**

**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: VietLegalQA: Unsupervised Legal Question Answering for Vietnamese using Cloze Translation Approach

Vietnamese:

Abbreviation: NLP-SA

- **Description**: Question Answering is a task that a model generate or extract answers given a document related to the input question or query. Obtaining training data for Question Answering (QA) is often labor-intensive and resource-consuming, especially for specialized domains like legal texts and the number of data samples in most existing Vietnamese datasets are restricted. In this study, we address the challenge of Legal QA for the Vietnamese language by proposing an unsupervised approach called VietLegalQA. Our method leverages a Cloze Translation Approach to generate high-quality training data without the need for manual annotation. We first extract context paragraphs from legal documents and randomly select noun phrases or named entity mentions as answers. Subsequently, we transform these answers into "fill-in-the-blank" cloze questions and further translate them into natural questions. We explore various unsupervised techniques for cloze-to-natural question translation, including training an unsupervised Neural Machine Translation (NMT) model using non-aligned corpora of natural and cloze questions, as well as a rule-based approach.

Top of Form

- **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:

* Get opportunities to build a legal chatbot
* Developing an unsupervised Legal QA system requires students to tackle complex challenges related to data generation, translation, and model training. This process fosters problem-solving skills and encourages students to think critically to overcome obstacles and optimize the performance of their system. Specifically, Vietnamese Question Answering or in-domain QA task such as Legal
* By developing and implementing this approach, students can make a valuable contribution to the field of Natural Language Processing (NLP) and Question Answering (QA) for the Vietnamese language. They have the chance to explore innovative techniques and methodologies in unsupervised QA and apply them to the domain of legal texts.
* The implementation of VietLegalQA involves various NLP and machine learning techniques, such as data preprocessing, text representation, translation, and training QA models. Students can develop and refine their skills in these areas, gaining hands-on experience with state-of-the-art NLP frameworks and libraries.

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

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)* |