ĐẠI HỌC KINH TẾ THÀNH PHỐ HỒ CHÍ MINH

UEH UNIVERSITY

Trường Công nghệ và Thiết kế

KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH

BÁO CÁO CUỐI KỲ

THUẬT TOÁN AFFINITY PROPAGATION

Môn học: Máy học

Giảng viên hướng dẫn: TS. Nguyễn An Tế

Mã lớp học phần: 24C1INF50904401

Nhóm sinh viên thực hiện:

- Nguyễn Phúc Minh Trâm - 31221024796
- Trần Vọng Triển - 31221021725
- Trầm Thái Tú - 31221022394
- Nguyễn Thành Vinh - 31221025662
- Nguyễn Vân Phi Yến - 31221021785

Hồ Chí Minh, ngày 30 tháng 11 năm 2024

LỜI CẢM ƠN

Để hoàn thành bài tiểu luận này, nhóm chúng em xin gửi lời tri ân sâu sắc đến Thầy TS. Nguyễn An Tế, giảng viên hướng dẫn môn học Machine Learning trực thuộc khoa Công nghệ Thông tin Kinh doanh.

Nhóm chúng em chân thành cảm ơn Thầy vì sự tận tâm trong giảng dạy, cũng như những kiến thức quý báu mà Thầy đã truyền đạt trong môn học Machine Learning. Không chỉ giới hạn trong khuôn khổ bài giảng, Thầy còn khuyến khích chúng em tìm hiểu và nghiên cứu thêm các thuật toán mới như Affinity Propagation, mở rộng góc nhìn và khả năng ứng dụng vào thực tiễn.

Trong quá trình thực hiện bài tiểu luận, nhóm chúng em đã cố gắng vận dụng những kiến thức đã được học và nghiên cứu thêm nhiều tài liệu để hoàn thiện bài tiểu luận này. Tuy nhiên, do kiến thức còn hạn chế và thiếu kinh nghiệm thực tiễn nên nội dung bài tiểu luận khó tránh khỏi những thiếu sót. Chúng em rất mong nhận được những ý kiến đóng góp quý báu từ Thầy để nhóm có thể hoàn thiện bài làm hơn.

Nhóm chúng em xin chân thành cảm ơn Thầy vì sự đồng hành và hỗ trợ trong suốt quá trình học tập và nghiên cứu.

Trân trọng,
Nhóm thực hiện

PHÂN CÔNG CÔNG VIỆC

Nhiệm vụ	Thành viên	Mức độ hoàn thành
- Tìm hiểu lý thuyết thuật toán AP - Phụ trách nội dung: Chương 2.1 và 2.2 - Phụ trách làm slide thuyết trình	Nguyễn Phúc Minh Trâm	100%
- Thu thập bộ dữ liệu - Biểu diễn trực quan, nhận xét sau khi phân cụm. - Phụ trách nội dung: Chương 5	Trầm Thái Tú *(Nhóm trưởng)*	100%
- Biểu diễn trực quan, nhận xét sau khi phân cụm. - Phụ trách nội dung: Chương 6, 2.3 và 2.4	Trần Vọng Triển	100%
- Tiền xử lý dữ liệu - Phụ trách nội dung: Chương 3, 4	Nguyễn Thành Vinh	100%
- Tìm hiểu lý thuyết thuật toán AP - Phụ trách nội dung: Chương 1 - Phụ trách làm slide thuyết trình	Nguyễn Vân Phi Yến	100%

MỤC LỤC

PHỤ LỤC HÌNH ẢNH

CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI

1.1. Giới thiệu đề tài

Trong thời đại cạnh tranh gay gắt giữa các thương hiệu, đặc biệt là các ngành kinh doanh dịch vụ và ẩm thực, việc thấu hiểu và xây dựng mối quan hệ lâu dài với khách hàng là một trong những yếu tố then chốt để doanh nghiệp phát triển bền vững và đem lại lợi thế hơn so với các đối thủ cạnh tranh.

Một trong những phương pháp phổ biến để phân khúc khách hàng là *Mô hình RFM (Recency, Frequency, Monetary)* - mô hình phân tích và phân khúc khách hàng theo các đặc điểm hành vi tiêu dùng dựa trên các dữ liệu giao dịch trong lịch sử và ba yếu tố: *gần đây nhất (Recency), tần suất (Frequency)* và *giá trị chi tiêu (Monetary)*.

Nhận thấy được tiềm năng ứng dụng, nhóm đã tiến hành phân tích trên bộ dữ liệu được trích xuất từ bộ dữ liệu đề thi của cuộc thi RMIT Business Analytics Champion (RBAC) 2023 do nhà tài trợ ra đề Pizza Hut Việt Nam cung cấp.

1.2. Mục tiêu nghiên cứu

Mục tiêu của đề tài nhằm ứng dụng thuật toán Affinity Propagation để phân khúc khách hàng dựa trên bộ dữ liệu từ Pizza Hut. Bài làm tập trung vào tiền xử lý dữ liệu và áp dụng thuật toán Affinity Propagation nhằm xác định số lượng cụm và cải thiện độ chính xác trong phân khúc khách hàng. Kết quả phân cụm nhằm rút ra được đặc điểm chung của từng cụm, từ đó đề xuất các chiến lược kinh doanh để nâng cao trải nghiệm của khách hàng.

