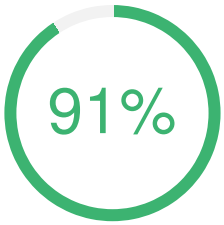
BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

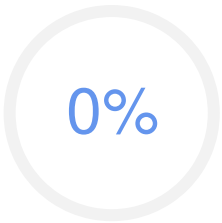
Tên tài liệu:	63CNTT4_2151062703_NguyenDucAnh_DATN_v2-content
Tác giả:	Ngành CNTT
Điểm trùng lặp:	9
Thời gian tải lên:	14:22 18/01/2026
Thời gian sinh báo cáo:	14:24 18/01/2026
Các trang kiểm tra:	42/42 trang

Kết quả kiểm tra trùng lặp

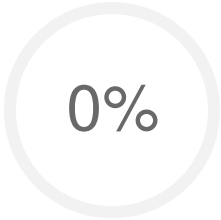
Có 9% nội dung trùng lặp



Có 91% nội dung không trùng lặp



Có 0% nội dung người dùng loại trừ



Có 0% nội dung hệ thống bỏ qua

Nguồn trùng lặp tiêu biểu

123docz.net ptithcm.edu.vn wecan-group.com

Danh sách các câu trùng lặp

Câu 1. Trang 1: trong bối cảnh các lĩnh vực công nghệ, ngày càng phát triển mạnh mẽ, việc tự động nhận diện cảm xúc trong văn bản tiếng Việt đã và đang được ứng dụng rộng rãi trong nhiều lĩnh vực như quản tri doanh nghiệp, xây dựng và phát triển thương hiệu chăm sóc khách hàng, khảo sát ý kiến và phân tích đánh giá, phản hồi từ người dùng

Độ trùng lặp: *51%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *các lĩnh vực công nghệ việc nhận diện cảm xúc Trong văn bản tiếng Việt được ứng dụng Trong nhiều lĩnh vực như quản tri doanh nghiệp, quản tri thương hiệu, sản phẩm, quản tri quan hệ khách hàng, khảo sát ý kiến khách hàng, hay dễ hiểu hơn là phân tích đánh giá ý kiến phản hồi*

Câu 2. Trang 1: Ý kiến và cảm nhận của khách hàng hiện nay đóng Vai trò vô cùng quan trọng trong việc định hướng sản phẩm cũng như chiến lược kinh doanh

Độ trùng lặp: *50%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *định hướng sản phẩm trong marketing là yếu tố vô cùng cần thiết định vị sản phẩm nên dựa vào lợi ích sản phẩm đưa ra, khách hàng là ai, định vị sản phẩm của đối thủ cạnh tranh như thế nào vai trò của định vị sản phẩm định vị sản phẩm/dịch vụ có vai trò vô cùng quan trọng trong chiến lược marketing của doanh nghiệp, ảnh hưởng đến việc định hướng hoạt động của các chiến lược kế hoạch marketing được đề ra như chiến lược*

Câu 3. Trang 2: Trong nhiều nghiên cứu, bài toán này thường được đơn giản hóa thành bài toán phân lớp

Độ trùng lặp: *50%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *được đơn giản hóa thành bài toán phân*

Câu 4. Trang 2: một số phương pháp tiêu biểu được áp dụng để xử lý bài toán phân tích cảm xúc bao gồm

Độ trùng lặp: *52%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Một số phương pháp tiền xử lý dữ liệu và áp dụng them các thuật toán phân lớp hay tối ưu các thuật toán phân lớp hiện có để mô hình giải quyết bài toán nhận diên cảm xúc trong văn bản tiếng Việt được tốt hơn 2 3 Kiến nghị phân tích cảm xúc*

Câu 5. Trang 3: Phương pháp học máy (machine Learning)

Độ trùng lặp: *100%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Phương pháp học máy (Machine learning*

Câu 6. Trang 3: Một trong những hướng tiếp cận hiện đại là Sử dụng các mô hình học máy như Support Vector Machine (SVM), Logistic Regression

Độ trùng lặp: *50%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *sử dụng các mô hình học máy như Support Vector Machine (SVM),*

Câu 7. Trang 3: Đồ án kết hợp giữa nghiên cứu lý thuyết và Xây dựng mô hình thực nghiệm

