

THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):
- Link slides (dạng .pdf đặt trên Github):

- | | |
|--|--|
| <ul style="list-style-type: none">• Họ và Tên: Quách Bảo Hưng• MSSV: 18520809 | <ul style="list-style-type: none">• Tự đánh giá (điểm tổng kết môn): 8.5/10• Số buổi vắng: 1• Link Github: |
|--|--|



ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT TRIỂN MỘT HỆ THỐNG TÌM KIẾM THÔNG MINH SỬ DỤNG CÁC PHƯƠNG PHÁP MACHINE LEARNING VÀ MÔ HÌNH VECTOR HÓA VĂN BẢN ĐỂ BIỂU DIỄN VÀ TÌM KIẾM VĂN BẢN HIỆU QUẢ

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

MACHINE LEARNING TECHNIQUES AND TEXT VECTORIZATION MODELS FOR EFFECTIVE TEXT REPRESENTATION AND RETRIEVAL TO CREATE AN INTELLIGENT SEARCH SYSTEM.

TÓM TẮT (*Tối đa 400 từ*)

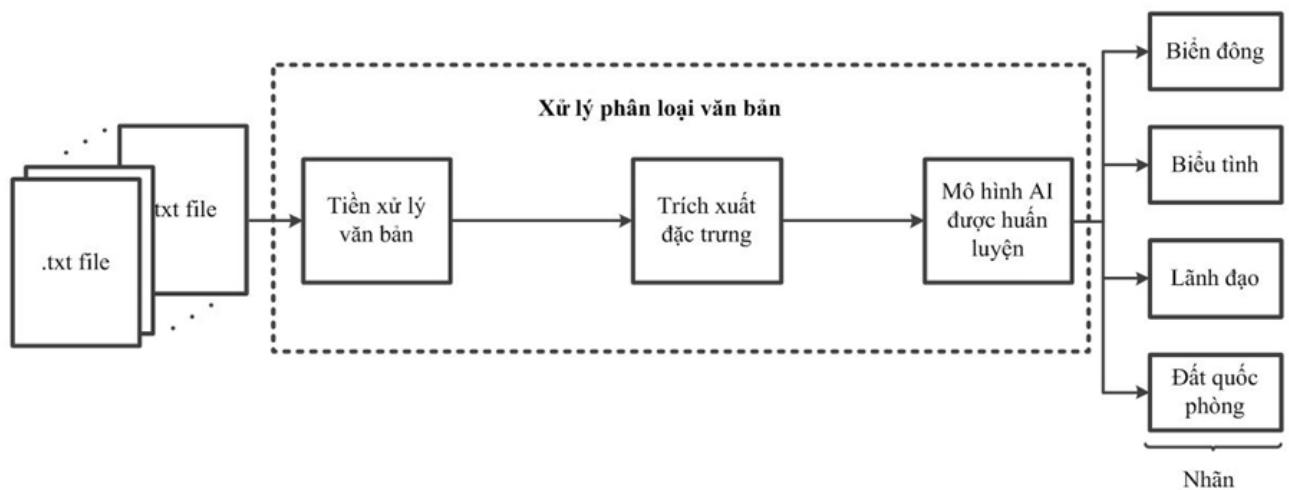
Nghiên cứu này tập trung vào việc thiết kế hệ thống tìm kiếm thông minh bằng máy học kết hợp mô hình vector hóa văn bản để cải thiện khả năng biểu diễn và tìm kiếm văn bản. Mục tiêu là thiết lập một hệ thống tìm kiếm thông minh nhanh chóng và hiệu quả nhằm hỗ trợ sự phát triển và tiến bộ của tìm kiếm thông tin. Mục tiêu của nghiên cứu là tạo ra một hệ thống tìm kiếm thông minh hiện đại sử dụng kỹ thuật học máy cùng với mô hình vector hóa văn bản sử dụng Word2Vec hoặc BERT. Xử lý và tiền xử lý cơ sở dữ liệu văn bản, phát triển các mô hình vector dữ liệu văn bản và sử dụng thuật toán tìm kiếm thông minh là tất cả các bước trong quá trình phân tích. Để cải thiện hiệu suất hiển thị và tìm kiếm văn bản mạnh mẽ, phương pháp học máy sẽ được sử dụng để tinh chỉnh và phân tích hệ thống. Kết quả mong đợi của dự án là một hệ thống tìm kiếm thông minh hoàn chỉnh sẽ đảm bảo khả năng tìm kiếm hiệu quả và phù hợp với thị hiếu người tiêu dùng. Trong lĩnh vực truy xuất thông tin, nghiên cứu này cũng sẽ hỗ trợ sự phát triển của học máy và vector hóa văn bản. Hệ thống tìm kiếm thông minh này có lợi cho các ứng dụng tìm kiếm trên web và hệ thống quản lý dữ liệu trong nhiều ngành như kinh doanh, y học và khoa học.

