

# INFORMATION RETRIEVAL

## 1. Thông tin về học phần (General Information)

**Tên học phần (Course name):** Information Retrieval

**Mã học phần (Course code):** INT\_E14125

**Số tín chỉ (Number of credits):** 3

**Loại học phần (Course type):** Compulsory

**Học phần tiên quyết (Prerequisites):**

**Học phần trước (Previous courses):**

**Học phần song hành (Parallel courses):**

**Các yêu cầu đối với học phần (Course requirements):**

- Lecture room: Projector, black board, microphone and speaker
- Laboratory:

**Giờ tín chỉ đối với các hoạt động (Teaching and Learning hours):**

- Lý thuyết (Lectures): 34h
- Bài tập (Exercises): 6h
- Bài tập lớn (Projects): 4h
- Thực hành (Labs): 0h
- Tự học (Individual reading): 1h

**Địa chỉ Khoa/Bộ môn phụ trách học phần (Address of the Faculty/Department in charge of the course):**

- Address: Faculty of Information Technology 1 - Posts and Telecommunications Institute of Technology, Km10, Nguyen Trai Street, Ha Dong District, Hanoi.
- Phone number: (024) 33510432

## 2. Mục tiêu học phần (Objectives)

**Về kiến thức (Knowledge):**

The goal of this course is to provide learners with the fundamental concepts and methods of information retrieval, including:

- basic concepts and fundamental tasks in information retrieval
- applications in information retrieval
- methods for capturing, representing, storing, organizing, and retrieving unstructured or loosely structured information.

**Kỹ năng (Skills):**

The aim of this course is to equip learners with skills in:

- applying the learned knowledge to solve some information retrieval tasks
- determining the effectiveness of an information retrieval system using a known document corpus.

**Thái độ, Chuyên cần (Attitude):**

Students are required to attend the classes and complete exercises and assignments.

### **3. Tóm tắt nội dung học phần (Description)**

Information retrieval systems are systems that provide the ability to search for specific data or information within a collection. This course will emphasize technologies and fundamental concepts that underlie all information retrieval systems. The course consists of five chapters as follows. Chapter 1 presents the overview of information retrieval. Chapter 2 introduces concepts of text processing. Chapter 3 presents ranking with indexes with inverted index, index construction and compression, and query processing. Chapter 4 presents all about queries and interfaces. Chapter 5 describes different models using in information retrieval. And chapter 6 introduces evaluation methods which help to determine the effectiveness of an information retrieval system.

### **4. Nội dung chi tiết học phần (Outlines)**

#### **Chapter 1. Overview of Information Retrieval**

- 1.1. What is information retrieval?
- 1.2. Examples of information retrieval problems
- 1.3. Approaches to information retrieval
- 1.4. Architecture of a search engine
- 1.5. Performance of an information retrieval system

#### **Chapter 2. Processing Text**

- 2.1. Text statistics
  - 2.1.1. Vocabulary growth
  - 2.1.2. Estimating collection and result set sizes
- 2.2. Document Parsing
  - 2.2.1. Tokenizing
  - 2.2.2. Stopping
  - 2.2.3. Stemming
  - 2.2.4. Phrases and N-grams
  - 2.2.5. Document structure and markup
- 2.3. Link Analysis
  - 2.3.1. Anchor text
  - 2.3.2. PageRank
  - 2.3.3. Link quality
- 2.4. Information extraction

#### **Chapter 3. Ranking with Indexes**

- 3.1. Abstract model of ranking
- 3.2. Inverted indexes
  - 3.2.1. Documents
  - 3.2.2. Counts
  - 3.2.3. Positions
  - 3.2.4. Fields and extents
  - 3.2.5. Scores
  - 3.2.6. Ordering
- 3.3. Index construction
  - 3.3.1. Blocked sort-based indexing
  - 3.3.2. Single-pass in-memory indexing

- 3.3.3. Distributed indexing
- 3.3.4. Dynamic indexing
- 3.3.5. Other types of indexes
- 3.4. Index compression
  - 3.4.1. Statistical properties of terms
  - 3.4.2. Dictionary compression
  - 3.4.3. Postings file compression
- 3.5. Query processing
  - 3.5.1. Document-at-a-time evaluation
  - 3.5.2. Term-at-a-time evaluation
  - 3.5.3. Optimization techniques
  - 3.5.4. Structured queries
  - 3.5.5. Distributed evaluation
  - 3.5.6. Caching

## **Chapter 4. Queries and Interfaces**

- 4.1. Information needs and queries
- 4.2. Query transformation and refinement
  - 4.2.1. Stopping and stemming revisited
  - 4.2.2. Spell checking and suggestions
  - 4.2.3. Query expansion
  - 4.2.4. Relevance feedback and pseudo relevance feedback
  - 4.2.5. Global methods for query reformulation
  - 4.2.6. Context and personalization
- 4.3. Showing the results
  - 4.3.1. Result pages and snippets
  - 4.3.2. Advertising and search
  - 4.3.3. Clustering the results
  - 4.3.4. Translation

## **Chapter 5. Retrieval Models**

- 5.1. The vector space model
  - 5.1.1. Dot products
  - 5.1.2. Queries as vectors
  - 5.1.3. Computing vector scores
- 5.2. Probabilistic information retrieval
  - 5.2.1. The probability ranking principle
  - 5.2.2. The binary independence model
- 5.3. Language models for information retrieval
  - 5.3.1. Language models
  - 5.3.2. The query likelihood model
  - 5.3.3. Language modeling versus other approaches
  - 5.3.4. Extended language modeling approaches

## **Chapter 6. Evaluation**

- 6.1. Information retrieval system evaluation

- 6.2. Standard test collections
- 6.3. Evaluation of unranked retrieval sets
- 6.4. Evaluation of ranked retrieval sets
- 6.5. Assessing relevance
- 6.6. System quality and user utility

## 5. Học liệu (Textbooks)

### 5.1. Học liệu bắt buộc (Required Textbooks)

- [1]. Manning, Christopher, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2009.

### 5.2. Học liệu tham khảo (Optional Textbooks)

- [2]. Stefan Büttcher, Charles Clarke, and Gordon Cormack, *Information retrieval: Implementing and evaluating search engines*. Mit Press, 2016.
- [3]. Bruce Croft, Donald Metzler, and Trevor Strohman, *Search Engines: Information Retrieval in Practice*, Addison-Wesley, 2009.

## 6. Phương pháp, hình thức kiểm tra – đánh giá kết quả học tập học phần (Grading Policy)

Grading method	Percentage	Group/Individual
- Attendance	10%	Individual
- Exercises	10%	Individual
- Mid-term projects/exam	20%	Group or individual
- Final examination	60%	Individual

**Trưởng Bộ môn**  
**(Head of Department)**

**Giảng viên biên soạn**  
**(Lecturer)**

**Ngô Xuân Bách**

**Nguyễn Ngọc Diệp**