Độ trùng lặp: *75%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *kết hợp giữa nghiên cứu lý thuyết và đánh giá kiểm chứng lý thuyết thông qua thực nghiệm Về lý thuyết xây dựng mô hình*

Câu 8. Trang 3: Về mặt lý thuyết, nghiên cứu tổng quan Về lĩnh vực phân tích, cảm xúc trong văn bản tiếng việt, các phương pháp phổ biến trong nhận diện cảm xúc và một số mô hình tiên tiến được ứng dụng trong các công trình khoa học

Độ trùng lặp: *59%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *tổng quan Về cảm xúc trong văn bản tiếng Việt, các phương pháp nhận dạng cảm xúc đồng thời cũng trình bày một số mô hình nhận diện cảm xúc được tổng*

Câu 9. Trang 5: Để thực hiện phân tích, cảm xúc, ở quy mô lớn việc Ứng dụng Xử lý ngôn ngữ tự nhiên (Natural Language Processing NLP) là phù hợp vì NLP Cho phép máy tính Xử lý và khai thác ý nghĩa từ văn bản

Độ trùng lặp: *50%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Xử lý ngôn ngữ tự nhiên (Natural Language Processing NLP) cho phép máy tính hiểu, phân tích thao tác và tạo ra ngôn ngữ của con người Hiểu ngôn ngữ tự nhiên (Natural Language Understanding NLU) Giúp máy tính hiểu ý nghĩa của văn bản hoặc lời nói Tạo ngôn ngữ tự nhiên (Natural Language*

Câu 10. Trang 6: Xây dựng quy trình xử lý dữ liệu bình luận thu thập, Làm sạch, chuẩn hóa và

Độ trùng lặp: *62%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *quy trình xử lý và khai thác dữ liệu Chức năng chính của module là thu thập, đồng bộ và tích hợp dữ liệu từ cơ sở dữ liệu dân cư quốc gia và nhiều nguồn cung cấp dữ liệu khác nhau Module xử lý dữ liệu Data Processing làm sạch, chuẩn hóa và*

Câu 11. Trang 7: Học máy (Machine Learning) là cho phép máy tính Học ra quy luật từ một hay nhiều tập dữ liệu giúp dự đoán hoặc quyết định mà không cần lập trình thủ công

Độ trùng lặp: *61%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Học máy (Machine Learning) là công nghệ cho phép máy tính Học từ dữ liệu và đưa ra dự đoán hoặc quyết định mà không cần được lập trình cụ thể cho từng nhiệm vụ Hệ thống Học máy sẽ phân tích các mẫu dữ liệu nhận ra quy luật*

Câu 12. Trang 9: trong đồ án, Logistic Regression được sử dụng như một mô hình baseline quan trọng để so sánh với các mô hình khác khi kết hợp với TF IDF

Độ trùng lặp: *51%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *được sử dụng như một mô hình nền tảng (baseline) để so sánh với các mô hình*

Câu 13. Trang 10: các điểm quan trọng, nằm gần biên được gọi là support vectors và có vai trò quyết định vị trí siêu phẳng

Độ trùng lặp: *63%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *được gọi là support vectors Đây là những điểm có vai trò quan trọng quyết định vị trí*

Câu 14. Trang 10: TF IDF là một kỹ thuật biểu diễn văn bản thành vector số trong đó mỗi từ (hoặc cụm từ n gram) được gán một trọng số phản ánh mức độ quan trọng của từ đó đối với một văn bản Cụ thể, trong toàn bộ tập dữ liệu TF IDF đánh giá giá trị thông tin của từ đó trên hai yếu tố tần suất xuất hiện trong văn bản và mức độ phổ biến của từ trong toàn bộ tập văn bản Cụ thể, TF IDF là Sự kết hợp của TF (Term Frequency) số lần một từ xuất hiện trong một văn bản và IDF(Inverse Document Frequency) độ hiếm của từ đó trên toàn bộ tập dữ liệu

