

<p style="text-align: center;">COMP 479/6791 Information Retrieval and Web Search Department of Computer Science and Software Engineering Fall 2023</p>	
<p>Course Instructor: <i>Sabine Bergler</i></p> <p>Contact : for emergencies and private communication: <i>sabine.bergler@concordia.ca</i> for course related communication: <i>Moodle messages</i></p>	
<p>Office Hours: <i>Tuesdays and Wednesdays, 11am</i></p>	
<p>Labs: COMP 479/6791 D DI Th 2:45-4:35pm COMP 479/6791 D DJ We 2:45-4:35pm COMP 479/6791 D DK Tu 2:45-4:35pm</p>	
<p>Course 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.</p>	
<p>Prerequisites: COMP 233 or ENGR 371; COMP 352 Co-requisites: <i>N/A</i></p>	
<p>Specific Knowledge and Skills Needed for this Course: Students taking this course are expected to have sufficient knowledge of the following topics. Should you have difficulties in any of these topics, you are strongly encouraged to review them before the DNE deadline.</p> <p><i>Programming concepts: recursive programming. Lists and vector data structures and their manipulation. Basic familiarity with Python. Processing large files and managing local file structures. Manipulation of ASCII files without tools.</i></p> <p><i>Knowledge: discrete mathematics, especially relations. Basic analysis of complexity of algorithms.</i></p> <p><i>Skills: writing an informative academic report.</i></p>	

Course materials

Required 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

Midterm 1	25%	17.10.2023
Midterm 2	35%	23.11.2023
Project 1	5%	20.9.2023
Project 2	10%	9.10.2023
Project 3	10%	6.11.2023
Project 4	15%	5.12.2023

Tentative Course Schedule

Week	Topic	Readings IIR Ch
1	Boolean retrieval, Term vocabulary and postings lists	1, 2 --- NLTK Ch.3
2	Dictionaries and tolerant retrieval	3
3	Index construction	4, 6
4	Index compression	5
5	Scoring, term weighting, and the vector space model	
Midterm break		
6	Evaluation in information retrieval, Relevance feedback	7, 8, 9
7	Midterm	
8	Pseudo-relevance feedback and query expansion, Naïve Bayes classification	9, 13
9	Vector space classification	14
10	Support vector machines	15
11	Flat clustering, Midterm	16
12	Hierarchical clustering, Cluster evaluation	17

Lab Details

- The NLTK toolkit with accompanying textbook is the **required** preprocessing environment for all projects: *Natural Language Processing with Python* by Steven Bird, Ewan Klein, and Edward Loper, O'Reilly. See <http://www.nltk.org/book/>
- Projects increasingly leave design options open. Students are to discuss different ideas during Labs. Students bring discussion topics to the Lab Instructor. Offline student discussions outside the Moodle portal are discouraged, on Moodle they are encouraged

Other information

- Exams test theoretical knowledge, Projects test practical implementation of the principles taught. Thus full marks in projects can only be obtained, when the project abides by the principles introduced in the lectures. Projects are individual projects.
- For students who do better on the second midterm, the second midterm counts for the entire 60% of exam marks.

Graduate Attributes:

The following is the list of graduate attributes (skills) that students use, learn and/or apply throughout the term.

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: Projects 1 and 2 require individual implementation of given algorithm s and pipelines. Project 2 requires an individual experiment using Project 2 code. The final Project requires the design and implementation of a complex project including web crawling, indexing web pages, ranking web pages, sentiment analysis. Project 1 and the Final Project have to be demonstrated to the lab instructor.

Course Learning Outcomes (CLOs):

By the end of this semester, students are expected to master the following concepts.

- basic text processing
- inverted index and its implementation
- search
- vector space model
- uses of the vector space model for classification and clustering
- web crawling

On Campus Resources

Please visit [Student services at Concordia University](#) for the services available Gina Cody School students.

