COMP 479/6791
Information Retrieval and Web Search
Department of Computer Science and Software Engineering
Fall 2024

Course Instructor:
Sabine Bergler, PhD

Contact :
for emergencies and private communication: *sabine.bergler@concordia.ca*
for course related communication: *Moodle messages*

Office Hours (in ER 1143):
Mondays or Tuesdays, 15:00-16:00 or by appointment

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

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.

Prerequisites: COMP 233 or ENGR 371; COMP 352
Co-requisites: *N/A*

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 and see the professor.

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.

Knowledge: discrete mathematics, especially relations. Basic analysis of complexity of algorithms. Graphs. Vectors and matrices.

Skills: writing an informative academic report.

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	20%	22.10.2024
Midterm 2	40%	21.11.2024
Project 1	15%	10.10.2024
Project 2	25%	28.11.2024

Tentative Course Schedule

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

Lab Details

- The NLTK toolkit 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 leave design options open where better choices lead to more points. 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.

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.