Độ trùng lặp: *54%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *tần suất xuất hiện của một từ trong một văn bản và ước lượng mức độ quan trọng của từ đó trong ngữ liệu cụ thể TF (Term Frequency) đo lường tần suất xuất hiện của một từ trong một mục, trong khi IDF (Inverse Document Frequency) đo lường tần suất xuất hiện của từ đó trong toàn bộ tập dữ liệu sự kết hợp giữa TF và IDF tạo ra một trọng số, cho mỗi từ thể hiện độ quan trọng của từ đó đối với một mục cụ thể Cosine Similarity Cosine Similarity là một phương pháp phổ biến để đo lường độ tương đồng giữa hai vectơ trong Content Based Recommendation System, vectơ thường biểu diễn trong một văn bản và ước lượng mức độ quan trọng của từ đó trong ngữ liệu cụ thể TF (Term Frequency) đo lường tần suất xuất hiện của một từ trong một mục, trong khi IDF (Inverse Document Frequency) đo lường tần suất xuất hiện của từ đó trong toàn bộ tập dữ liệu sự kết hợp*

Câu 15. Trang 11: Khi đó mỗi văn bản sẽ được biểu diễn dưới dạng một vector có số chiều kích thước từ điển V

Độ trùng lặp: *70%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *biểu diễn dưới dạng một vector n chiều Theo mô hình này, mỗi văn bản sẽ được biểu diễn trong một không gian vector có số chiều lớn, trong đó mỗi chiều của không gian tương ứng với một từ*

Câu 16. Trang 11: tính TF Term Frequency (tần suất từ trong văn bản)

Độ trùng lặp: *88%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *TF Term Frequency*

Câu 17. Trang 11: với mỗi văn bản d tf đo mức độ xuất hiện của từ t trong chính văn bản đó

Độ trùng lặp: *62%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Với văn bản d, trong một tập các văn bản d, tfidf (t d, D) TF (t d,) idf (t D) ( )*

TF (Term Frequency) Gọi f (t d,) là số lần xuất hiện của từ t trong

Câu 18. Trang 11: TF(t,d)=f(t,d) f(t, d) số lần từ txuất hiện trong văn bản d

Độ trùng lặp: *100%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *hiện trong văn bản d*

Câu 19. Trang 11: Tính IDF inverse document frequency (độ hiếm của từ trong tập dữ liệu)

Độ trùng lặp: *58%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Tính IDF Inverse Document Frequency Tần số nghịch của từ trong tập*

Câu 20. Trang 12: Nlà tổng số văn bản trong tập d df(t)là số văn bản có chứa từ t

Độ trùng lặp: *80%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *văn bản có chứa từ t N tổng số văn bản trong tập dữ*

Câu 21. Trang 12: trọng số TF IDF, của từ t trong văn bản d được tính bằng tích của TF và

Độ trùng lặp: *70%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *TF IDF Trọng số từ là tích của tần suất từ (TF) và tần suất nghịch đảo của từ đó (IDF) Đây là phương pháp kết hợp được ưu điểm của cả hai phương pháp trên Trọng số được tính bằng tần suất xuất hiện của từ t trong văn bản d*

Câu 22. Trang 13: 2 3 Các mô hình nhận diện cảm xúc trong văn bản

Độ trùng lặp: *84%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Các mô hình nhận diện cảm xúc trong văn bản 15 CHƯƠNG 3*

Câu 23. Trang 13: Từ đó, chọn Lựa các đặc trưng phù hợp để đưa vào mô hình phân tích

Độ trùng lặp: *64%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *lựa chọn các đặc trưng phù hợp để đưa vào mô hình*

Câu 24. Trang 14: học có giám sát (Supervised Learning) Hệ thống học từ dữ liệu đã được gán nhãn, để dự đoán nhãn, cho dữ liệu mới

Độ trùng lặp: *78%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *giám sát (Supervised Learning) Hệ thống được đào tạo trên một tập dữ liệu đã được gán nhãn Học cách dự đoán nhãn cho các dữ liệu mới dựa trên mẫu đã Học Ví dụ, một Hệ thống Học có giám sát*

Câu 25. Trang 14: Học không giám sát (Unsupervised Learning) dữ liệu đầu vào không có nhãn

Độ trùng lặp: *91%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Dữ liệu đầu vào Học không giám sát (Unsupervised Learning) Dữ liệu không có nhãn*

Câu 26. Trang 14: Học bán giám sát (Semi supervised Learning) kết hợp giữa hai phương pháp trên, một phần dữ liệu được gán nhãn phần còn lại không có nhãn

