The Hong Kong University of Science and Technology

UG Course Syllabus (Spring 2025-26)

[Course Title] Search Engines for Web and Enterprise Data

[Course Code] COMP 4321

[No. of Credits] 3-credit

[Any pre-/co-requisites] Prerequisite(s): COMP 2011 OR COMP 2012 OR COMP 2012H

Name: Dr. Kenneth LEUNG

Email: kwtleung@cse.ust.hk

Course Description

Text retrieval models, vector space model, document ranking, performance evaluation; indexing, pattern matching, relevance feedback, clustering; web search engines, authority-based ranking; enterprise data management, content creation, meta data, taxonomy, ontology; semantic web, digital libraries and knowledge management applications.

List of Topics

1. Introduction and course overview	6. Document preprocessing
2. Business models	7. Query expansion and relevance feedback
3. Information retrieval models and Inverted Files	8. Machine learning for document ranking
4. Web-based information retrieval	9. Enterprise search
5. Retrieval effectiveness, benchmarking	10. Applications: text summarization

Intended Learning Outcomes (ILOs)

By the end of this course, students should be able to:

1. Understand the history, evolution, impacts and challenges of web-scale search engine.
2. Understand information retrieval models, document indexing, searching and ranking.
3. Evaluate the performance of search algorithms using performance metrics.
4. Obtain hands-on experience by implementing a complete search engine.
5. Understand state-of-the-art in the industry.

Assessment and Grading

This course will be assessed using criterion-referencing and grades will not be assigned using a curve. Detailed rubrics for each assignment are provided below, outlining the criteria used for evaluation.

[List specific assessed tasks, exams, quizzes, their weightage]**Assessment Task**	**Contribution to Overall Course grade (%)**
Quizzes	20%

Course Participation	10%
Group Project	40%
Final examination	30%
Total	100%

Assessments:

[List specific assessed tasks, exams, quizzes, their weightage, and due dates; perhaps, add a summary table as below, to precede the details for each assessment.]

Assessment Task	Contribution to Overall Course grade (%)	Due date
Quizzes	20%	By the end of each month*
Course Participation	10%	By the end of each online zoom meeting*
Group Project	10%	By the end of the semester*
Final examination	60%	Final Exam Period*

* Assessment marks for individual assessed tasks will be released within two weeks of the due date.

Mapping of Course ILOs to Assessment Tasks

Assessed Task	Mapped ILOs	Explanation
Quizzes (20%)	ILO1, ILO2, ILO3, ILO4, ILO5	ILO1: Quiz questions cover search engine history, evolution, and key challenges. ILO2: Conceptual and applied questions on retrieval models, indexing, and ranking techniques. ILO3: Interpretation of evaluation metrics (e.g., precision, recall) through MCQs or short problems. ILO4: Some quizzes include practical code snippets or algorithm flow analysis to reinforce system-level thinking. ILO5: Items testing understanding of modern practices such as semantic search, knowledge graphs, and AI-enhanced search.
Course Participation (10%)	ILO1, ILO2, ILO3, ILO4, ILO5	ILO1: Students engage in discussions about the societal impact and evolution of search technologies. ILO2: Active problem-solving and conceptual explanation during lectures and labs enhance understanding of indexing and ranking models. ILO3: Participation in peer-review and performance comparison during labs strengthens metric-based thinking. ILO4: Engagement in collaborative coding demos and project planning discussions reflects practical experience. ILO5: Involves

		reflection on guest talks or news about industry search technologies.
Group Project (10%)	ILO2, ILO3, ILO4, ILO5	ILO2: Students apply retrieval models and design systems for indexing, ranking, and querying. ILO3: The project includes testing and analyzing performance using formal IR metrics. ILO4: The core of the project is building a search engine—from crawling and indexing to ranking and interface. ILO5: Teams explore advanced IR topics like clustering, semantic similarity, and ontology integration.
Final examination (60%)	ILO1, ILO2, ILO3, ILO5	ILO1: Essay and short-answer questions assess knowledge of search engine development and impact. ILO2: Theory-based questions examine mastery of models, algorithms, and techniques. ILO3: Scenario-based questions require calculating or interpreting IR metrics. ILO5: Covers state-of-the-art tools, concepts, and technologies in current search engines and enterprise systems.

