NGHIÊN CỨU PHƯƠNG PHÁP CẢI THIỆN PHÂN CỤM DỮ LIỆU CÁC CỬA HÀNG TRÊN SÀN THƯƠNG MẠI ĐIỆN TỬ TIKI

Ứng dụng thuật toán khai thác mẫu thường xuyên và học máy

Nguyễn Văn Đạt - 2186400229

GVHD: TS. Bùi Danh Hường
Đại học Công nghệ Tp.HCM

Nội Dung

1. Tổng quan đề tài
2. Chuẩn bị dữ liệu
3. Khám phá dữ liệu
4. Phân cụm trước khi sử dụng khai thác mẫu thường xuyên
5. Phân cụm sau khi sử dụng khai thác mẫu thường xuyên
6. Kết luận

Tổng quan đề tài

Lý do chọn đề tài

- Tầm quan trọng của hiểu biết thị trường: Giúp đưa ra quyết định chiến lược hiệu quả.
- Phân tích phân cụm: Nhận diện nhóm cửa hàng có đặc điểm tương đồng.
- Ứng dụng thực tiễn:
  - Tối ưu hóa hoạt động kinh doanh.
  - Cải thiện dịch vụ khách hàng.
  - Nâng cao hiệu quả marketing.
- Ý nghĩa phát triển chiến lược:
  - Hỗ trợ doanh nghiệp nắm bắt xu hướng.
  - Phát triển chiến lược kinh doanh bền vững.

Mục tiêu đề tài

- Thu thập và xử lý dữ liệu.
- Phân cụm dữ liệu.
- Tối ưu hóa hiệu quả phân cụm.
- Phân tích kết quả phân cụm.
- Đề xuất chiến lược phù hợp.

Chuẩn bị dữ liệu

Thu thập và tiền xử lý dữ liệu

Thu thập dữ liệu:

- Dữ liệu thu thập từ sàn thương mại điện tử Tiki bằng phương pháp API Scraping.
- Dữ liệu ban đầu gồm: 73.000 mẫu gồm thông tin về các cửa hàng, thông tin về sản phẩm của cửa hàng và các đánh giá của khách hàng trên từng sản phẩm.
- Dữ liệu sau khi tiền xử lý dữ liệu.

Revenue	YearJoined	Followers	ChatResponse	RatQua	PosQua	NegQua
1326000	5	982	0.00	4.621877	0.787091	0.125943
25545495	4	1500	0.00	4.631574	0.706230	0.210805
10023000	6	11708	1.00	4.686931	0.845361	0
3148294	7	479	0.00	4.350326	0.752689	0.191492
6572000	6	565	1.00	4.560354	0.792613	0.158939
529000	6	137	0.00	4.648886	0.643643	0.149938
2261000	5	955	0.66	4.545801	0.786455	0.138842
0	4	73	0.00	4.190539	0.689961	0.043443
816900	8	181	0.00	4.168226	0.548946	0.186020
966000	8	3013	0.50	4.539589	0.770787	0.152678

Table: Bảng dữ liệu hoàn chỉnh.

Thu thập và tiền xử lý dữ liệu

Tiền xử lý dữ liệu:

- Xóa các dữ liệu trùng lặp.
- Đối với dữ liệu feedback(text):
  - Bóc tách dữ liệu từ định dạng JSON
  - Sử dụng thư viện underthesea để:
    - Chuẩn hóa dữ liệu tiếng Việt (Normalize)
    - Phân đoạn câu (Tokenize)
    - Phân tích cảm xúc đánh giá của khách hàng (Sentiment)

$\Longrightarrow$ *Positive*, *Negative*

Khám phá dữ liệu

Trực quan hóa dữ liệu

Xử lý Outliers

Chuẩn hóa dữ liệu

Phân tích độ tương quan

Phân cụm trước khi sử dụng khai thác mẫu thường xuyên

Loại bỏ nhiễu từ DBSCAN

Tìm số cụm tối ưu bằng Silhouette, AIC, BIC

Phân cụm K-Means

Phân cụm GMM (Gaussian Mixture Model)

Phân cụm Agglomerative

Phân cụm sau khi sử dụng khai thác mẫu thường xuyên

Rời rạc hóa dữ liệu K-Bins Discretizer

Áp dụng khai thác mẫu thường xuyên FP-Max

Phân cụm K-Means

Phân cụm GMM (Gaussian Mixture Model)

Phân cụm Agglomerative

Kết luận