Độ trùng lặp: *71%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *giám sát (Semi supervised Learning) Học bán giám sát Kết hợp cả hai phương pháp Học có giám sát và Học không giám sát trong đó một phần dữ liệu được gán nhãn, và phần còn lại không*

Câu 27. Trang 14: học tăng cường (Reinforcement Learning) mô hình, học qua quá trình thử sai, nhận phần thưởng hoặc hình, phạt, để điều chỉnh hành vi

Độ trùng lặp: *63%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Học tăng cường (Reinforcement Learning RL) dựa trên quá trình thử nghiệm và nhận phản hồi để tối ưu hóa hành vi Trong Reinforcement Learning, một tác nhân (agent) thực hiện hành động trong môi trường, nhận phần thưởng hoặc hình phạt*

Câu 28. Trang 15: quá trình nghiên cứu được thực hiện theo các bước cơ bản sau

Độ trùng lặp: *91%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Quá trình nghiên cứu có thể được thực hiện theo các bước cơ bản sau (*

Câu 29. Trang 15: o Tập huấn luyện được sử dụng để đào tạo mô hình học máy

Độ trùng lặp: *94%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Tập huấn luyện Được sử dụng để đào tạo mô hình và Tập kiểm tra Được sử dụng để đánh giá hiệu suất của mô hình Bước 3 Xây dựng mô hình học máy*

Câu 30. Trang 15: mô hình học từ mối quan hệ giữa đặc trưng và nhãn để tìm ra quy luật phân loại

Độ trùng lặp: *64%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *quy luật và mối quan hệ giữa*

Câu 31. Trang 15: o tập kiểm thử Sau khi mô hình học xong, tập dữ liệu kiểm thử được sử dụng để đánh giá khả năng dự đoán của mô hình với dữ liệu chưa từng thấy

Độ trùng lặp: *59%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *để đánh giá khả năng dự đoán của mô hình học máy Việc sử dụng Tập kiểm tra giúp đánh giá mức độ tổng quát của mô hình đối với dữ liệu chưa biết Tập huấn luyện Training*

Set Là một Tập các quan sát được sử dụng để

Câu 32. Trang 16: Việc đánh giá Hiệu suất mô hình được thực hiện qua các chỉ số đo lường chất lượng như độ chính xác (accuracy), độ thu hồi (recall), độ chính xác (precision) F1 score

Độ trùng lặp: *53%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *mô hình được đo lường thông qua các chỉ số đánh giá như độ chính xác (accuracy), độ*

Câu 33. Trang 17: CHƯƠNG 3 NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT 3 1 Tiền xử lý Dữ liệu VĂN BẢN

Độ trùng lặp: *89%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *CHƯƠNG 3 NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT 3 1 Tiền xử lý ngữ liệu dữ liệu*

Câu 34. Trang 17: Tách từ (word segmentation) là một bước không thể thiếu trong tiền xử lý ngữ liệu tiếng Việt

Độ trùng lặp: *52%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *liệu là một bước không thể thiếu trong*

Câu 35. Trang 17: Điều này khiến cho việc tách từ trong tiếng Việt trở nên phức tạp và là một thách thức trong các bài toán xử lý ngôn ngữ tự nhiên

Độ trùng lặp: *58%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *trong tiếng Việt là một bài toán quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên Đặc thù của tiếng Việt đóng góp vào độ khó của bài toán này tiếng Việt có nhiều biến thể ngôn ngữ bao gồm việc sử dụng dấu thanh, dấu câu và các biến thể từ ngữ Điều này làm cho việc xử lý và phân tích văn bản tiếng Việt trở nên phức tạp*

Câu 36. Trang 17: Tuy nhiên, khi ghép lại chúng tạo thành một từ mang ý nghĩa khác biệt

Độ trùng lặp: *57%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *tạo thành một từ mang ý nghĩa*

Câu 37. Trang 18: Ăn cơm không được uống rượu có thể tách thành o Ăn / cơm / không / được / uống / rượu

Độ trùng lặp: *81%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Ăn / cơm / không / được / uống / rượu Ăn / cơm không / được / uống / rượu*

Câu 38. Trang 18: ví dụ Ẩm thực Việt Nam, ẩm_thực Việt_Nam

Độ trùng lặp: *100%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Ví dụ ẩm thực Việt Nam ẩm thực*

