**TRƯỜNG ĐH SƯ PHẠM KỸ THUẬT TPHCM CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**KHOA CÔNG NGHỆ THÔNG TIN Độc lập – Tự do – Hạnh Phúc**

**---- ------**

# PHIẾU NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Họ và tên sinh viên:

1. Trần Thành Quang MSSV: 19133047
2. Cao Anh Văn MSSV: 19133067

Ngành: Kỹ thuật dữ liệu

Tên đề tài: Tìm hiểu kỹ thuật embedding cho xử lý ngôn ngữ tự nhiên

Giáo viên phản biện: TS. Nguyễn Thiên Bảo

## **NHẬN XÉT**

1. Về nội dung đề tài & khối lượng thực hiện:

**Lý thuyết:**

* Tìm hiểu về xử lý ngôn ngữ tự nhiên.
* Tìm hiểu về word embedding (TF-IDF,word2vec).
* Tìm hiểu về bài toán tóm tắt văn bản.
* Tìm hiểu thuật toán kmean.
* Tìm hiểu độ đo cho bài toán tóm tắt văn bản là rouge , bert score.

**Thực nghiệm:**

* Dữ liệu ( khoảng 2.385.532 câu tiếng việt với nhiều chủ đề khác nhau).
* Xây dựng model word2vec với thư viện Gensim Word2Vec của python.
* Dùng kmean để phân cụm và chọn số câu đầu ra.
* Đánh giá kết quả với độ đo bert score.

**Kết quả:**

* Hiểu được cơ sở lý thuyết về xử lý ngôn ngữ tự nhiên.
* Hiểu được kỹ thuật Word embedding.
* Áp dụng lý thuyết vào thực hành xử lý bài toán tóm tắt văn bản tiếng việt.
* Xây dựng một ứng dụng web để tóm tắt văn bản.
* Deploy ứng dụng web lên AWS.

**Links:**

* Github: <https://github.com/vanac17122001/NLP_VietNamese_Summaries_Text>
* Drive: <https://drive.google.com/drive/folders/1UFCRI4tiqUYGmWkNd4ohYJTFaQ-izzIp?usp=sharing>

1. Ưu điểm:

* Hiểu được tầm quan trọng của xử lý ngôn ngữ tự nhiên trong đời sống.
* Word embedding giúp ta xử lý ngôn ngữ tự nhiên tốt hơn.
* Đã xây dựng một ứng dụng web tóm tắt văn bản tiếng việt.

1. Khuyết điểm:

* Sử dụng thuật toán phân cụm kmean khá đơn giản.
* Chọn tham số K cho thuật toán kmean chưa tối ưu.
* Kết quả của bài toán chưa thực sự tốt.
* Chưa tìm hiểu sâu và rộng hơn với kỹ thuật embedding.
* Giao diện ứng dụng web chưa đẹp.

1. Đề nghị cho bảo vệ hay không?

1. Đánh giá loại:

Tp*. Hồ Chí Minh, ngày 04 tháng 01 năm 2023*

Giáo viên hướng dẫn

*(Ký & ghi rõ họ tên)*

**TS. Nguyễn Thiên Bảo**