1.3. Phương pháp nghiên cứu

- Phân tích tổng quan các biến
- Tiền xử lý dữ liệu
- Sử dụng thuật toán Affinity Propagation để phân cụm khách hàng

1.4. Ngôn ngữ sử dụng

- Ngôn ngữ lập trình Python

CHƯƠNG 2. TỔNG QUAN VỀ THUẬT TOÁN AFFINITY PROPAGATION

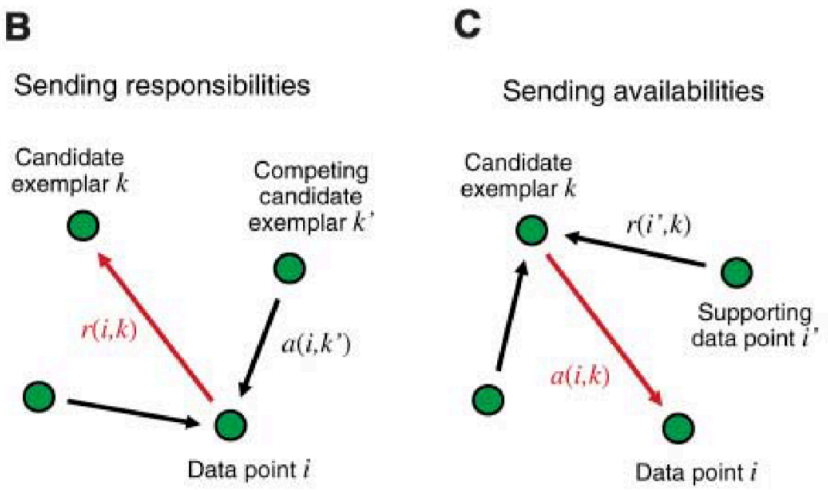
2.1. Giới Thiệu Về Thuật Toán Affinity Propagation

Thuật toán Affinity Propagation là một thuật toán phân cụm hiệu quả được đề xuất và phát triển bởi Frey và Dueck trong bài báo với tiêu đề *"Clustering by Passing Messages between Data Points"* (2007).

Affinity Propagation không yêu cầu chỉ định số cụm mà thuật toán sẽ tự động tìm ra các cụm dựa trên cơ chế "truyền thông điệp" (message-passing) giữa các điểm dữ liệu. Thuật toán này bắt đầu bằng việc sử dụng các độ đo tương đồng giữa các cặp điểm dữ liệu và liên tục trao đổi thông điệp dạng số thực giữa các điểm dữ liệu cho đến khi phát hiện được điểm đại diện của cụm (exemplar) và tự động phân các điểm dữ liệu vào các cụm tương ứng dựa trên sự tương đồng.

Thuật toán sử dụng ba ma trận chính để thực hiện việc phân cụm:

1. Ma trận tương đồng (Similarity Matrix - S): Thể hiện mức độ giống nhau giữa các điểm dữ liệu, được tính toán dựa trên khoảng cách Euclidean bình phương âm.
2. Ma trận trách nhiệm (Responsibility Matrix - R): Biểu diễn mức độ "phù hợp" của một điểm dữ liệu khi trở thành exemplar cho một điểm khác.
3. Ma trận khả dụng (Availability Matrix - A): Phản ánh khả năng của một điểm dữ liệu được chọn làm exemplar, dựa trên sự cạnh tranh giữa các điểm.



Hình 2.1. Thuật toán Affinity Propagation (Frey, B. J., & Dueck, D. (2007))

Quá trình truyền thông điệp được lặp lại liên tục giữa các điểm dữ liệu cho đến khi các ma trận hội tụ, tức là khi các giá trị trong ma trận ổn định. Kết quả cuối cùng của Affinity Propagation là danh sách các exemplar và việc gán cụm tương ứng, giúp phân chia dữ liệu thành các nhóm phù hợp.

2.2. Quy Trình Hoạt Động Của Thuật Toán

2.2.1. Các định nghĩa

2.2.1.1. Preference

Preference là tham số đại diện cho mức độ "mong muốn" của một điểm dữ liệu để trở thành một điểm đại diện (exemplar). Đồng thời nó cũng là giá trị ở đường chéo chính của ma trận tương đồng.

Preference có thể được coi như "ngưỡng" để thuật toán quyết định một điểm có thể là đại diện của cụm hay không. Giá trị của preference càng cao, số lượng cụm càng lớn và ngược lại, giá trị càng thấp sẽ dẫn đến ít cụm hơn.

Đây là *siêu tham số* của thuật toán Affinity Propagation, người dùng có thể điều chỉnh preference để tạo ra số cụm mong muốn.

2.2.1.2. Ma trận tương đồng

Để xác định mức độ tương đồng của các điểm dữ liệu, thuật toán Affinity Propagation tính toán "điểm tương đồng" dựa trên các đặc điểm của chúng. Ma trận điểm tương đồng của các cặp điểm dữ liệu được gọi là "ma trận tương đồng (S)". "Điểm tương đồng" được tính bằng cách sử dụng khoảng cách bình phương âm giữa các điểm dữ liệu. Tức là, lấy khoảng cách giữa các điểm (có thể là khoảng cách Euclid, cosine similarity…), bình phương khoảng cách đó, sau đó làm cho kết quả âm.

2.2.1.3. Ma trận trách nhiệm

Ma trận được sử dụng biểu diễn tính phù hợp của điểm dữ liệu để làm đại diện cụm cho điểm dữ liệu khác. $R(i, k)$ là giá trị cho biết điểm $\boldsymbol{k}$ phù hợp như thế nào để làm đại diện cụm cho điểm dữ liệu $\boldsymbol{i}$, so với các đại diện cụm ứng viên khác cho điểm $\boldsymbol{i}$.

$$R(i,\ k)\ =\ S(i,\ k)\ -\ max\{A(i,\ k')\ +\ S(i,\ k')\}\ (với\ k\ <>\ k')$$

2.2.1.4. Ma trận khả dụng

Ma trận khả dụng được sử dụng để thể hiện tính "khả dụng" của mỗi điểm dữ liệu để làm đại diện cho các điểm dữ liệu khác. ***A(i, k)*** biểu diễn mức độ "phù hợp" khi điểm ***i*** chọn điểm ***k*** làm đại diện cụm.

