Midterm Project

Introduction to NLP course

**Câu 1 (4đ):**

Hãy tìm hiểu các phương pháp Word2Vec, chọn 1 hoặc 2 phương pháp trong các phương pháp sau:

* CBOW và Skip-gram
* Fasttext
* Glove

1. Trình bày lý thuyết các phương pháp đó: ý nghĩa, mô hình, huấn luyện mô hình như thế nào?
2. Hãy ứng dụng để thực hiện bài toán sau:

Cho môt tập các câu gọi là S = {s1, s2, …, sn}

Nhập vào 1 câu X, hãy tìm câu si nào đó trong tập S mà tương tự với X nhiều nhất

**Câu 2 (6đ):**

Chúng ta đang thực hiện xây dựng mô hình phát hiện một text có thuộc các dạng sau không:

* Harrasment
* Hate speech
* Sextually explicit
* Dangerous content

1. Xây dựng dữ liệu, cụ thể là gán nhãn cho bài toán trên. Hãy sử dụng công cụ được cung cấp và làm theo hướng dẫn (trong file đi kèm)
2. Hãy sử dụng các mô hình phân loại text khác nhau để thực hiện bài toán này với mục tiêu độ chính xác càng cao càng tốt.