**Midterm Project**

**Introduction to Natural Language Processing**

Hướng dẫn nộp bài:

* Báo cáo bằng tiếng Anh (đối với lớp hệ CLC) và bằng tiếng Việt đối với lớp hệ Tiêu chuẩn;
* Báo cáo phương pháp (lý thuyết) nộp file pdf
* Code + data
* Slide tóm tắt lý thuyết + kết quả code

+ 3 file đó cho vào thư mục (đặt tên là sinh viên trong nhóm), nén lại và nộp

+ Tất cả sinh viên trong nhóm đều phải nộp bài

+ Tất cả sinh viên trong nhóm đều phải hiểu tất cả bài (tránh trường hợp mỗi sinh viên chỉ biết phần mình làm).

**Bài 1 (6 điểm):**

Xây dựng mô hình nhận dạng thực thể tên (Named Entity Recognition: NER) cho câu tiếng Anh hoặc tiếng Việt. Sử dụng các phương pháp sau:

* Phương pháp dùng mô hình phân loại
* Phương pháp dùng mô hình LSTM (Long Short Term Memory)

Yêu cầu:

* Hãy mô tả phương pháp (lý thuyết)
* Xây dựng chương trình (programming), đánh giá độ chính xác của mô hình trên tập test.
* Sử dụng dữ liệu huấn luyện tự collect

**Bài 2 (4 điểm):**

* Hãy trình bày các phương pháp Fasttext cho việc biểu biểu diễn từ bằng vector (Word2Vec và Word Embeddings).
* Ứng dụng biểu diễn đó cho bài toán 1, tức là input vào cho mỗi từ là 1 vector embedding theo phương pháp fasttext, cho phương pháp LSTM và phương pháp phân loại.

**Question 1 (6 points):**  
Build a Named Entity Recognition (NER) model for English or Vietnamese sentences using the following methods:

* Classification model approach
* Long Short Term Memory (LSTM) model approach

**Requirements:**

* Describe the methods (theory)
* Build the program, evaluate the model's accuracy on the test set
* Use the training data collected yourself

**Question 2 (4 points):**

* Present the Fasttext methods for word representation using vectors (Word2Vec and Word Embeddings)
* Apply this representation to problem 1, i.e., input for each word should be an embedding vector using the Fasttext method for both the LSTM and classification approaches.