$$a(i,\ k) \leftarrow min[0,\ r(k,\ k)\ +\ \sum max\{0, r(i',\ k)\}]\ \forall i, k$$

$$a(k, k) \leftarrow \sum max\{0, r(i', k)\}\ \forall i\ \neq\ k$$

2.2.2. Quy trình thuật toán

Khởi tạo: Tính ma trận S

Bước 1: Khởi tạo ma trận **R** và **A** bằng ma trận số không có kích thước ***n x n*** (với n là số lượng điểm dữ liệu)

Bước 2: Cập nhật ma trận R

$$R(i,\ k)\ =\ S(i,\ k)\ -\ max\{A(i,\ k')\ +\ S(i,\ k')\}\ (với\ k \neq k')$$

Bước 3: Cập nhật ma trận A

- Với i ≠ k:

$$A(i,\ k)\ =\ min\{0,\ R(k,\ k)\ +\ \sum_{i' \notin \{i,\ k\}} max\{0,\ R(i',\ k)\}\}$$

- Với i = k:

$$A(k,\ k)\ =\ \sum_{i' \neq k} max(0,\ R(i',\ k))$$

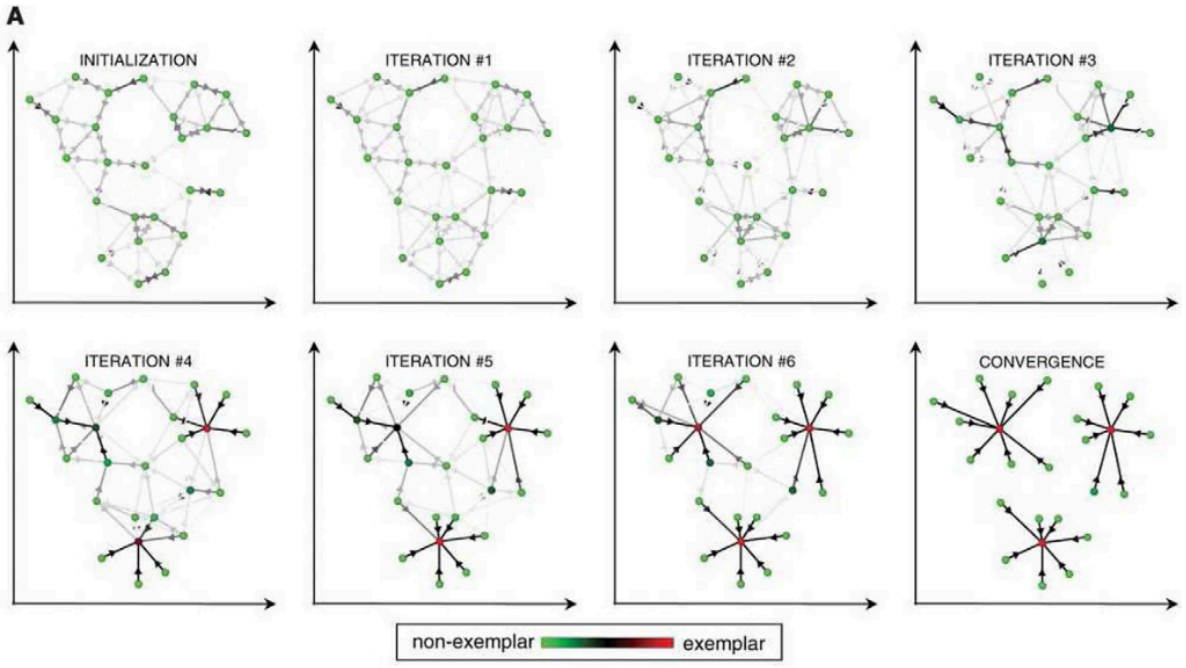
Bước 4: Tính tổng hợp

Tính tổng $T\ =\ R(i,\ k)\ +\ A(i,\ k)$:

- Nếu $T\ >\ 0$: k được chọn làm đại diện của i
- Nếu $T\ \leq\ 0$: k không được chọn làm đại diện của i

Bước 5: Lặp lại

Lặp lại **bước 2** đến **4** cho đến khi các ma trận **không thay đổi** qua số lần lặp được xác định trước.



Hình 2.2. Lược đồ mô tả quy trình hoạt động của thuật toán gom cụm Affinity Propagation (Frey, B. J., & Dueck, D. (2007))

2.3. Ưu và Nhược điểm của Thuật toán Affinity Propagation

2.3.1. Ưu điểm

Trong nghiên cứu của mình, Brendan J. Frey và Delbert Dueck (2007) đã chứng minh hiệu quả của thuật toán Affinity Propagation (AP) qua nhiều bài toán phân cụm. Cụ thể, trong việc phân cụm các putative exons, tức các đoạn DNA được giả định là gen, họ đã áp dụng AP trên ma trận tương đồng thưa (sparse similarity matrix) được trích xuất từ dữ liệu microarray. Trong quá trình so sánh, họ đối chiếu AP với phương pháp phân cụm K-centers, vốn yêu cầu xác định số lượng cụm trước khi thực hiện. Kết quả cho thấy AP không chỉ nhận diện các cụm gen chính xác hơn mà còn thực hiện nhanh hơn đáng kể. Khi so sánh hiệu quả của AP với K-centers clustering và phân cụm phân cấp (Hierarchical Agglomerative Clustering), Frey và Dueck chỉ ra rằng AP đạt được TP rates cao hơn, trong khi FP thấp hơn, điều này rất quan trọng trong các ứng dụng sinh học. Bên cạnh đó, AP có khả năng xử lý các tình huống dữ liệu không đối xứng (asymmetric, tức $s(i,k) \neq s(k,i)$) hoặc không thỏa mãn bất đẳng thức tam giác (tức $s(i,k) < s(i,j) + s(j,k)$).

