

Master Course Syllabus
School of Engineering and Computer Science
Washington State University Vancouver

CS 454
Information Retrieval
3 Semester Hours
(3 lecture hours)

Catalog Description

Fundamental principles and algorithms, related to information retrieval: text querying, indexing, and retrieval methods, evaluation metrics, ranking, semi-structured data, crawling and scraping of the web using link-based algorithms, and user interface issues.

Prerequisite Courses

- CS 223 with a C or better
- CS 224 with a C or better

Prerequisite Topics

- Programming proficiency
- Object oriented concepts and modeling
- Use of Unix environment for coding, compilation, debugging, and testing
- Function, relation, and set theory
- Sorting and searching algorithms

Measured Course Outcomes

Students taking this course will:

1. Analyze a document corpus to identify correct document retrieval methods using keyword search (Contributes to performance criterion 1-b.)
2. Interact professionally with team members (Contributes to performance criterion 5-b.)
3. Effectively contribute as a team member to create an efficient and effective information retrieval search engine. (Contributes to performance criterion 5-c.)
4. Solve information retrieval problems by selecting or implementing correct retrieval, ranking, and satisfaction algorithms. (Contributes to performance criterion 6-c.)

Required Textbooks

Introduction to Information Retrieval, Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, Cambridge University Press, 2008.

Major Topics Covered in the Course

1. Web scraping and web data collection policies and methods
2. Retrieval methods including boolean and vector models
3. Ranking returned tuples to users based on relevance
4. Determining user satisfaction of returned results
5. Providing suggestions based on similar users or history

6. Interfacing with a database through a web client

Projects

Programming Project Area	Weeks
Web scraping/crawling	2
Ranking results for a given corpus using static methods	2
Determine user satisfaction using standard metrics	2
Implement an inverted index and process keyword queries	2
Information retrieval search engine project	4

Design, Implementation and Analysis

Students are guided through the challenges that unstructured and semi-structured data present when handling keyword queries along with some solutions. First, students will scrape or crawl the web to create their own corpus. Students then use some corpus to process keyword queries and rank the results using algorithms covered in class. Using a search log, students then determine how well the ranked results satisfy the users information needs. Using the corpus previously scraped or crawled, students will create an inverted index and use popular techniques (stopping, stemming, etc.) to efficiently search a large collection of documents. Finally, students will combine all these steps into a complete search engine with a web interface.

This course has multiple projects, each designed to tackle a portion of the information retrieval model. When implementing the inverted index, students use their own data that was scraped or crawled earlier. User satisfaction and ranking of results is performed over some provided corpus with the necessary information, such as judgment scores. The final group project asks students to create a full stack web search engine, combining all the techniques covered in the class so far.

CS2013

This course provides coverage of CS2013 knowledge areas. Values listed are minimum course hours dedicated to the topic, percentages indicate the fraction of CS2013 knowledge area topics covered (acceptable values are: <25%, 25-75%, >75%, or 100%).

Area	Tier 1	Tier 2	Elective
IM/Information Management Concepts	2 (<25%)	6 (25-75%)	
IM/Data Modeling		1 (<25%)	
IM/Indexing			6 (>75%)
IM/Information Storage and Retrieval			9 (25-75%)
IS/Natural Language Processing			2 (<25%)
PBD/Web Platforms			2 (<25%)

Course Coordinator:	Ben McCamish
Last Updated:	October 5, 2020
Syllabus Version Number:	1.0