

BÁO CÁO CUỐI KỲ

Môn học

CS2205.CH1501 -

PHƯƠNG PHÁP LUẬN NCKH

Giảng viên

PGS.TS. LÊ ĐÌNH DUY

Thời gian

03/2021 - 06/2021

----- *Trang này cố tình để trống* -----

HƯỚNG DẪN


Yêu cầu:

- *Bước 1: Chọn File/Make a copy để tạo ra một file theo template mẫu https://docs.google.com/document/d/1pu86lH6STGaVk2JH70n3jWx8qt9Eue_imVTQhg3s/. Đặt tên tập tin này là: CS2205.CH1501.RM.FinalReport.MSHV*
- *Bước 2: Điền các thông tin về đề cương đề tài vào file GDocs trên. Tối đa 6 trang.*
- *Bước 3: Copy toàn bộ nội dung đề cương đề tài và Paste vào cuối tập tin này (tránh ghi đè lên nội dung của HV khác).*
- *Bước 4: Nộp bài (Turn in) theo yêu cầu trên Classroom. Chọn Add or Create và chọn Link đến file Google ở trên. Lưu ý đặt quyền Anyone with the link - Viewer. Trong phần Private Comment, cung cấp thông tin của github repos, thông tin các thành viên của nhóm và các ghi chú khác nếu có. Lưu một phiên bản pdf của đề cương trên github repos*

Lưu ý:

- *Việc tuân thủ các hướng dẫn, các yêu cầu theo mẫu là bắt buộc và được đánh giá trong điểm tổng kết của đồ án môn học.*
- ***Deadline: 25/07/2021***

----- *Trang này cố tình để trống* -----

Họ và tên (IN HOA)	TÔ QUỐC HUY
Ảnh	
Số buổi vắng	0
Bonus	20
Tên đề tài (VN)	NGHIÊN CỨU THỰC NGHIỆM CÁC MÔ HÌNH BERT CHO TÁC VỤ TÓM TẮT ĐA VĂN BẢN TRÊN TIẾNG VIỆT
Tên đề tài (EN)	EMPIRICAL STUDY OF TEXT SUMMARIZATION ON MULTI-DOCUMENTS IN VIETNAMESE USING BERT
Giới thiệu	<p>Tóm tắt văn bản là việc cô đọng thông tin từ một hoặc nhiều đoạn văn bản thành một đoạn văn bản ngắn hơn. Tuy giảm thiểu số lượng câu chữ nhưng vẫn phải đảm bảo các yếu tố như thông tin và ý nghĩa về mặt nội dung. Các ứng dụng của tóm tắt văn bản tự động bao gồm: phân loại văn bản lớn, Question Answering, tóm tắt văn bản pháp lý, tóm tắt tin tức, tạo tiêu đề tự động.</p> <p>Việc tự động hóa công việc tóm tắt đang ngày càng phổ biến và độ hiệu quả cải thiện dần theo thời gian. Trên Tiếng Việt, các mô hình tóm tắt tự động đã</p>

	<p>được đề xuất như TSGVi [1], CFVi [2] và một số mô hình khác dựa trên thuật toán TextRank.</p> <p>Đề tài nghiên cứu độ hiệu quả của mô hình pre-train BERT, một mô hình hiện đại đã được thực nghiệm và chứng minh độ hiệu quả trên các tác vụ khác trong lĩnh vực xử lý ngôn ngữ tự nhiên. Tuy đã được thực nghiệm trên Tiếng Anh [3] nhưng chưa được thực nghiệm và đánh giá trên Tiếng Việt.</p>
Mục tiêu	<ul style="list-style-type: none"> • Nghiên cứu các mô hình BERT, các kỹ thuật có liên quan cho bài tóm tắt đa văn bản trên Tiếng Việt. • Chạy thực nghiệm để kiểm chứng độ chính xác và đánh giá hiệu suất của các mô hình. • Cải thiện độ chính xác của mô hình, chọn ra mô hình tốt nhất cho bài toán tóm tắt đa văn bản.
Nội dung và phương pháp thực hiện	<p>Nội dung nghiên cứu:</p> <ul style="list-style-type: none"> • Tạo tự động các đoạn tóm tắt sử dụng các mô hình BERT đa ngôn ngữ và đơn ngôn ngữ kết hợp với thuật toán K-Means clustering trên bộ dữ liệu VietnameseMDS. • Đánh giá và so sánh độ hiệu quả trên độ đo ROUGE giữa các mô hình BERT và với các mô hình đã được đề xuất trước đó. • Tối ưu mô hình dựa trên các kết quả phân tích để cho kết quả tốt nhất. <p>Phương pháp thực hiện:</p> <ul style="list-style-type: none"> - Nghiên cứu phương pháp thực nghiệm: <ul style="list-style-type: none"> • Thực nghiệm trên các mô hình pre-trained BERT đa ngôn ngữ và đơn ngôn ngữ, điều chỉnh tham số để tìm ra mô hình phù hợp với bài toán đặt ra, như:

	<ul style="list-style-type: none"> ○ mBERT ○ XML-Roberta ○ DistilBERT ○ PhoBERT ○ ViBERT4News ● Đồng thời áp dụng thuật toán như K-Means Clustering để trích xuất các câu có liên quan với chủ đề. - Các phương pháp được thực nghiệm trên bộ dữ liệu VietnameseMDS, bao gồm 200 cụm văn bản chuyên dùng cho tóm tắt đa văn bản. - Tối ưu các giá trị tham số của mô hình dựa trên kết quả phân tích từ kết quả độ đo đánh giá trên từng mô hình.
Kết quả dự kiến	<ul style="list-style-type: none"> - Đề xuất được mô hình pre-train BERT tốt nhất cho tác vụ tóm tắt đa văn bản trên Tiếng Việt. - Web demo với đầu vào là các văn bản Tiếng Việt dưới định dạng file chữ (text) và cho ra kết quả là một đoạn văn tóm tắt
Tài liệu tham khảo	<p>[1] Tu-Anh Nguyen-Hoang, Hoang Khai Nguyen, Quang Vinh Tran: An Efficient Vietnamese Text Summarization Approach Based on Graph Model. RIVF 2010: 1-6</p> <p>[2] Van-Giau Ung, An-Vinh Luong, Nhi-Thao Tran, Minh-Quoc Nghiem: Combination of Features for Vietnamese News Multi-document Summarization. KSE 2015: 186-191</p> <p>[3] Yang Liu, Mirella Lapata: Text Summarization with Pretrained Encoders. EMNLP/IJCNLP (1) 2019: 3728-3738</p>

- *Các bài nộp bắt đầu từ trang 7.*

- *Lưu ý Paste vào cuối file để tránh ảnh hưởng đến các bài nộp trước đó.*