GIỚI THIỆU (*Tối đa 1 trang A4*)

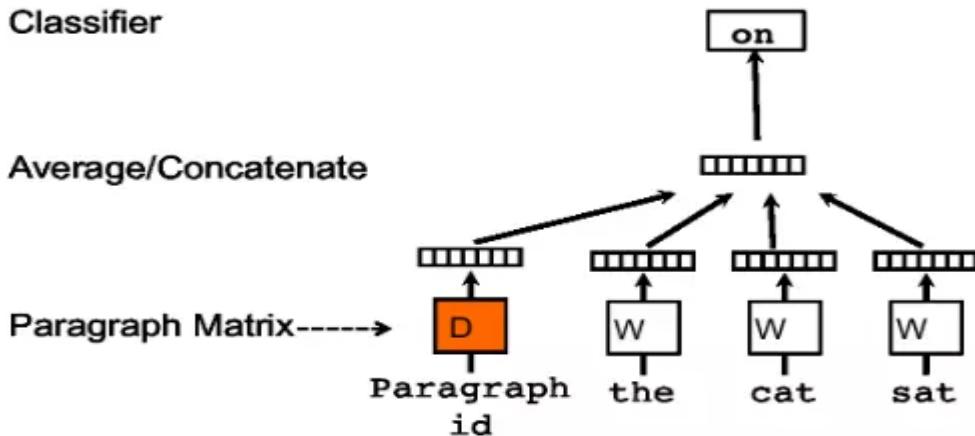
Một lĩnh vực nghiên cứu quan trọng là phát triển một hệ thống tìm kiếm thông minh

sử dụng mô hình vector hóa văn bản và phương pháp học máy. Nó hứa hẹn cải thiện hiệu suất biểu diễn và tìm kiếm văn bản. Trong thời đại thông tin phát triển nhanh chóng này, người dùng khó tìm và truy xuất thông tin. Hạn chế của hệ thống tìm kiếm truyền thống dựa trên từ khóa đã khiến nó không chính xác và khó hiểu ý định của người dùng.

Nghiên cứu này tập trung vào việc tạo ra một hệ thống tìm kiếm thông minh nhằm giải quyết những hạn chế này. Mục tiêu chính của nghiên cứu này là xây dựng một hệ thống tìm kiếm thông minh tiên tiến bằng cách sử dụng phương pháp học máy. Xử lý và tiền xử lý dữ liệu văn bản, sử dụng các mô hình vector hóa văn bản và sử dụng thuật toán tìm kiếm phù hợp sẽ là những bước trong quá trình nghiên cứu. Huấn luyện và tối ưu hóa mô hình, đảm bảo khả năng biểu diễn và tìm kiếm văn bản sẽ được thực hiện bằng phương pháp học máy.



Hình 1: Mô hình phân loại văn bản sử dụng Machine Learning



Hình 2: Mô hình vector hóa văn bản

Nghiên cứu này sẽ tạo ra một hệ thống tìm kiếm thông minh mạnh mẽ cung cấp kết quả tìm kiếm chính xác và phù hợp với nhu cầu của người dùng. Ngoài ra, mục tiêu của nghiên cứu này là cải thiện và phát triển các phương pháp học máy và mô hình vector hóa văn bản trong lĩnh vực tìm kiếm thông tin. Điều này sẽ mang lại lợi ích rõ rệt cho các ứng dụng tìm kiếm thông minh, bao gồm các công cụ tìm kiếm trực tuyến và các hệ thống quản lý dữ liệu trong nhiều lĩnh vực, chẳng hạn như kinh doanh, y tế và khoa học.

Input: Các văn bản đầu vào có thể bao gồm tài liệu, bài viết, bản tin và các nguồn thông tin khác. Các văn bản được tiền xử lý và trình bày như vector số hóa.

Output: Độ chính xác và phù hợp với nhu cầu người dùng đánh giá kết quả tìm kiếm. Hệ thống trả văn bản cho từ khóa hoặc yêu cầu tìm kiếm của người dùng.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Xây dựng một công cụ tìm kiếm tự động: xây dựng một hệ thống tìm kiếm mạnh mẽ, đáng tin cậy và chính xác có khả năng tìm kiếm và truy xuất dữ liệu từ một lượng lớn văn bản.
- Phương pháp học máy và mô hình vector hóa văn bản nên được sử dụng: Biểu diễn và phân tích văn bản bằng máy học là trọng tâm của nghiên cứu. Các mô

hình vector hóa văn bản như Word2Vec và BERT sẽ được sử dụng để hiển thị ý nghĩa của văn bản và tạo không gian vector đa chiều cho việc tìm kiếm dựa trên độ tương đồng.

