Big Data Course

Capstone Project   
Action Plan

For students (instructor’s review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

|  |  |
| --- | --- |
| Course | Big Data Course |
| Team Name | BigData02 – Nhóm 2 |
| Team Leader/  Members | Lê Minh Kiệt /  Võ Trần Yến Như, Dương Ngọc Minh Huy, Trần Hữu Hậu, Trần Trọng Nhân, Nguyễn Thanh Nam, Nguyễn Thanh Nở |
| Project Title | Phân tích dữ liệu tài chính và thương mại điện tử |
| Goal |  |
| Phát triển hệ thống giúp khách hàng nhận dạng nhanh chóng được sản phẩm chất lượng hay không dựa vào bình luận của khách hàng, làm giảm nhu cầu đọc bình luận thủ công của khách hàng | |
| Abstract |  |
| Trong bối cảnh mua sắm trực tuyến ngày càng phổ biến, việc đánh giá chất lượng sản phẩm dựa trên nhận xét của người tiêu dùng đã trở thành yếu tố quan trọng trong quyết định mua hàng. Tuy nhiên, với số lượng nhận xét quá lớn, người tiêu dùng thường mất nhiều thời gian để chọn lọc thông tin hữu ích. Dự án này tập trung vào việc phát triển một hệ thống thông minh, sử dụng các thuật toán học máy như SVM, Logistic Regression và Random Forest, để tự động phân tích và phân loại nhận xét của người dùng. Hệ thống sẽ giúp người tiêu dùng nhanh chóng nhận diện các sản phẩm có chất lượng tốt hoặc kém dựa trên những nhận xét này. Trong quá trình thử nghiệm, mô hình Random Forest đã chứng tỏ được hiệu quả cao nhất và được tích hợp vào ứng dụng web Flask, cho phép người dùng dễ dàng thực hiện phân tích chỉ với vài thao tác đơn giản | |

|  |  |
| --- | --- |
| Method |  |
| Trong dự án này, chúng tôi áp dụng phương pháp học máy để giải quyết bài toán phân loại nhận xét của người dùng về sản phẩm thành ba nhóm: chưa tốt, trung bình, tốt. Chúng em tạo ra một mô hình Random Forest với 200 cây quyết định. Mỗi cây được huấn luyện trên một tập con dữ liệu khác nhau, nhằm đảm bảo rằng mô hình có thể tận dụng tối đa thông tin từ tập dữ liệu và đưa ra dự đoán đáng tin cậy. Kết quả của mô hình Random Forest sau khi huấn luyện cho thấy nó có hiệu suất tốt nhất trong việc phân loại nhận xét người dùng, so với các thuật toán khác mà chúng tôi đã thử nghiệm. Mô hình được huấn luyện trên tập dữ liệu đã được chia thành hai phần: tập huấn luyện và tập kiểm tra. Tập huấn luyện chiếm 80% dữ liệu và tập kiểm tra chiếm 20% còn lại. | |

|  |  |
| --- | --- |
| Data |  |
| Dữ liệu của chúng tôi là dữ liệu dạng chữ (ngôn ngữ tự nhiên). Chúng tôi thu thập dữ liệu từ sàn thương mại điện tử Lazada, dữ liệu thu thập về bao gồm nội dung bình luận và số sao mà người mua hàng đã đánh giá. Do chính sách bảo mật thông tin khách hàng nên quá trình thu thập dữ liệu đã gặp một chút khó khăn. Để giảm thiểu tối đa tình trạng về sự cố bảo mật thì tôi đã chia nhỏ việc thu thập dữ liệu cho tất cả thành viên. | |
| Expected  Outcome |  |
| Dữ liệu đầu ra là một con số (%) - biểu thị mức độ chất lượng của sản phẩm. Số càng cao thì mức độ chất lượng của sản phẩm càng tốt. Dữ liệu đầu ra của chúng tôi sẽ có thể giúp người mua hàng có thêm độ tin cậy và thuận tiện khi tham gia mua hàng trên các sàn thương mại điện. Đưa ra được lời khuyên mang tính tương đối để giúp khách hàng dễ dàng đưa ra lựa chọn hơn. | |
| Role by  Member |  |
| Lê Minh Kiệt: Thu thập dữ liệu – Thuật toán Logistic – Xử lí ngôn ngữ tự nhiên – Thiết kế web – Làm sạch dữ liệu  Võ Trần Yến Như: Thu thập dữ liệu – Làm sạch dữ liệu – Xử lý ngôn ngữ tự nhiên  Dương Ngọc Minh Huy: Thu thập dữ liệu – Thuật toán SVM – Xử lý ngôn ngữ tự nhiên  Trần Hữu Hậu: Thu thập dữ liệu – Thuật toán Random Forest – Xử lý ngôn ngữ tự nhiên  Nguyễn Thanh Nam: Thu thập dữ liệu – Làm sạch dữ liệu – Thiết kế web – Xử lý ngôn ngữ tự nhiên  Trần Trọng Nhân: Thu thập dữ liệu – Vector hóa dữ liệu – Thuật toán KNN – Xử lý ngôn ngữ tự nhiên  Nguyễn Thanh Nở: Thu thập dữ liệu – Thiết kế web – Xử lý ngôn ngữ tự nhiên | |

|  |  |
| --- | --- |
| Schedule  Summary |  |
|  | |
| Comment &  Assessment |  |
| <Comment and assessment **by the instructor.**> | |