Câu 39. Trang 18: hiện nay có bốn hướng tiếp cận chính trong việc tách từ tiếng Việt

Độ trùng lặp: *58%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *việc tách từ tiếng Việt (Hình) Hình Các hướng tiếp cận cơ bản trong việc phân đoạn văn bản tiếng Hoa và các hướng tiếp cận Hiện nay*

Câu 40. Trang 18: dựa vào từ điển so khớp các chuỗi từ trong văn bản với từ điển từ vựng có sẵn

Độ trùng lặp: *57%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Dựa vào từ điển từ có sẵn rồi dùng các biện pháp So khớp để tách ra các từ cụm từ trong văn bản*

Câu 41. Trang 19: kết hợp từ điển và thống kê Phương pháp kết hợp giúp tận dụng cả ưu điểm của từ điển và tính thực tiễn từ dữ liệu

Độ trùng lặp: *53%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Phương pháp Kết hợp Phương pháp này Kết hợp sử dụng cả từ điển và thống kê để tận dụng ưu điểm của*

Câu 42. Trang 19: Cùng một từ ngữ có thể được viết dưới nhiều dạng khác nhau, chẳng hạn

Độ trùng lặp: *61%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *có thể được viết dưới nhiều dạng khác nhau, một hiện tượng hóa học được gọi tên bằng những thuật ngữ*

Câu 43. Trang 20: Trong nghiên cứu này, chúng tôi sử dụng phương pháp TF IDF (Term Frequency Inverse Document Frequency) để vector hóa văn bản

Độ trùng lặp: *74%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *dụng phương pháp TF IDF (Term Frequency Inverse Document Frequency) để*

Câu 44. Trang 20: TF IDF đo lường mức độ quan trọng của một từ dựa trên hai yếu tố (một) tần suất xuất hiện của từ trong một văn bản Cụ thể, và (hai) mức độ hiếm của từ đó trong toàn bộ tập văn bản

Độ trùng lặp: *62%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *trong một văn bản và ước lượng mức độ quan trọng của từ đó trong ngữ liệu cụ thể, TF (Term Frequency) đo lường tần suất xuất hiện của một từ trong một mục, trong khi IDF (Inverse Document Frequency) đo lường tần suất xuất hiện của từ đó trong toàn bộ tập dữ liệu Sự kết hợp giữa TF và IDF tạo ra một trọng số cho mỗi từ thể, hiện độ quan trọng của từ đó đối với một mục cụ thể, Cosine Similarity Cosine Similarity là một phương pháp phổ biến để đo lường độ*

tương đồng giữa hai

Câu 45. Trang 21: 3 2 1 Các phương pháp biểu diễn văn bản cổ điển (Classic Word Embedding)

Độ trùng lặp: *64%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Các phương pháp biểu diễn văn bản*

Câu 46. Trang 21: Mỗi văn bản được biểu diễn bằng một vector có số chiều tương ứng với kích thước của bộ từ vựng

Độ trùng lặp: *61%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *được biểu diễn bằng một one hot vector có dạng ( ) = [ ] ( ) ( ) = [ ] ( ) Mỗi hàng của ma trận W là một biểu diễn vector có số chiều là N tương ứng với một từ w trong tập từ vựng Ma trận h với kích thước*

Câu 47. Trang 21: TF IDF là Phương pháp cải tiến từ BoW, Giúp đánh giá tầm quan trọng của một từ trong một văn bản so với toàn bộ tập tài liệu

Độ trùng lặp: *51%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *tầm quan trọng của một từ trong một văn bản dựa trên toàn bộ văn bản trong cơ sở dữ liệu phương pháp TF IDF TF IDF (Term frequency inverse document frequency) giúp thống kê các từ các đoạn từ trong đoạn văn bản (hay trong các trường của dữ liệu trong dữ liệu của bài này) Term frequency (TF) là tần số xuất hiện của một từ Số lần xuất hiện của từ đó so với*

Câu 48. Trang 21: TF (Term frequency) tần suất xuất hiện của một từ trong một văn bản

Độ trùng lặp: *100%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *từ trong một văn bản thông qua thống kê TF (Term frequency) Tần suất xuất hiện của một từ trong*

