TEXT MINING AND ANALYTICS

1.Thông tin về học phần (General Information)

Tên học phần (Course name): Text Mining and Analytics

Mã học phần (Course code): INT_E14127

Số tín chỉ (Number of credits): 3

Loại học phần (Course type): Elective

Học phần tiên quyết (Prerequisites):

- Probability and Statistics (BAS1226)

Học phần trước (Previous courses):

Học phần song hành (Parallel courses):

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

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

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

- Lý thuyết (Lectures): 32h
- Bài tập (Exercises): 8h
- 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 aim of this course is to provide learners with important knowledge about natural language text and methods for mining and analyzing text data, including:

- natural language basics (languages, words and word classes, syntactics, semantics, corpora)
- text mining problems (text processing, text classification, text summarization, text clustering, semantic analysis)
- methods for mining and analyzing text data (Naïve Bayes, SVM, LSI, LSA, LDA, matrix factorization, k-means, TextRank, and so on)

Kỹ năng (Skills):

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

- applying the learned knowledge to mine and analyze natural language text
- evaluation of text mining models.

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)

This course introduces learners to basic knowledge about natural language text (language, words and word classes, syntactics, semantics, corpora), text mining problems (text processing, text classification, text summarization, text clustering, semantic analysis), and methods for mining and analyzing text data (Naïve Bayes, SVM, LSI, LSA, LDA, matrix factorization, k-means, TextRank, and so on). The course also teaches students how to evaluate text mining models.

4. Nội dung chi tiết học phần (Outlines)

Chapter 1: Natural Language Basics

1.1. Natural language
1.2. Language syntax and structure
    1.2.1. Words
    1.2.2. Phrases
    1.2.3. Clauses
    1.2.4. Grammars
1.3. Language semantics
    1.3.1. Lexical semantic relations
    1.3.2. Representation of semantics
1.4. Text corpora
    1.4.1. Corpora annotation and utilities
    1.4.2. Popular corpora
    1.4.3. Accessing text corpora
1.5. Natural language processing
    1.5.1. Fundamental tasks
    1.5.2. Applications

Chapter 2: Processing and Understanding Text

2.1. Text tokenization
    2.1.1. Sentence tokenization
    2.1.2. Word tokenization
2.2. Text normalization
    2.2.1. Cleaning text
    2.2.2. Tokenizing text
    2.2.3. Removing special characters and stopwords
    2.2.4. Correcting words
    2.2.5. Stemming
2.3. Understanding text syntax and structure
    2.3.1. Parts-of-speech tagging
    2.3.2. Shallow parsing
    2.3.3. Dependency parsing
    2.3.4. Constituency parsing

Chapter 3: Text Classification

3.1. Introduction to text classification
3.2. Feature extraction
    3.2.1. Bag of words
    3.2.2. TF-IDF
    3.2.3. Word vectors
3.3. Classification algorithms
    3.3.1. Naïve Bayes

3.3.2. Support vector machines
3.4. Evaluating classification models
3.5. Projects

Chapter 4: Text Summarization
4.1. Introduction to text summarization
4.2. Keyphrase extraction
4.2.1. Collocations
4.2.2. Weighted tag–based phrase extraction
4.3. Topic modeling
4.3.1. Latent semantic indexing
4.3.2. Latent Dirichlet allocation
4.3.3. Non-negative matrix factorization
4.3.4. Extracting topics from product reviews
4.4. Automated document summarization
4.4.1. Latent semantic analysis
4.4.2. TextRank
4.4.3. Multi document summarization

Chapter 5: Text Similarity and Clustering
5.1. Introduction
5.2. Analyzing term similarity
5.2.1. Hamming distance
5.2.2. Manhattan distance
5.2.3. Euclidean distance
5.2.4. Levenshtein edit distance
5.2.5. Cosine distance and similarity
5.3. Analyzing document similarity
5.3.1. Cosine similarity
5.3.2. Hellinger-Bhattacharya distance
5.3.3. Okapi BM25 Ranking
5.4. Document clustering
5.4.1. K-means clustering
5.4.2. Affinity propagation
5.4.3. Hierarchical clustering

Chapter 6: Semantic and Sentiment Analysis
6.1. Exploring WordNet
6.1.1. Understanding synsets
6.1.2. Analyzing lexical semantic relations
6.1.3. Word sense disambiguation
6.1.4. Named entity recognition
6.2. Analyzing semantic representations
6.3. Sentiment analysis
6.3.1. Preparing datasets
6.3.2. Feature extraction
6.3.3. Supervised machine learning techniques
6.3.4. Unsupervised lexicon-based techniques
6.3.5. Model evaluation
6.4. Projects

5. Học liệu (Textbooks)

5.1. Học liệu bắt buộc (Required Textbooks)

[1]. Dipanjan Sarkar. *Text Analytics with Python: A Practical Real-World Approach to*

Gaining Actionable Insights from your Data. Apress, 2016.

5.2. Học liệu tham khảo (Optional Textbooks)

[2]. Dan Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, 2nd edition, 2009.

[3]. Yoav Goldberg. *Neural Network Methods in Natural Language Processing - Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, 2017.

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	**Ngô Xuân Bách**