VNUHCM-UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY

COURSE SYLLABUS
<CSC15105> – Text Mining & Applications

1. GENERAL INFORMATION

Course name:	Text Mining & Applications
Course name (in Vietnamese):	Khai thác ngữ liệu văn bản & ứng dụng
Course ID:	CSC15105
Knowledge block:	Elective - Computer Science
Number of credits:	4
Credit hours for theory:	45
Credit hours for practice:	30
Credit hours for self-study:	90
Prerequisite:	Introduction to Natural Language Processing
Prior-course:	None
Instructors:	Le Thanh Tung

2. COURSE DESCRIPTION

The course is designed to provide students with an overview of the field of corpus mining, introducing concepts and methods to discover linguistic knowledge from large textual data. The course also helps to build the foundation knowledge for the Knowledge Technology major (Natural Language Processing field) in order to prepare for more specialized subjects in the following years. The course also helps students experience the application of data mining techniques to solve problems in life: business, education, health, ...

3. COURSE GOALS

At the end of the course, students are able to

ID	Description	Program LOs
G1	Work individually and in group collaboration to research, apply textual corpus mining, and present results	2.2, 2.3, 2.4.3

G2	Know the basic concepts of text corpus mining	1.7.2, 2.1.2, 2.1.3
G3	Understand and distinguish important terminologies in text corpus mining	1.7.2, 2.1.2, 2.1.3
G4	Understand and apply the text mining methods to apply to actual textual data	1.7.2, 2.1.2, 2.1.3, 2.1.5
G5	Recognize the role and importance of text corpus mining through applications in daily life	2.1.2, 2.1.3, 2.1.5

4. COURSE OUTCOMES

CO	Description	I/T/U
G1.1	Establish, organize, operate and manage groups	I,T
G1.2	Participate in group discussions on subject topics	U
G1.3	Research, collaborate, write reports and present projects	U
G2	Know the basic concepts of text corpus mining	I,T
G3	Understand and distinguish important terminologies in text corpus mining	T,U
G4.1	Understanding text corpus mining methods and their applications	T,U
G4.2	Understand the role of text corpus mining methods in practical applications: document classification, opinion detection, information retrieval, ...	U
G5	Understand the role of text corpus mining applications	U

5. TEACHING PLAN

THEORY

ID	Topic	Course outcomes	Teaching/Learning Activities (samples)	Assessments
1	Introduction to Text Mining	G1.2, G2, G3	**- At home:** ● Read Chapter 1 in Reference 1 **- In class:** ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A ● QZ1
2	- Text predictive analytics - Text representation	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	**- At home:** ● View Slides **- In class:** ● Group discussion ● Q&A ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A ● Project1
3	- Text classification - Representation-based learning	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	**- At home:** ● View Slides **- In class:** ● Group discussion ● Q&A ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A
4	- Model evaluation and experiment	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	**- At home:** ● View Slides **- In class:** ● Group discussion ● Q&A ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A

5	- Exploratory analysis: Clustering - Semantic analysis	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	**- At home:** ● View Slides **- In class:** ● Group discussion ● Q&A ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A
6	- Opinion mining - Text-based forecasting	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	**- At home:** ● View Slides **- In class:** ● Group discussion ● Q&A ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A
7	- Information mining - Bootstrapping approaches	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	**- At home:** ● View Slides **- In class:** ● Group discussion ● Q&A ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A
8	- Corpus pre-processing - Query on vector spaces	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	**- At home:** ● View Slides **- In class:** ● Group discussion ● Q&A ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A
9	- Semantic distribution - Text classification	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	**- At home:** ● View Slides **- In class:** ● Group discussion ● Q&A ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A

10	- Tagging and a corpus - Text mining applications	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	**- At home:** ● View Slides **- In class:** ● Group discussion ● Q&A ● Lecturing	**- At home:** ● Discussion **- In class:** ● Q&A
11	Project report	G1.1, G1.2, G1.3, G2, G3, G4.1, G4.2, G4.3, G5	Students: The group presents the results of the project, the applications of the project, discussions and Q&A.	Project2

LABORATORY

ID	Topic	Course outcomes	Teaching/Learning Activities (samples)	Assessments
1	Concepts, Terminologies	G2, G3	Explain and Group discussion	LW1
2-10	Methods of textual data mining and applications	G4.1, G4.2, G4.3, G5	Explain and Group discussion	LW2-10

6. ASSESSMENTS

ID	Topic	Description	Course outcomes	Ratio (%)
A	**Projects**			**100%**

1	Project1	Seminar: Group presentation + report on Information Retrieval (registered by the group on a list)	G1.1, G1.2, G1.3, G2, G4.1, G4.2, G4.3, G5	40%
2	Project2	Seminar: Group presentation + report on Open-ended Question Answering (registered by the group on a list)	G1.1, G1.3, G4.2, G4.3, G5	60%

7. RESOURCES

- **Textbooks**

  - Text Mining - Predictive Methods for Analyzing Unstructured Information, Weiss, S.M. et al., 2005, Springer New York. doi: 10.1007/978-0-387-34555-0
  - A comparison of statistical significance tests for information retrieval evaluation, Smucker, M. D., Allan, J. and Carterette, B.,2007, in International Conference on Information and Knowledge Management, Proceedings, pp. 623–632. doi: 10.1145/1321440.1321528.
  - Updating users about time critical events, Guo, Q., Diaz, F. and Yom-Tov, E., 2013, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 483–494. doi: 10.1007/978-3-642-36973-5_41.
  - A bootstrapping approach for identifying stakeholders in public-comment corpora, Arguello, J. and Callan, J., 2007, in Proceedings of the 8th annual international conference on Digital government research, Digital Government Society of North America, pp. 92-101.
  - Reading the markets: Forecasting public opinion of political candidates by news analysis, Lerman, K. et al., 2008, in Coling 2008 - 22nd International Conference on Computational Linguistics, Proceedings of the Conference. Manchester, pp. 473–480. doi: 10.5555/1599081.1599141

- **Others**
  - Link Youtube
  - Coursera
  - Online tutorials

8. GENERAL REGULATIONS & POLICIES

- All students are responsible for reading and following strictly the regulations and policies of the school and university.
- Students who are absent for more than 3 theory sessions are not allowed to take the exams.
- For any kind of cheating and plagiarism, students will be graded 0 for the course. The incident is then submitted to the school and university for further review.
- Students are encouraged to form study groups to discuss the topics. However, individual work must be done and submitted on your own.