Câu 49. Trang 22: trong đó để đánh giá tầm quan trọng của từ trong văn bản,

Độ trùng lặp: *86%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *để đánh giá tầm quan trọng của từ Trong văn bản*

Câu 50. Trang 22: o ni số lần từ i xuất hiện trong văn bản o ni tổng số từ trong văn bản

Độ trùng lặp: *70%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *số lần từ i xuất hiện trong văn bản k N là tổng số các văn bản M là tổng số các từ khác nhau ni*

Câu 51. Trang 22: IDF (Invert Document Frequency) Hay tần số văn bản nghịch đảo được dùng

Độ trùng lặp: *55%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *tần số văn bản nghịch đảo. (IDF Inverse Document Frequency) giá trị trọng số của từ được*

Câu 52. Trang 22: TF IDF (Term frequency Invert Document Frequency) Sự kết hợp của tần số từ TF và tần số văn bản nghịch đảo IDF

Độ trùng lặp: *59%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *kết hợp của tần số từ khóa (TF Term Frequency) và nghịch đảo số văn bản chứa từ khóa (IDF Inverted Document Frequency)*

Câu 53. Trang 22: Word2Vec là một trong những phương pháp embedding hiện đại phổ biến nhất hiện nay,

Độ trùng lặp: *78%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *là một trong những phương pháp kế hoạch hóa gia đình hiện đại phổ biến nhất hiện nay*

Câu 54. Trang 23: CBOW (Continuous Bag of Words) Dùng các từ xung quanh để dự đoán từ

Độ trùng lặp: *71%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *CBOW (Continuous Bag of Words) Lấy từ mục tiêu làm đầu ra dự đoán từ đầu vào là các từ xung quanh nó để*

Câu 55. Trang 23: Skip gram Dùng một từ để dự đoán các từ xung quanh nó trong một cửa sổ ngữ cảnh

Độ trùng lặp: *70%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *dự đoán các từ xung quanh từ mục tiêu trong một cửa sổ ngữ cảnh [*

Câu 56. Trang 23: FastText là phiên bản cải tiến của Word2Vec, được phát triển bởi Facebook AI

Độ trùng lặp: *64%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *là phiên bản cải tiến của mô hình ngôn ngữ GPT 3, được phát triển bởi OpenAI*

Câu 57. Trang 24: với TF IDF giúp gán trọng số cho từng từ, theo mức độ quan trọng trong toàn bộ tập dữ liệu

Độ trùng lặp: *51%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *trong toàn bộ tập dữ liệu Sự kết hợp giữa TF và IDF tạo ra một trọng số cho mỗi từ thể hiện độ quan trọng của từ đó đối Với*

Câu 58. Trang 24: một phần dữ liệu được sử dụng để huấn luyện phần còn lại dùng để đánh giá độ chính xác của các mô hình,

Độ trùng lặp: *79%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *sử dụng để huấn luyện, và xác nhận chéo mô hình nhằm lựa chọn siêu tham số của các mô hình ML phần dữ liệu thử nghiệm sử dụng để đánh giá độ chính xác của các mô hình được*

Câu 59. Trang 25: việc tự thu thập gán nhãn và xử lý dữ liệu, là công việc mất nhiều thời gian, công sức, và tài nguyên

Độ trùng lặp: *57%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *thu thập, và xử lý dữ liệu đây mới là công Việc chính của người làm kế toán, Việc nhập dữ liệu không phải là công Việc mất nhiều thời gian, kỹ năng và công sức*

Câu 60. Trang 27: Tổng số lượng mẫu trong của cả data_train và data_test là 31 460 mẫu dữ liệu bình luận 4 2 Tiền xử lý dữ liệu

Độ trùng lặp: *56%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Tổng số lượng mẫu của trong cả data_train và data_test là 50 000 mẫu dữ liệu bình luận (*

Câu 61. Trang 28: Do đó, việc làm sạch và chuẩn hóa dữ liệu là một bước không thể thiếu nhằm đảm bảo chất lượng tập huấn luyện

Độ trùng lặp: *56%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Do đó, việc làm sạch và xác thực dữ liệu là một bước không thể thiếu trong quá trình này Quá trình làm sạch dữ liệu giúp loại bỏ các lỗi và thông tin không hợp lệ, đảm bảo*

