# **THÔNG TIN CHUNG CỦA NHÓM**

* Link YouTube video của báo cáo (tối đa 5 phút):   
  *(ví dụ: https://www.youtube.com/watch?v=AWq7uw-36Ng)*
* Link slides (dạng .pdf đặt trên Github của nhóm):   
  *https://github.com/kiendoo4/CS519.O11/blob/main/Ki%C3%AAn%20Nguy%E1%BB%85n%20Trung%20-%20xCS519.DeCuong.FinalReport.Template.Slide.pptx*
* *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
* *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

| * Họ và Tên: Nguyễn Trung Kiên * MSSV: 21521024 | * Lớp: CS519.O11 * Tự đánh giá (điểm tổng kết môn): 8.75/10 * Số buổi vắng: 2 * Số câu hỏi QT cá nhân: 12 * Link Github: https://github.com/kiendoo4/CS519.O11 * Toàn bộ nội dung đề cương, video, slide và poster được lên ý tưởng, biên soạn và thực hiện bởi sinh viên Nguyễn Trung Kiên. |
| --- | --- |

# **ĐỀ CƯƠNG NGHIÊN CỨU**

| **TÊN ĐỀ TÀI (IN HOA)**  TRUY VẤN SÁCH TIẾNG VIỆT DỰA VÀO NGỮ NGHĨA MÔ TẢ CỦA NGƯỜI DÙNG ÁP DỤNG MÔ HÌNH VIETNAMESE-SBERT |
| --- |
| **TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**  SEMANTIC-BASED VIETNAMESE BOOK RETRIEVAL USING VIETNAMESE-SBERT |
| **TÓM TẮT** *(Tối đa 400 từ)*  Hiện nay, các công cụ tìm kiếm văn bản sách như Google Books hoạt động bằng cách sử dụng phương pháp so khớp từ khóa. Phương pháp này có ưu điểm là giúp độc giả tìm kiếm được cuốn sách bằng nội dung câu từ còn nhớ được. Tuy nhiên, phương pháp này cũng có một hạn chế là không thể tìm kiếm được cuốn sách nếu truy vấn không dựa trên câu từ còn nhớ trong sách, bởi lẽ các công cụ tìm kiếm này không được xây dựng để hiểu được nội dung và ý nghĩa mô tả của truy vấn. Do đó, nếu người đọc không nhớ được chính xác những gì trong sách, họ sẽ không thể tìm kiếm được cuốn sách mong muốn. Để giải quyết vấn đề hạn chế của phương pháp so khớp từ khóa, chúng tôi đề xuất một phương pháp tìm kiếm sách tiếng Việt dựa vào ngữ nghĩa mô tả của truy vấn được lấy cảm hứng từ bài toán phát hiện sự đồng nghĩa giữa 2 câu khác nhau. Cụ thể, phương pháp của chúng tôi kết hợp việc lọc bớt những cuốn sách chứa nội dung không hoặc ít liên quan đến truy vấn bằng BM25 [1], sau đó sử dụng mô hình Vietnamese-SBERT [2] để tìm kiếm ngữ nghĩa cho từng câu trong sách và cuối cùng là tính toán mức độ liên quan giữa truy vấn và nội dung trong sách bằng một cách thức được chúng tôi đề xuất. Ngoài ra, chúng tôi cũng thực hiện việc xây dựng một tập ngữ liệu sách tiếng Việt để phục vụ không chỉ riêng cho đề tài này mà có thể dùng vào các nghiên cứu về truy vấn văn bản tiếng Việt trong tương lai. |
| **GIỚI THIỆU** *(Tối đa 1 trang A4)*  Việc quên sách là một trong những vấn đề thường gặp trong cộng đồng những người yêu đọc sách. Nhiều độc giả từng đọc những cuốn sách hay và giá trị nhưng sau một thời gian họ lại quên mất cuốn sách đó, đến mức chỉ còn nhớ được một vài chi tiết hoặc nội dung chính một cách mơ hồ.  Để giải quyết vấn đề, một số các công cụ tìm kiếm sách đã được tạo ra và phát triển. Tuy nhiên, các công cụ này vẫn đang hoạt động dựa trên việc so khớp từ khóa là chính, do đó ngữ nghĩa của truy vấn có thể bị bỏ qua. Điều này khiến cho việc nếu mô tả của độc giả không căn cứ vào nội dung của sách, hay nếu ta muốn tìm kiếm sách dựa trên mô tả sở thích thì cũng không thể sử dụng các công cụ này vì chúng vốn không được xây dựng để hiểu được nội dung và ý nghĩa của mô tả.  