COURSE SYLLABUS  
**CSC15006 – Introduction to Natural Language Processing**

# GENERAL INFORMATION

|  |  |
| --- | --- |
| Course name: | Introduction to Natural Language Processing |
| Course name (in Vietnamese): | Nhập môn xử lý ngôn ngữ tự nhiên |
| Course ID: | CSC15006 |
| Knowledge block: | Fundamental and specialized knowledge |
| Number of credits: | 4 |
| Credit hours for theory: | 45 |
| Credit hours for practice: | 30 |
| Credit hours for self-study: | 90 |
| Prerequisite: |  |
| Prior-course: |  |
| Instructors: | Nguyen Hong Buu Long |

# COURSE DESCRIPTION

This is an in-depth course about the problems of reducing the ambiguities of natural language problems. This course will focus on the corpus-based approach to address the problem of reducing the ambiguities of natural language in terms of morphology (such as Vietnameses word boundary), text-type, syntax and the semantics of the words. This module will also introduce some advanced applications of natural language processing such as automatic translation, text summary, ...

# COURSE GOALS

At the end of the course, students are able to

|  |  |  |
| --- | --- | --- |
| **ID** | **Description** | **Program LOs** |
| G1 | Understanding and applying knowledge about the knowledge of natural language processing according to the corpus-based approach in language problems such as: part-of-speech tagging and sentence parsing by statistical method; word sense disambiguation; information retrieval; text classification; statistical machine translation. | LO1, LO2 |
| G2 | Understanding the aspects of natural language such as morphology, grammar, and semantics | LO6, LO7 |
| G3 | Applying morphological, grammatical and semantic analysis techniques for text, especially Vietnamese text | LO6, LO7 |
| G4 | Assessing the performance of natural language processing systems | LO3, LO4, LO5, LO7 |
| G5 | Enhance ability to read English documents, work in groups and present reports | LO12 |

# COURSE OUTCOMES

|  |  |  |
| --- | --- | --- |
| **CO** | **Description** | **I/T/U** |
| G1.1 | Know and understand specialized English terminology of the subject. | I |
| G1.2 | Ability to apply English language processing methods for Vietnamese. | I |
| G1.3 | Classifying basic concepts and terms in natural language | T |
| G1.4 | List the similarities and differences between English and Vietnamese in terms of automatic language processing | T |
| G2.1 | Understand some problems in English and Vietnamese natural language processing. | TU |
| G2.2 | Applying knowledge of Natural language processing to deploy application problems | IU |
| G3 | Using a number of programming languages, organizing appropriate data structures, understanding and implementing available algorithms. | I |
| G4.1 | Ability to self-learn other methods (in addition to the learned content) for Natural Language Processing problems. | I |
| G4.2 | Applying teamwork skills | U |
| G4.3 | Applying skills to search, read and understand professional documents, and write reports | U |
| G5.1 | Read and understand English materials related to lectures. | U |
| G5.2 | Written and oral skills, and presentation skills related to the subject's topics. | U |

# TEACHING PLAN

**THEORY**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Topic** | **Course outcomes** | **Teaching/Learning Activities (samples)** | **Assessments** |
| 1 | Introduction to natural language processing: the characteristics of natural language, English characteristics, Vietnamese characteristics | G1.1,G1.2,  G1.3,G1.4  G2.1 | Lecturing  Q&A, Group discussion |  |
| 2 | Linguistic-based approach: linguistic criteria, labeled linguistic material | G1.1,G1.3,  G2.1 | Lecturing  Q&A, Group discussion |  |
| 3 | Language model in natural language processing | G1.1,G1.2,  G1.3,G1.4,  G2.1 | Lecturing  Q&A, Group discussion |  |
| 4 | Morphological analysis techniques in language | G1.2,G1.3,  G1.4,G2.1,  G2.2 | Lecturing  Q&A, Group discussion |  |
| 5 | Analytical techniques of grammar studies in language | G1.2,G1.3,  G1.4,G2.1,  G2.2 | Lecturing  Q&A, Group discussion |  |
| 6 | Analytical techniques for semantics in language | G1.2,G1.3,  G1.4,G2.1,  G2.2 | Lecturing  Q&A, Group discussion |  |
| 7 | Corpus linguistics: collecting, constructing and labeling data | G1.2,G1.3,  G1.4,G2.1,  G2.2 | Lecturing  Q&A, Group discussion |  |
| 8 | Open source tools for natural language processing | G2.1,G2.2,  G3,G4.1 | Lecturing  Q&A, Group discussion |  |
| 9 | Components in a natural language processing system | G2.2, G4.1,G4.2,  G4.3,G5.1 | Lecturing  Q&A, Group discussion |  |
| 10 | Coursework report | G3, G4.1,  G4.2,G4.3,  G5.1,G5.2 | Demonstrations, Debates | Project |