Câu 62. Trang 29: Tách từ Đây là bước đặc biệt quan trọng trong xử lý tiếng Việt, do tiếng Việt, sử

Độ trùng lặp: *54%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *đặc biệt quan trọng trong xử lí tiếng Việt, do tiếng Việt, là*

Câu 63. Trang 29: Do đó, việc làm sạch và chuẩn hóa dữ liệu là một bước không thể thiếu nhằm đảm bảo chất lượng tập huấn luyện

Độ trùng lặp: *56%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Do đó, việc làm sạch và xác thực dữ liệu là một bước không thể thiếu trong quá trình này Quá trình làm sạch dữ liệu giúp loại bỏ các lỗi và thông tin không hợp lệ, đảm bảo*

Câu 64. Trang 30: Nhờ cơ chế này TF IDF làm nổi bật các từ mang tính phân biệt, cao và giảm ảnh hưởng của các từ xuất hiện quá phổ biến

Độ trùng lặp: *50%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *giảm ảnh hưởng của các từ phổ biến nhưng ít mang tính phân biệt đồng thời làm*

nổi bật các từ

Câu 65. Trang 30: từ các đặc trưng TF IDF đã thu được hệ thống sử dụng các thuật toán học có giám sát để huấn luyện mô hình phân loại cảm xúc

Độ trùng lặp: *54%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *sử dụng các thuật toán học máy có giám sát hoặc không có giám sát để huấn luyện mô hình rút gọn mô phỏng sở thích của người dùng các thuật toán học*

Câu 66. Trang 32: để Kiểm tra hiệu quả của mô hình các nhãn dự đoán được So sánh với nhãn thực tế trong tập test

Độ trùng lặp: *52%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Để kiểm tra mô hình. thị giác máy tính của bạn và hiểu hiệu suất của nó Chạy dự đoán Sử dụng mô hình. Để đưa ra dự đoán trên tập dữ liệu thử nghiệm so sánh các dự đoán kiểm tra xem các dự đoán của mô hình.*

Câu 67. Trang 32: Hai thuật toán học máy SVM và Logictic được sử dụng để huấn luyện mô hình trên tập dữ liệu đã chuẩn hóa

Độ trùng lặp: *62%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *thuật toán học máy không giám sát để huấn luyện mô hình trên tập dữ liệu đã được tiền xử lý và phân tích đặc trưng Đánh giá mô hình Đánh giá hiệu suất của mô hình bằng cách sử dụng*

Câu 68. Trang 32: Toàn bộ thí nghiệm được thực hiện trên thiết bị có thông số kỹ thuật như sau

Độ trùng lặp: *50%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *thiết bị có thông số kỹ thuật như sau*

Câu 69. Trang 33: Phân loại bằng mô hình học máy SVM và Logistic Regression (trên tập dữ liệu kiểm tra)

Độ trùng lặp: *50%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *tập dữ liệu kiểm tra*

Câu 70. Trang 37: về precision tức độ chính xác của mô hình khi dự đoán một nhãn

Độ trùng lặp: *64%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Về độ chính xác của mô hình khi dự đoán*

Câu 71. Trang 37: Hình 4 6 Kết quả dữ liệu kiểm tra với PP Logistic Regression

Độ trùng lặp: *75%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *dữ liệu kiểm tra với PP Logistic Regression 35 Hình 4 6*

Câu 72. Trang 38: Về F1 Score Thước đo cân bằng giữa Precision và Recall

Độ trùng lặp: *90%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *thước đo cân bằng giữa precision và recall*

Câu 73. Trang 42: [2] Võ Trường Duy, Bài 1 Giới thiệu về Machine Learning Machine Learning cơ bản 26/12/2016

Độ trùng lặp: *54%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Machine Learning , cơ bản, Mạng neuron tích chập Bài 1 Giới thiệu về Machine Learning , Bài 2*

Câu 74. Trang 42: Võ Trường Duy, bài 33 Các phương pháp đánh giá một hệ thống phân lớp, (Precision, Recall, F1,) , Machine Learning Cơ Bản, 31/08/2017

Độ trùng lặp: *58%*

Nguồn: *Dữ liệu nội sinh*

Nội dung nguồn: *Bài 33 Các phương pháp đánh giá một hệ thống phân lớp Machine Learning cơ bản,*

--- Hết ---