Nhận thấy sự hạn chế của các hệ thống tìm kiếm văn bản nói chung và sách nói riêng khi áp dụng so khớp từ khóa, chúng tôi đề xuất một giải pháp tìm kiếm sách dựa vào mô tả của người dùng lấy cảm hứng từ bài toán phát hiện sự đồng nghĩa giữa 2 câu khác nhau (Semantic textual similarity, hay STS).  Một số các bài báo khoa học nhằm giải quyết bài toán STS đã được công bố như Sentence-BERT [3] đã có được những kết quả state-of-the-art trên nhiều bộ dataset khác nhau và đã trở thành tiền đề cho rất nhiều những công trình nghiên cứu khoa học sau này, trong đó có Vietnamese-SBERT [2] - một mô hình ngôn ngữ tiền huấn luyện giúp giải quyết bài toán STS cho ngôn ngữ Việt.  Áp dụng sự hiệu quả của mô hình Vietnamese-SBERT [2], giải pháp mà chúng tôi đề xuất là một giải pháp tìm kiếm ngữ nghĩa trên toàn bộ nội dung cuốn sách nhằm tìm ra các cuốn sách liên quan nhất. Cụ thể:  *Input:*   * Một câu truy vấn mô tả cuốn sách muốn tìm kiếm. * Ngữ liệu: chứa thông tin của rất nhiều những cuốn sách, mỗi sách bao gồm tiêu đề và toàn bộ nội dung của nó.   *Output:*   * Một danh sách các sách được sắp xếp từ trên xuống dưới tương ứng với mức độ liên quan của cuốn sách đó với truy vấn đã nhập. |
| **MỤC TIÊU**   * Tìm hiểu, nghiên cứu và áp dụng BM25 [1] và mô hình ngôn ngữ Vietnamese-SBERT [2], để tìm kiếm các cuốn sách chứa nội dung liên quan đến mô tả tìm kiếm. * Sử dụng độ đo nDCG@k [4] để đánh giá và so sánh hiệu quả của phương pháp với một số các phương pháp truy vấn văn bản khác. * Xây dựng tập dữ liệu cho bài toán tìm kiếm văn bản tiếng Việt. |
| **NỘI DUNG VÀ PHƯƠNG PHÁP**  **Nội dung:**   * Tạo một bộ ngữ liệu lấy tên là UIT-VNBook chứa đựng thông tin của các cuốn sách tiếng Việt. Ngữ liệu này sẽ là nguồn để thực hiện việc truy xuất dựa trên mô tả và do đó, các kết quả trả về cho truy vấn cũng sẽ là những cuốn sách nằm trong ngữ liệu. * Sử dụng hàm truy vấn BM25 [1] để lọc bớt những cuốn sách chứa nội dung gần như không liên quan đến mô tả. Công việc này cũng giúp giảm bớt số lượng sách để xử lý khi áp dụng mô hình Vietnamese-SBERT [2] vì nếu số lượng sách lớn, mỗi sách còn có số câu lớn thì chi phí tính toán và thời gian xử lý sẽ vô cùng cao. * Nghiên cứu và áp dụng mô hình Vietnamese-SBERT [2] để thực hiện việc tìm kiếm ngữ nghĩa trên từng câu của nội dung cuốn sách với truy vấn nhập vào. * Đề xuất một độ đo duy nhất đặt tên là mDCpQ dùng để xác định mức độ liên quan giữa nội dung của sách và truy vấn. * So sánh và đánh giá hiệu quả của mô hình được đề xuất so với một số các phương pháp đã được công bố khác như CEDR [5], DRMM [6]. * Xây dựng một ứng dụng web minh họa đề tài.   **Phương pháp:**   * Xây dựng bộ ngữ liệu đặt tên là UIT-VNBook chứa đựng các thông tin cơ bản của một cuốn sách tiếng Việt bao gồm tựa đề sách, tên tác giả, ngày xuất bản, nhà xuất bản và thể loại bằng cách dùng thư viện BeautifulSoup và sử dụng thư viện PyPDF2 để lấy được nội dung bên trong file pdf của cuốn sách trên các trang web đọc sách điện tử. * Sử dụng thuật toán BM25 [1] để loại ra những cuốn sách chứa nội dung gần như không liên quan đến truy vấn. Từ điểm BM25 của từng cuốn sách ứng với truy vấn được nhập, chúng