- Mục tiêu là tăng hiệu suất và độ chính xác: Mục tiêu là tăng hiệu suất và độ chính xác tìm kiếm văn bản. Hệ thống sẽ được tối ưu hóa để cung cấp kết quả tìm kiếm chính xác, thích hợp và đáp ứng nhu cầu của người dùng.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

1. Nội dung

- Xử lý và tiền xử lý dữ liệu văn bản: Bước này bao gồm thu thập và tiền xử lý dữ liệu văn bản từ nhiều nguồn. Các thành phần không cần thiết, chẳng hạn như ký tự đặc biệt và từ dừng, sẽ được loại bỏ từ dữ liệu và chuyển đổi nó về dạng phù hợp.
- Mục tiêu của việc trình bày văn bản bằng các mô hình vector hóa là trình bày ý nghĩa của văn bản như các vector số hóa. Để tạo không gian vector đa chiều cho việc tìm kiếm và phân loại, các mô hình vector hóa văn bản như Word2Vec, GloVe và BERT sẽ được sử dụng.
- Áp dụng phương pháp học máy: Quá trình này sẽ huấn luyện mô hình tìm kiếm bằng cách sử dụng Học có giám sát (Supervised Learning) hoặc Học không giám sát (Unsupervised Learning). Mô hình sẽ tìm hiểu cách đánh giá và phân loại văn bản dựa trên độ tương đồng và sự phù hợp với yêu cầu tìm kiếm.
- Tối ưu hóa và thử nghiệm mô hình: Trong quá trình này, các tham số của mô hình phải được tối ưu hóa để tăng hiệu suất và độ chính xác. Các kỹ thuật tối ưu hóa như Gradient Descent, Regularization và Cross-validation sẽ được sử dụng. Đồng thời, mô hình sẽ được kiểm tra bằng các chỉ số như độ chính xác, độ phủ và độ đo F1 để đảm bảo rằng kết quả tìm kiếm là chính xác và liên quan.

2. Phương pháp

- Tiền xử lý và biểu diễn dữ liệu văn bản: Tiền xử lý loại bỏ các phần không cần thiết từ dữ liệu văn bản và chuyển đổi nó sang dạng phù hợp. Văn bản sẽ được trình bày thành các vector số hóa bằng cách sử dụng các mô hình vector hóa văn bản như Word2Vec hoặc BERT.
- Xây dựng và huấn luyện mô hình tìm kiếm: Các phương pháp và thuật toán học máy sẽ được sử dụng để xây dựng một mô hình tìm kiếm thông minh. Mục đích của việc huấn luyện mô hình bằng cách sử dụng tập dữ liệu tiền xử lý và biểu diễn là học cách phân loại và đánh giá văn bản dựa trên độ tương đồng và phù hợp.
- Tối ưu hóa mô hình: Các phương pháp tối ưu hóa như giảm gradient, quản lý và đối chiếu sẽ được sử dụng để điều chỉnh các tham số của mô hình. Mục tiêu là cải thiện hiệu suất và độ chính xác của mô hình tìm kiếm.
- Đánh giá và kiểm tra mô hình: Mô hình tìm kiếm sẽ được đánh giá bằng các tập dữ liệu kiểm tra. Kết quả tìm kiếm sẽ được so sánh với các kết quả đã biết trước hoặc các hệ thống tìm kiếm khác để đánh giá hiệu suất và độ chính xác của mô hình.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Tăng hiệu suất tìm kiếm: Hệ thống tìm kiếm thông minh sẽ cung cấp kết quả tìm kiếm chính xác và liên quan để đáp ứng nhu cầu của người dùng. Việc sử dụng các mô hình vector hóa văn bản và phương pháp học máy dự kiến sẽ mang lại kết quả tìm kiếm vượt trội so với các hệ thống tìm kiếm truyền thống.
- Độ phủ và độ chính xác đáng kinh ngạc: Để đánh giá mô hình tìm kiếm thông minh, các chỉ số độ chính xác, độ phủ và độ đo F1 sẽ được sử dụng. Để đảm bảo rằng kết quả tìm kiếm bao gồm đầy đủ các văn bản liên quan đến yêu cầu của người dùng, chúng tôi hy vọng đạt được độ chính xác và độ phủ cao.
- Tối ưu hóa hiệu suất và độ chính xác là mục tiêu: Đánh giá và tối ưu hóa kết

quả sẽ giúp cải thiện hiệu suất và độ chính xác của hệ thống. Chúng tôi sẽ điều chỉnh các tham số và đảm bảo rằng mô hình hoạt động tốt nhất trong biểu diễn và tìm kiếm văn bản bằng cách sử dụng các phương pháp tối ưu hóa như Gradient Descent, Regularization và Cross-validation.

- **Ứng dụng thực tế:** Kết quả của nghiên cứu cho thấy các ứng dụng tìm kiếm thông minh sẽ có lợi. Hệ thống tìm kiếm thông minh có thể cải thiện trải nghiệm tìm kiếm của người dùng trong nhiều lĩnh vực, từ công cụ tìm kiếm trên internet cho đến các hệ thống quản lý dữ liệu trong các lĩnh vực như y tế, khoa học và kinh doanh.

TÀI LIỆU THAM KHẢO (*Định dạng DBLP*)

- [1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
- [2] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
- [3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies (NAACL-HLT) (Vol. 1, pp. 4171-4186).
- [4] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of machine learning research, 3(Feb), 1137-1155.
- [5] Goldberg, Y., & Levy, O. (2014). word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.
- [6] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1746-1751).
- [7] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information

retrieval. Cambridge University Press.

- [8] Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. Cambridge University Press.
- [9] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.
- [10] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196).
- [11] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Google Brain. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.
- [12] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in neural information processing systems (pp. 8026-8037).
- [13] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Ghemawat, S. (2016). TensorFlow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI) (pp. 265-283).