**LABORATORY**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Topic** | **Course outcomes** | **Teaching/Learning Activities (samples)** | **Assessments** |
| 1 | Lab work: Regular Expression | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Group discussion | LW#1 |
| 2 | Lab work: Pre-processing | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Group discussion | LW#2 |
| 3 | Lab work: Correcting Spelling Errors | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Group discussion | LW#3 |
| 4 | Lab work: Minimum Edit Distance | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Group discussion | LW#4 |
| 5 | Lab work: Text Classification | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Group discussion | LW#5 |
| 6 | Laboratory mid term exam | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Implement projects | Laboratory mid term exam |
| 7 | Lab work: Extracting keyword | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Group discussion | LW#6 |
| 8 | Lab work: Vietnamese sentence and word segmentation | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Group discussion | LW#7 |
| 9 | Lab work: Earley parser | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Group discussion | LW#8 |
| 10 | Laboratory final exam | G2.1, G2.2, G3, G4.1, G4.2, G4.3, G5.1 | Implement projects | Laboratory final exam |

.

# ASSESSMENTS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Topic** | **Description** | **Course outcomes** | **Ratio (%)** |
| **A1** | **Assignments** |  |  | **16%** |
| A11 | Weekly labwork: LW1–LW8 | LW1: Regular Expression  LW2: Pre-processing  LW3: Correcting Spelling Errors  LW4: Minimum Edit Distance  LW5: Text Classification  LW6: Extracting keyword  LW7: Vietnamese sentence and word segmentation  LW8: Earley parser | G2.1, G2.2, G3,G4.1,  G4.2,G4.3,  G5.1 | 16% |
| **A2** | **Projects** |  |  | **25%** |
| A21 | Project |  | G3, G4.1,  G4.2,G4.3,  G5.1,G5.2 | 25% |
| **A3** | **Exams** |  |  | **59%** |
| A31 | Laboratory mid term exam | Basic language processing exercises | G2.1,G2.2,  G3,G4.1,  G4.2,G4.3,  G5.1 | 4% |
| A32 | Laboratory final exam | Advanced language processing exercises and applications | G2.1,G2.2,  G3,G4.1,  G4.2,G4.3,  G5.1 | 5% |
| A33 | Theory final exam | Oral exam: Understanding and solving problems based on topics: ambiguities, morphological, grammar, semantic, Vietnamese language | G1.1,G1.2,  G1.3,G1.4  G2.1,G4.1,  G4.3,G5.1,  G5.2 | 50% |

# RESOURCES

# Textbooks

* Daniel Jurafsky, James H. Martin. “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”. 2023 (3rd ed. Draft) (<https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>)
* Sebastian Raschka, Build a Large Language Model (From Scratch), September 2024 ISBN 9781633437166, 368 pages.
* Đinh Điền. “Xử Lý Ngôn Ngữ Tự Nhiên”. NXB ĐHQG, 2006

# Others

* Christopher D. Manning, Hinrich Schütze. “Foundation of Statistical Natural Language Processing”. MIT Press, 2001
* Vaclav Brezina. “Statistics for Corpus Linguistics”. Cambridge University Press, 2018
* Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem, and Piotr Fuglewicz. “Computational Linguistics: Applications”. Springer, 2013
* Koehn, P. (2009). Statistical Machine Translation. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511815829
* Quim Motger, Xavier Franch, Jordi Marco (2021). Conversational Agents in Software Engineering: Survey, Taxonomy and Challenges. https://arxiv.org/abs/2106.10901
* Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. 2023. Mapping Brains with Language Models: A Survey. In Findings of the Association for Computational Linguistics: ACL 2023, pages 9748–9762, Toronto, Canada. Association for Computational Linguistics.
* Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 447–459, Suzhou, China. Association for Computational Linguistics.
* [*http://nltk.sourceforge.net/*](http://nltk.sourceforge.net/)

# 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 on the topics. However, individual work must be done and submitted on your own.