**

**CAPSTONE PROJECT REGISTER**

Class: Duration time: from 5/9/2022 to 12/11/2022

(\*) Profession: <Artificial Intelligence> Specialty: <ES> <IS> <JS>

(\*) Kinds of person make registers: Lecturer: Students

x

1. Register information for supervisor (if have)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Full name** | **Phone** | **E-Mail** | **Title** |
| Supervisor 1 |  |  |  |  |
| Supervisor 2 |  |  |  |  |

2. Register information for students (if have)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Full name** | **Student code** | **Phone** | **E-mail** | **Role in Group** |
| Student 1 |  |  |  |  |  |
| Student 2 |  |  |  |  |  |
| Student 3 |  |  |  |  |  |
| Student 4 |  |  |  |  |  |

3. Register content of Capstone Project

(\*) 3.1. Capstone Project name:

English: GraphSum: Long Document Summarization with Graph approaches on Sentence-level

Vietnamese: GraphSum: tóm tắt tài liệu dài sử dụng các phương pháp về đồ thị

Abbreviation: NLP-SA

- **Description**: Long document summarization is the process of generating or extracting concise and informative summaries from lengthy documents. It involves extracting essential information, key ideas, and main points from the original document while preserving its overall meaning and coherence. Transformer-based models require length limitation for the input, scientific papers’ length, for example, can range from 2000 to 7000 words is a challenge for models and expensive computation because of the quadratic computation complexity of attention with respect to the input tokens length of Transformer-based models. Moreover, long documents typically cover diverse topics and have richer structural information than short news, which is difficult for sequential models to capture. Therefore, the solution is that long documents’ perspectives such as topic, keyword, even semantic aspect of each section should be better captured by using graph in order to create connections between key components in a long document which is difficult for sequential models to capture.

- **Objectives**: After working with this project, students will obtain knowledge and skills related to Artificial Intelligence, Machine Learning, NLP, and the implementation and deployment of AI in productions (AI-based softwares). Recently, these fields attract concerns from both academic and industrial environments. Therefore, it would bring great opportunities for students' career.

(\*) 3.2. Main proposal content (including result and product)

1. Theory and practice (document): Students will be able to:

* Understand how to combine traditional algorithm, graph theory, to deep learning, specifically, Natural Language Processing
* Approach Graph Neural Network, do experiments on Graph Convolutional Neural Network or Graph Attention Neural Network
* Implement text mining algorithms such as topic extraction, keyword extraction, …
* Understand the different between word-level and sentence-level that NLP models process.
* Deep dive into Transformer architecture, popular NLP models’ architecture, to modify and customize for a specific task.

1. Program: students will combine both knowledge and skills to build a simple end-to-end application that supports the task long document summarization.

4. Other comment (propose all relative thing if have)

* Students are willing to learn new research topics, especially machine learning and deep learning for NLP.
* Students are willing to work hard such as report every week with new progress, read and present knowledge from the documents in the field.
* There is at least one student in the group who has solid background in software engineering for implementing the tool.
* The number of students for this project varies from 2 to 4. The project output may differ depending on this number.

|  |  |
| --- | --- |
| **Supervisor (If have)**  *(Sign and full name)* | Ho Chi Minh city, date 25/07/2022  **On behalf of Registers**  *(Sign and full name)* |