Như vậy, việc áp dụng Affinity Propagation trong phân cụm mang lại một số ưu điểm nổi bật, bao gồm:

- Tự động phát hiện đại diện cụm và số lượng cụm: AP tự động xác định đại diện cụm và số lượng cụm mà không cần chỉ định trước, phù hợp với các tập dữ liệu có số lượng cụm không xác định hoặc thay đổi. Điều này rất hữu ích trong các bài toán phân cụm với dữ liệu phức tạp và quy mô lớn.
- Xử lý các dạng dữ liệu phức tạp: AP có thể làm việc với các dữ liệu không phải là đại lượng đo lường (nonmetric), dữ liệu bất đối xứng (asymmetric), cũng như những trường hợp không tuân theo bất đẳng thức tam giác.

2.3.2. Nhược điểm

Độ phức tạp tính toán: Thuật toán **Affinity Propagation** có độ phức tạp tính toán $O(N^2)$, trong đó N là số lượng điểm dữ liệu. Điều này có nghĩa rằng khi số lượng điểm dữ liệu tăng lên, thời gian tính toán sẽ tăng theo bình phương. Đối với các tập dữ liệu lớn với hàng nghìn hoặc hàng triệu điểm, thuật toán trở nên cực kỳ chậm và không hiệu quả.

Nhạy cảm với tham số đầu vào: Một điểm yếu quan trọng của AP là tính nhạy cảm với giá trị preference - một tham số quan trọng được sử dụng để điều khiển quá trình phân cụm. Việc lựa chọn preference không phù hợp có thể dẫn đến kết quả phân cụm không chính xác. Hiện tại chưa có phương pháp tiêu chuẩn để xác định giá trị preference tối ưu cho mọi bộ dữ liệu.

Không ổn định và khó dự đoán: Thuật toán Affinity Propagation không đảm bảo tìm được nghiệm toàn cục tối ưu. Điều này có nghĩa là các lần chạy khác nhau với cùng một bộ dữ liệu có thể cho ra các kết quả phân cụm khác nhau. Tính không ổn định này làm giảm độ tin cậy của thuật toán trong các ứng dụng yêu cầu độ ổn định cao.

2.4. Một Vài Ứng Dụng Của Thuật Toán Affinity Propagation

Affinity Propagation (cùng các thuật toán mở rộng và biến thể từ nó) đã được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, bao gồm khoa học kỹ thuật, tin sinh học, kinh tế xã hội, thị giác máy tính,... Một số lĩnh vực và nghiên cứu chứng minh hiệu suất và tính linh hoạt của thuật toán này:

Tin sinh học: Affinity Propagation hiệu quả với các bài toán tin sinh học, đặc biệt là khi làm việc với dữ liệu sinh học như dữ liệu microarray, tiêu biểu là các bài toán gene-expression (Leone, M., et al., 2007) và bài toán phân loại không giám sát tế bào thần kinh (Santana, R., et al., 2013). Các nghiên cứu cho thấy Affinity Propagation hoạt động hiệu quả hơn so với các thuật toán khác như Ward's method.

Thị giác máy tính: Khi các đặc trưng hình ảnh hoặc các đặc trưng trong các tác vụ thị giác không thể được đo lường theo kỹ thuật metric-based, Affinity Propagation có thể sử dụng phương pháp non-metric để nhóm các đối tượng lại với nhau.
Một số nghiên cứu cho thấy Affinity Propagation có tốc độ xử lý nhanh hơn rất nhiều so với các phương pháp phân cụm khác khi xử lý dữ liệu lớn, như trong các bài báo: Non-metric affinity propagation for unsupervised image categorization (Dueck, D., & Frey, B. J., 2007); Band selection for hyperspectral imagery using affinity propagation (Qian, Y., et al., 2010).

Khoa học máy tính: Affinity Propagation là một thuật toán phân cụm mạnh mẽ trong khoa học máy tính, với nhiều ứng dụng trong các lĩnh vực như tối ưu hóa, học máy và mạng cảm biến không dây. Ví dụ, trong nghiên cứu của Wang, J., et al. (2019), một phương pháp phân cụm dựa trên Affinity Propagation đã được áp dụng cho các mạng cảm biến không dây; trong nghiên cứu của Liu, Y., et al. (2020), thuật toán này đã được sử dụng trong tối ưu bầy đàn (PSO) để giải quyết các bài toán dynamic optimization.

Kinh tế xã hội: Affinity Propagation có thể giúp xác định các nhóm đối tượng tương đồng trong các bài toán kinh tế xã hội, từ đó tạo ra các chiến lược phân bổ tài nguyên hợp lý và hiệu quả hơn, hỗ trợ các tổ chức quyết định chính sách dựa trên dữ liệu. Ví dụ, trong nghiên cứu của Asriny, N. I., et al. (2021), thuật toán phân cụm K-Affinity propagation (một biến thể của thuật toán AP kết hợp với K-means**)** được áp dụng để phân loại những người lao động bán thời gian sử dụng internet, trong nghiên cứu của A'yuni, T. Q., et al. (2023), phương pháp này đã được sử dụng để phân nhóm các doanh nghiệp nhỏ và vừa (MSMEs) dựa trên mức độ ưu tiên phân phối hỗ trợ kinh doanh.

Khoa học kỹ thuật: Trong các lĩnh vực khoa học kỹ thuật như định vị không gian, xử lý tín hiệu và mã hóa thông tin, Affinity Propagation có thể cải thiện hiệu quả và độ

chính xác trong các bài toán cần xử lý nhiều hệ thống phức tạp. Ví dụ, trong nghiên cứu của Karegar, P. A. (2018), phương pháp phân cụm Affinity Propagation đã được sử dụng trong định vị trong nhà bằng fingerprint wireless.

Xử lý ngôn ngữ tự nhiên: Affinity Propagation cung cấp một phương pháp phân cụm hiệu quả trong các bài toán như Segmentation, Classification và Semantic Analysis. Những cải tiến và mở rộng của Affinity Propagation giúp tăng cường khả năng ứng dụng trong các tác vụ NLP phức tạp, đồng thời duy trì tính chính xác và giảm thiểu độ phức tạp tính toán, như trong nghiên cứu của Kazantseva, A., & Szpakowicz, S. (2011), nghiên cứu này sử dụng Affinity Propagation để phân đoạn văn bản (segmentation).

