1BÁO CÁO DỰ ÁN
MÔN HỌC: ADY201m - AI, DATA SCIENCE WITH PYTHON & SQL

REPORT 1: RESEARCH PROPOSAL & SYSTEM ARCHITECTURE

Chủ đề: PHÂN TÍCH CẢM XÚC TRÊN NỀN TẢNG GIAO ĐỒ ĂN
(Sentiment Analysis on Food Delivery Platform)

1. TỔNG QUAN & LÝ DO CHỌN ĐỀ TÀI

1.1. Bối cảnh

Trong ngành dịch vụ ăn uống , đánh giá trực tuyến (Online Review) đã trở thành yếu tố quyết định hành vi tiêu dùng. Nền tảng **Foody.vn** hiện sở hữu kho dữ liệu khổng lồ với hàng triệu bình luận trải dài khắp 63 tỉnh thành Việt Nam.

Tuy nhiên, các doanh nghiệp F&B thường chỉ nhìn vào điểm số định lượng (Rating 1-10) mà bỏ qua nội dung văn bản (Review Text). Điểm số đôi khi gây hiểu lầm: Một khách hàng chấm 8 điểm (khá tốt) nhưng vẫn để lại phàn nàn nghiêm trọng về vệ sinh, hoặc chấm 5 điểm chỉ vì giá cao dù món ăn ngon.

1.2. Vấn đề cần giải quyết

Dữ liệu trên Foody tồn tại dưới dạng **phi cấu trúc (unstructured data)** và có khối lượng quá lớn để đọc thủ công. Cần có một hệ thống tự động hóa để:

1. Thu thập dữ liệu quy mô lớn (Big Data Collection).
2. Khai phá dữ liệu văn bản để hiểu rõ "insight" khách hàng.
3. Phát hiện các xu hướng tiềm ẩn mà các báo cáo doanh thu không chỉ ra được.

2. MỤC TIÊU NGHIÊN CỨU

Đồ án này không chỉ dừng lại ở việc thu thập dữ liệu (Crawling), mà tập trung vào **Phân tích dữ liệu (Data Analytics)** nhằm trả lời các câu hỏi quản trị quan trọng:

1. **Vùng miền:** Khách hàng miền Bắc, Trung, Nam có tiêu chuẩn đánh giá và cảm xúc khác nhau như thế nào?
2. **Hành vi:** Yếu tố nào (Món ăn, Giá cả, Phục vụ) tác động mạnh nhất đến sự hài lòng hoặc giận dữ của khách hàng?
3. **Cảnh báo:** Dấu hiệu nào cho thấy một quán ăn đang đi xuống về chất lượng?

3. HỆ THỐNG GIẢ THUYẾT NGHIÊN CỨU (RESEARCH HYPOTHESES)

Đồ án đặt ra 5 giả thuyết khoa học để kiểm chứng thông qua mô hình phân tích dữ liệu:

- **H1: Sự thiên kiến vùng miền (Regional Sentiment Bias)**
    - *Giả thuyết:* Có sự chênh lệch về chuẩn mực đánh giá giữa các vùng. Cụ thể, khách hàng miền Bắc thường khắt khe hơn (tỷ lệ review tiêu cực/trung tính cao hơn) so với khách hàng miền Nam.
    - *Phương pháp:* So sánh phân phối điểm Rating trung bình giữa Hà Nội, Đà Nẵng và TP.HCM.
- **H2: Tác động của khía cạnh dịch vụ (Aspect Impact)**
    - *Giả thuyết: "Thái độ phục vụ"* có tác động đến cảm xúc tiêu cực mạnh hơn *"Chất lượng món ăn"*. Khách hàng có thể tha thứ cho món ăn không hợp khẩu vị, nhưng sẽ đánh giá cực thấp nếu nhân viên có thái độ tệ.
    - *Phương pháp:* Phân tích tương quan giữa các từ khóa (Keywords) và điểm số.
- **H3: Mối quan hệ Độ dài - Cảm xúc (Length-Intensity Relation)**
    - *Giả thuyết:* Độ dài bình luận tỷ lệ thuận với cường độ cảm xúc. Những review rất dài (>200 từ) thường rơi vào hai cực: Rất hài lòng (9-10đ) hoặc Rất bức xúc (0-3đ).
    - *Phương pháp:* Phân tích biểu đồ phân tán giữa Word Count và Rating.
- **H4: Sự đánh đổi giữa Quy mô và Chất lượng (Scale-Quality Trade-off)**
    - *Giả thuyết:* **"Chất lượng dịch vụ có xu hướng đi xuống khi quán trở nên quá đông khách."** Khi số lượng review tăng lên đột biến (hype), điểm rating trung bình sẽ giảm dần theo thời gian do quá tải vận hành.
    - *Phương pháp:* Phân tích xu hướng (Trend Analysis) theo thời gian thực của các quán Hot.
- **H5: Từ khóa tử huyệt (The Fatal Keywords)**
    - *Giả thuyết:* Tồn tại một bộ từ khóa "chết chóc" quyết định việc quán bị 1 sao. Các từ này thường liên quan đến vệ sinh và gian lận (ví dụ: *"gián", "tóc", "chửi", "lừa đảo"*).
    - *Phương pháp:* Trích xuất đặc trưng văn bản (Text Mining & TF-IDF) trên tập dữ liệu nhãn 1 sao.

III. KIẾN TRÚC HỆ THỐNG

Crawler → MinIO (raw) → ETL → PostgreSQL → Analysis → ML Models

3 Docker containers:

1. MinIO (Port 9000/9001)
2. PostgreSQL (Port 5432)
3. Python App (Port 8888)