tôi chọn ra những cuốn sách có điểm cao nhất cho mục tiếp theo. * Nội dung của mỗi cuốn sách không bị lọc ra sẽ được tách thành một tập hợp các câu, trong đó mỗi câu sẽ được đưa vào mô hình Vietnamese-SBERT [2] để chuyển về thành một feature vector có kích thước 768x1. Thực hiện chuyển hết toàn bộ các câu nằm trong nội dung của các cuốn sách đã được lọc và cả câu truy vấn được nhập vào về dạng feature vector, chúng tôi sử dụng giá trị cosine để tính toán mức độ tương đồng giữa từng câu trong tập hợp các câu với câu truy vấn. Giá trị cosine của hai feature vector sẽ nằm trong đoạn [-1;1], cosine càng gần 1 chứng tỏ hai câu càng đồng nghĩa với nhau.và ngược lại. * Sắp xếp tất cả các câu được tách thành một danh sách L dựa vào giá trị cosine đã tính toán từ lớn đến bé. Chúng tôi đề xuất độ đo MDCpQ để xác định mức độ liên quan giữa nội dung của sách và truy vấn bằng cách lấy tổng giá trị của thương giữa cosine từng câu trong sách với logarit cơ số 10 của vị trí câu đó trong danh sách L. Sắp xếp giá trị của từng cuốn sách theo thứ tự từ cao xuống thấp và trả về output là danh sách thông tin các sách theo thứ tự đó. * So sánh và đánh giá hiệu quả của mô hình được đề xuất so với một số các phương pháp đã được công bố khác như CEDR [5], DRMM [6] bằng độ đo nDCG@k [4] trên một bộ dữ liệu văn bản khác xây dựng bằng cách thức tương tự UIT-VNBOOK. Bộ dữ liệu sẽ được dựa trên một truy vấn nào đó và các cuốn sách nằm trong bộ dữ liệu này cũng sẽ được sắp xếp từ trên xuống dưới tương ứng với mức độ liên quan nhiều hay ít với truy vấn đó. Để đảm bảo tính khách quan, chúng tôi sẽ xây dựng một số tiêu chí chi tiết để sắp xếp sách. * Tìm hiểu và áp dụng Flask để xây dựng một ứng dụng web sử dụng giúp mọi người tìm kiếm các cuốn sách đã bị lãng quên hoặc theo mong muốn dựa trên mô tả của người dùng đó. |
| **KẾT QUẢ MONG ĐỢI**   * Kết quả độ đo nDCG@k [4] của phương pháp được chúng tôi đề xuất vượt qua các phương pháp đã được công bố trước đó như CEDR [5], DRMM [6] cho bài toán Ad-hoc Document Retrieval. * Giải pháp được thiết kế có thể được ứng dụng không chỉ cho việc tìm kiếm mà có thể dùng để tìm kiếm sách dựa trên mô tả sở thích của người đọc. Đóng góp vào sự phát triển của lĩnh vực Truy vấn thông tin và Xử lý ngôn ngữ tự nhiên, làm tiền đề cho các công trình nghiên cứu khoa học sau này. |
| **TÀI LIỆU THAM KHẢO** *(Định dạng DBLP)*  [1]. Andrew Trotman, Antti Puurula, Blake Burgess:  Improvements to BM25 and Language Models Examined. ADCS 2014: 58  [2]. Hoang-Quoc Nguyen-Son, Nam-Phong Tran, Ngoc-Vien Pham, Minh-Triet Tran, Isao Echizen: Vietnamese Paraphrase Identification Using Matching Duplicate Phrases and Similar Words. FDSE 2018: 172-182  [3]. Nils Reimers, Iryna Gurevych: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP/IJCNLP (1) 2019: 3980-3990  [4]. Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu: A Theoretical Analysis of NDCG Type Ranking Measures. COLT 2013: 25-54  [5]. Sean MacAvaney, Andrew Yates, Arman Cohan, Nazli Goharian:  CEDR: Contextualized Embeddings for Document Ranking. CoRR abs/1904.07094 (2019)  [6]. Jiafeng Guo, Yixing Fan, Qingyao Ai, W. Bruce Croft:  A Deep Relevance Matching Model for Ad-hoc Retrieval. CoRR abs/1711.08611 (2017) |