CHƯƠNG 3. TỔNG QUAN BỘ DỮ LIỆU

3.1. Sơ lược bộ dữ liệu

Bộ dữ liệu sử dụng trong đồ án này được trích xuất từ bộ dữ liệu đề thi của cuộc thi *RMIT Business Analytics Champion (RBAC) 2023* do nhà tài trợ ra đề Pizza Hut Việt Nam cung cấp. Đây là một bộ dữ liệu thực tế đã được điều chỉnh cho phù hợp, nhằm phản ánh các tình huống kinh doanh thực tế trong ngành thực phẩm và đồ uống tại thị trường Việt Nam.

Trong phạm vi đồ án này, nhóm chỉ sử dụng một phần của bộ dữ liệu với mục đích phục vụ nghiên cứu học thuật. Nhóm cam kết chỉ sử dụng dữ liệu cho mục đích học thuật trong phạm vi thực hiện đồ án, tôn trọng các quy định về bản quyền và bảo mật thông tin từ ban tổ chức RBAC 2023.

3.2. Mô tả thuộc tính

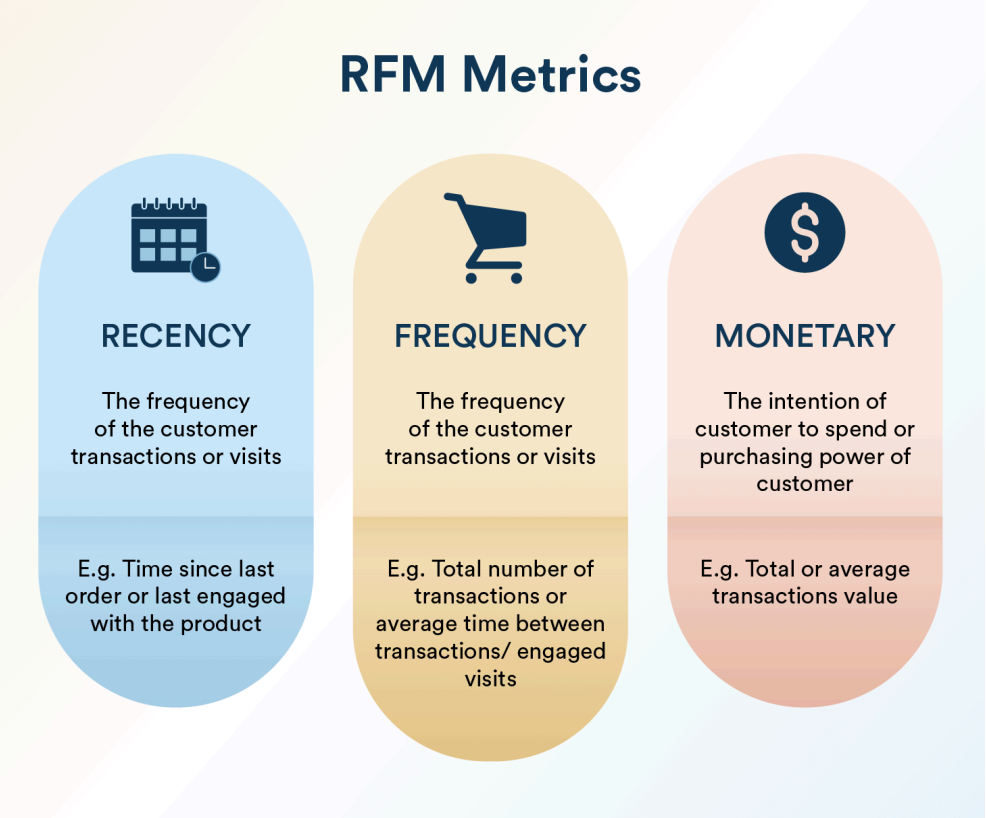
Dữ liệu bao gồm 1,500 dòng, không có giá trị khuyết, với 9 thuộc tính được mô tả chi tiết như sau:

- *BillID (int64):* Mã định danh duy nhất của mỗi đơn hàng.
- *Channel (object):* Kênh mà đơn hàng được tiêu thụ, bao gồm các giá trị: Dine In, Take Away, Delivery.
- *OrderFrom (object):* Kênh khách hàng đặt đơn hàng, bao gồm các giá trị: APP, CALL CENTER, STORE, hoặc WEBSITE.
- *TransactionDate (datetime64[ns]):* Ngày giao dịch của đơn hàng, định dạng theo kiểu dữ liệu ngày giờ.
- *SalesAmount (int64):* Số tiền khách hàng đã chi trả cho đơn hàng (đơn vị: VND).
- *CustomerID (int64):* Mã định danh duy nhất của mỗi khách hàng.
- *CustomerGender (object):* Giới tính của khách hàng, với các giá trị: Female, Male, hoặc Unknown.
- *VoucherStatus (object):* Đơn hàng có áp dụng voucher hay không, với hai giá trị: Yes (có áp dụng) và No (không áp dụng).

- *Province (object):* Tỉnh hoặc thành phố nơi thực hiện đơn hàng. Thuộc tính này chỉ có giá trị 'Ho Chi Minh City', do dữ liệu được trích xuất từ các chi nhánh tại Thành phố Hồ Chí Minh.

3.3. Mô hình RFM

RFM là một mô hình phân tích phổ biến trong Customer Relationship Management (CRM) dùng để phân nhóm khách hàng theo các đặc điểm về hành vi tiêu dùng dựa trên dữ liệu giao dịch trong quá khứ. Trong đó RFM là viết tắt của Recency, Frequency, Monetary.



