course Reevaluation through the Birks Student Centre.

7) Writing tests and exams represents a major component of your course work. These tests and exams have rigorous requirements such as:

- No cell phone or other communication enabling tool is allowed on the student during the examination period.

- Only specified faculty calculators are allowed during tests and exams unless otherwise indicated by the instructor.
- Usually, no materials are allowed in the exam unless otherwise announced.

Get used to signing in and out of your exam. Make sure that you leave your exam papers with the invigilator. There are rules concerning general exam issues in the UG Calendar. These requirements are there to eliminate any possible misunderstanding and you are asked to respect the rules. Disciplinary measures are taken when the rules are not followed.

8) Respect your colleagues and those that you meet during the class: tutors, instructors, lab instructors, technical personnel, assistants, etc. Use appropriate communication means and language. Be considerate for all human beings. This includes small things such as turning off cell-phones before a class begins. Concordia University is a very diverse group of people and a very large multicultural community.

9) Communication is part of your future profession. Learn how to communicate effectively and efficiently in the shortest time possible. Write short but meaningful e-mails, make effective phone calls, etc. If your instructor accepts emails make sure that your request is clear with the course number and your name in the Subject line. Do not ask for special treatment as instructors have to treat all students equitably.

10) Respect all the above and you will get closer to your future profession.