Grading Rubrics

Assessment Task	Excellent (A)	Good (B)	Satisfactory (C)	Marginal/Fail (D/F)
Quizzes (20%)	Consistently accurate, demonstrates strong conceptual and applied understanding across all ILOs including models, history, metrics, implementation, and industry trends.	Mostly correct with solid grasp of core concepts and moderate application; minor gaps in one or two ILOs.	Basic coverage of most topics; shows limited application or depth across several ILOs.	Many incorrect responses; lacks understanding of key topics; little or no evidence of ILO achievement.
Course Participation (10%)	Highly engaged in discussions and activities; contributes meaningful insights related to models, metrics, system implementation, history, and industry.	Participates actively in most sessions with generally relevant contributions across the ILOs.	Occasional contributions; superficial understanding evident; passive in collaborative discussions.	Rarely participates or off-topic; fails to demonstrate grasp of ILOs.

Group Project (10%)	Implements a functional search engine with solid use of retrieval models, evaluation metrics, and state-of-the-art features; excellent teamwork and documentation.	Functional system with good application of IR concepts and metrics; some innovation or research evident.	Partially working system with limited depth; minimal connection to modern techniques or evaluation.	Incomplete or poorly executed project; major gaps in system, metrics, or industry relevance.
Final Examination (60%)	Demonstrates mastery of historical, theoretical, evaluative, and industry-related concepts; clear, well-supported answers.	Good overall performance; answers show clear logic and reasonable understanding of IR concepts.	Covers essential content with basic explanations; minor misunderstandings or lack of depth.	Inaccurate or incomplete answers; lacks understanding of core ILOs (excluding implementation).

Final Grade Descriptors:

Grades	**Short Description**	**Elaboration on subject grading description**
A	Excellent Performance	[Example: Demonstrates a comprehensive grasp of subject matter, expertise in problem-solving, and significant creativity in thinking. Exhibits a high capacity for scholarship and collaboration, going beyond core requirements to achieve learning goals.]
B	Good Performance	[Example: Shows good knowledge and understanding of the main subject matter, competence in problem-solving, and the ability to analyze and evaluate issues. Displays high motivation to learn and the ability to work effectively with others.]
C	Satisfactory Performance	[Example: Possesses adequate knowledge of core subject matter, competence in dealing with familiar problems, and some capacity for analysis and critical thinking. Shows persistence and effort to achieve broadly defined learning goals.]
D	Marginal Pass	[Example: Has threshold knowledge of core subject matter, potential to achieve key professional skills, and the ability to make basic judgments. Benefits from the course and has the potential to develop in the discipline.]
F	Fail	[Example: Demonstrates insufficient understanding of the subject matter and lacks the necessary problem-solving skills. Shows limited ability to think critically or analytically and exhibits minimal effort towards achieving learning goals. Does not meet the threshold requirements for professional practice or development in the discipline.]

Course AI Policy

Generative AI is allowed for the course project only.

Communication and Feedback

Assessment marks for individual assessed tasks will be communicated via Canvas within two weeks of submission. Feedback on assignments will include [specific details, e.g., strengths, areas for improvement]. Students who have further questions about the feedback including marks should consult the instructor within five working days after the feedback is received.

Resubmission Policy

All students are required to adhere strictly to the deadlines posted on Canvas. Late submissions will not be accepted under any circumstances, unless prior approval has been granted for exceptional cases. Please plan your time carefully and check Canvas regularly for assignment due dates.

Required Texts and Materials

C.D. Manning, R. Raghavan, and H. Schutze *Introduction to Information Retrieval*. Cambridge University Press, 2007.

- The pre-publication manuscript of the book and the lecture slides used in a Stanford course are available online

Academic Integrity

Students are expected to adhere to the university's academic integrity policy. Students are expected to uphold HKUST's Academic Honor Code and to maintain the highest standards of academic integrity. The University has zero tolerance of academic misconduct. Please refer to Academic Integrity | HKUST – Academic Registry for the University's definition of plagiarism and ways to avoid cheating and plagiarism.

Additional Resources

Reference books

C.J. van Rijsbergen *Information Retrieval*. 2nd Edition, Butterworth & Co (Publishers) Ltd, 1979. Online Version.

Web site for the textbook **[BR] Baeza-Yates and Ribeiro-Neto**. It does not have the book online, but it contains many useful resources and an errata.

Web site for the reference book **[FB] Bill Frakes and Ricardo Baeza-Yates**. It does not have the book online, but it contains the source codes used in the book. The soruce code will be useful for your project.

R. Baeza-Yates and Berthier Ribeiro-Neto *Modern Information Retrieval*. Addison Wesley, Essex, England, 1999.

W.B. Frakes and R. Baeza-Yates (Eds.) *Information Retrieval: Data Structures and Algorithms*. Prentice- Hall, Englewood Cliffs, NJ, 1992.
G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.

G. Salton, and M.J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, 1983.