Hình 3.1. RFM Metrics (Makhija, P. (2018))

Recency (lần mua hàng gần nhất): đo lường khoảng thời gian kể từ lần mua hàng gần nhất của khách hàng cho đến ngày tiến hành phân tích. Khoảng thời gian này càng lớn thì khả năng cao khách hàng rời bỏ dịch vụ, sản phẩm. Ngược lại, khoảng thời gian này càng nhỏ thì khả năng họ quay lại mua sắm nhiều hơn.

R = ngày phân tích - ngày giao dịch gần nhất

Frequency (tần suất mua hàng): đo lường tổng số lần khách hàng đã mua hàng trong một khoảng thời gian cụ thể. Khách hàng mua hàng thường xuyên có khả năng trở thành khách hàng trung thành.

F = tổng số lần mua hàng

Monetary (tổng số tiền khách hàng chi trả): đo lường tổng số tiền mà khách hàng đã chi tiêu trong khoảng thời gian cụ thể. Những khách hàng mua nhiều lần, có sự tin tưởng nhất định vào thương hiệu mới có thể mua tiếp những đơn hàng có giá trị cao.

M = Cộng gộp số tiền mà khách hàng đã chi trả cho sản phẩm, dịch vụ

CHƯƠNG 4. TIỀN XỬ LÝ DỮ LIỆU

4.1. Quan sát bộ dữ liệu

Quan sát các cột dữ liệu trong bộ dữ liệu

	BillID	Channel	OrderFrom	TransactionDate	SalesAmount	CustomerID	CustomerGender	VoucherStatus	Province
0	925181	Delivery	CALL CENTER	2022-11-07	376676	1599602	Female	No	Ho Chi Minh City
1	875964	Take Away	WEBSITE	2022-10-13	109927	1761308	Unknown	No	Ho Chi Minh City
2	1021982	Delivery	WEBSITE	2022-12-29	2313260	1692638	Unknown	No	Ho Chi Minh City
3	1007513	Delivery	CALL CENTER	2022-12-22	1871407	1605462	Male	No	Ho Chi Minh City
4	1015400	Dine In	STORE	2022-12-25	689747	70040	Female	No	Ho Chi Minh City
...	...	...	...	...	...	...	...	...	...
1495	859105	Take Away	STORE	2022-10-03	89538	676327	Unknown	No	Ho Chi Minh City
1496	925602	Take Away	APP	2022-11-07	156713	1667719	Female	No	Ho Chi Minh City
1497	917288	Delivery	CALL CENTER	2022-11-03	405818	551659	Unknown	No	Ho Chi Minh City
1498	980535	Delivery	WEBSITE	2022-12-08	303575	920724	Unknown	No	Ho Chi Minh City
1499	1018475	Take Away	STORE	2022-12-26	302712	986592	Male	No	Ho Chi Minh City

1500 rows × 9 columns

Hình 4.1. Dữ liệu Pizza Hut

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   BillID           1500 non-null   int64
 1   Channel          1500 non-null   object
 2   OrderFrom        1500 non-null   object
 3   TransactionDate  1500 non-null   datetime64[ns]
 4   SalesAmount      1500 non-null   int64
 5   CustomerID       1500 non-null   int64
 6   CustomerGender   1500 non-null   object
 7   VoucherStatus    1500 non-null   object
 8   Province         1500 non-null   object
dtypes: datetime64[ns](1), int64(3), object(5)
memory usage: 105.6+ KB
```

Hình 4.2. Metadata

Bộ dữ liệu PizzaHut có 1500 dòng dữ liệu, không chứa dữ liệu Null. Với 5 biến phân loại, 3 biến số (bao gồm cả BillID và CustomerID) và 1 biến thời gian.

	BillID	TransactionDate	SalesAmount	CustomerID
count	1.500000e+03	1500	1.500000e+03	1.500000e+03
mean	9.421668e+05	2022-11-16 09:46:33.600000256	3.444380e+05	1.061367e+06
min	8.529200e+05	2022-10-01 00:00:00	2.094900e+04	1.120000e+02
25%	8.980422e+05	2022-10-23 00:00:00	1.952278e+05	5.149332e+05
50%	9.437440e+05	2022-11-18 00:00:00	2.897025e+05	1.049294e+06
75%	9.872235e+05	2022-12-11 00:00:00	4.042420e+05	1.600427e+06
max	1.028975e+06	2022-12-31 00:00:00	6.022211e+06	2.173664e+06
std	5.165086e+04	NaN	3.263135e+05	6.271797e+05

Hình 4.3. Mô tả các thuộc tính numeric

Dữ liệu này được tổng hợp trong khoảng thời gian Quý 4 năm 2022, với đơn hàng nhỏ nhất có giá trị 20.949 VNĐ và lớn nhất vào khoảng 6.022.211 VNĐ. Phần lớn giá trị đơn hàng rơi vào khoảng 200.000 - 400.000 VNĐ.

Để phục vụ cho việc phân cụm khách hàng. Dựa vào các giá trị RFM, chúng ta có thể phân loại khách hàng thành các nhóm khách hàng khác nhau nhằm phục vụ cho các chiến dịch marketing, chăm sóc khách hàng…

Đoạn chương trình tính giá trị RFM cho từng khách hàng:

```
df = df.loc[:, ['CustomerID', 'SalesAmount', 'TransactionDate']]
# Lấy ngày giao dịch cuối cùng của từng khách hàng
df['Last Order'] = df.groupby('CustomerID')['TransactionDate'].transform('max')
# Chuyển đổi định dạng ngày
df['TransactionDate'] = pd.to_datetime(df['TransactionDate'])
df['Last Order'] = pd.to_datetime(df['Last Order'])
# Tạo ngày báo cáo cố định
df['Report Date'] = '2023-1-1'
df['Report Date'] = pd.to_datetime(df['Report Date'])
# Tính recency (khoảng thời gian từ lần mua cuối đến ngày báo cáo)
df['Recency'] = (df['Report Date'] - df['Last Order']).dt.days
# Tính frequency (số lần giao dịch của từng khách hàng)
df['Frequency'] = df.groupby('CustomerID')['TransactionDate'].transform('count')
# Tính tổng GMV cho từng khách hàng
df['Monetary'] = df.groupby('CustomerID')['SalesAmount'].transform('sum')
✓ 0.1s
```

	CustomerID	Recency	Frequency	Monetary
0	1599602	55	1	376676
1	1761308	80	1	109927
2	1692638	3	1	2313260
3	1605462	10	1	1871407
4	70040	7	1	689747
...	...	...	...	...
1495	676327	90	1	89538
1496	1667719	55	1	156713
1497	551659	59	1	405818
1498	920724	24	1	303575
1499	986592	6	1	302712

1477 rows × 4 columns

Hình 4.4. Bảng dữ liệu sau khi tính các giá trị RFM

```
<class 'pandas.core.frame.DataFrame'>
Index: 1477 entries, 0 to 1499
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   CustomerID  1477 non-null   int64
 1   Recency     1477 non-null   int64
 2   Frequency   1477 non-null   int64
 3   Monetary    1477 non-null   int64
dtypes: int64(4)
memory usage: 57.7 KB
```

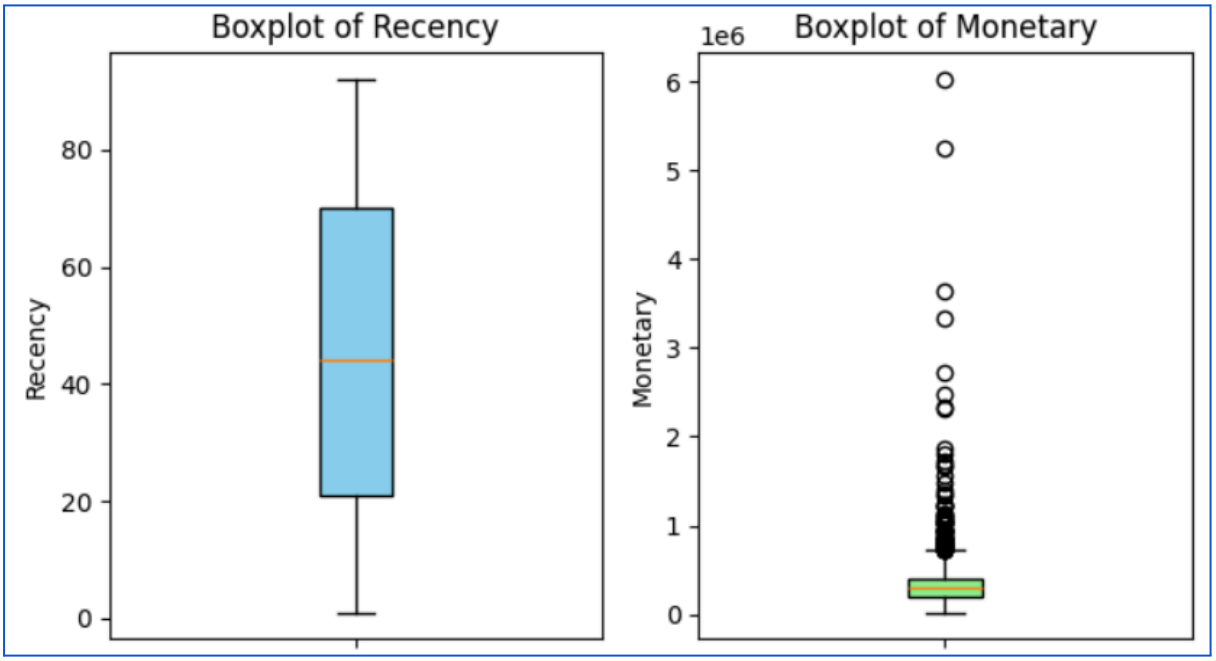
Hình 4.5. Metadata bảng dữ liệu sau khi tính RFM

Trên đây là tính toán các giá trị Recency, Frequency, Monetary từ bộ dữ liệu các Bill bán hàng của PizzaHut. Ta thấy có 1477 khách hàng từ 1500 hóa đơn, điều này đồng nghĩa với việc khách hàng rất ít khi quay lại trong khoảng thời gian Quý 4/2022 tại chi nhánh TPHCM này.

	CustomerID	Recency	Frequency	Monetary
count	1.477000e+03	1477.000000	1477.000000	1.477000e+03
mean	1.063928e+06	45.345972	1.015572	3.498017e+05
std	6.268409e+05	27.459556	0.129209	3.294608e+05
min	1.120000e+02	1.000000	1.000000	2.094900e+04
25%	5.186930e+05	21.000000	1.000000	2.011600e+05
50%	1.057924e+06	44.000000	1.000000	2.922320e+05
75%	1.600998e+06	70.000000	1.000000	4.084290e+05
max	2.173664e+06	92.000000	3.000000	6.022211e+06

Hình 4.5. Mô tả các thuộc tính RFM

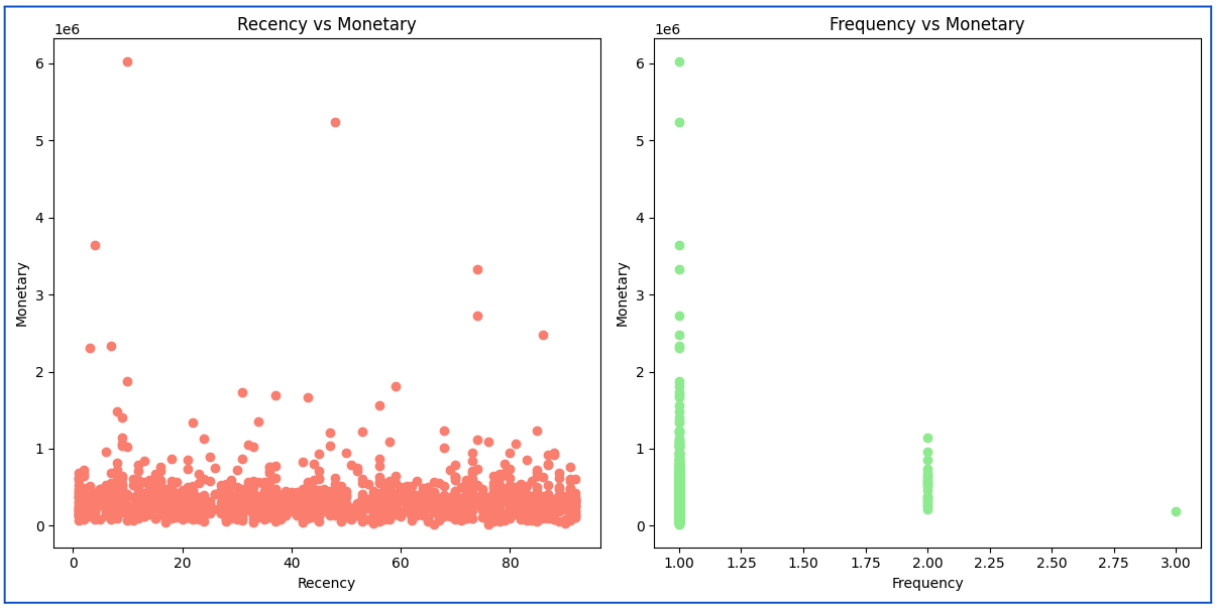
Như đã lập luận ở trên, khách hàng hiếm khi quay lại trong cùng Quý 4/2022 và chỉ có trường hợp quay lại nhiều nhất là 3 lần. Thời gian gần nhất và xa nhất một khách hàng lần cuối mua hàng tại quán lần lượt là 1 và 92 ngày. Phần lớn khách hàng đến quán lần cuối cùng cách ngày kiểm duyệt là 21 - 70 ngày. Tổng số tiền một khách hàng bỏ ra cho Quý 4/2022 tại chi nhánh rơi vào khoảng 200.000 - 400.000 VNĐ.



Hình 4.6. Biểu đồ Boxplot của Recency và Monetary

Dựa vào biểu đồ Boxplot trên, ta có thể thấy sự chênh lệch giữa tổng số tiền bỏ ra trong Quý 4/2022 của các khách hàng có sự chênh lệch lớn với các outlier rất rõ ở biến

Monetary. Điều này cho thấy, có các khách hàng thật sự tiêu dùng một khoảng rất lớn so với phần còn lại và cần xem xét quan tâm đến các đối tượng này.



Hình 4.7. Biểu đồ Scatterplot thể hiện mối tương quan giữa các biến RM, FM

Biểu đồ Scatter bên trên cho thấy có sự tương quan không mạnh giữa RM nhưng FM cho thấy sự tương quan ngược. Với các khách hàng tiêu dùng nhiều tiền ở chi nhánh lại có số lần quay lại quán rất ít (chỉ với 1 lần), tức các khách hàng tiêu nhiều có xu hướng mua hàng 1 lần với hóa đơn lớn hơn thay vì quay lại nhiều lần.

4.2. Chuẩn hóa dữ liệu

Vì dữ liệu hiện tại có khoảng dao động dữ liệu rất lớn và chênh lệch giữa các thuộc tính. Do đó, để thuật toán Affinity Propagation hoạt động tối ưu hơn, ta tiến hành Scale dữ liệu bằng phương pháp StandardScaler vì bộ dữ liệu có kích thước lớn, có thể xấp xỉ phân phối chuẩn nên phù hợp sử dụng phương pháp này.

```
    scaler = StandardScaler()
    scaled_features = scaler.fit_transform(df)
    scaled_features
```

```
array([[ 0.36184278, -0.17110587,  0.07422758],
       [ 1.27282663, -0.17110587, -0.73697626],
       [-1.53300362, -0.17110587,  5.9635257 ],
       ...,
       [ 0.5076002 , -0.17110587,  0.1628506 ],
       [-0.76777719, -0.17110587, -0.14807808],
       [-1.42368556, -0.17110587, -0.15070252]])
```

Hình 4.8. Dữ liệu sau khi đã được chuẩn hóa

CHƯƠNG 5. ÁP DỤNG GIẢI THUẬT

5.1. Phương Pháp Phân Cụm

5.1.1. Các Bước Tiền Xử Lý

Tính toán các giá trị RFM:

- *R = ngày phân tích - ngày giao dịch gần nhất*
- *F = tổng số lần mua hàng*
- *M = Cộng gộp số tiền mà khách hàng đã chi trả cho sản phẩm, dịch vụ*

5.1.2. Chuẩn Hóa Dữ Liệu

Sử dụng StandardScaler để chuẩn hóa các giá trị RFM

5.1.3. Sử dụng thuật toán Affinity Propagation để phân cụm

Thuật toán Affinity Propagation được áp dụng nhằm tối ưu hóa việc xác định các nhóm khách hàng dựa trên hành vi. Phương pháp được triển khai như sau:

- *Khoảng giá trị preference:* Tiến hành kiểm tra một số giá trị preference trong khoảng từ -250 đến -10. Với mục tiêu chọn số cụm tối đa là 5, nhóm lựa chọn giá trị preference thấp, điều này giúp làm giảm số lượng điểm được chọn làm exemplar. Khi chỉ một số ít điểm trở thành đại diện cụm, các điểm dữ liệu khác sẽ gộp chung vào các cụm lớn hơn. Điều này giúp kiểm soát số cụm, đảm bảo phù hợp với yêu cầu phân loại.
- *Đánh giá chất lượng phân cụm:* Với mỗi giá trị preference, thuật toán phân cụm được thực thi. Chỉ các kết quả có số cụm từ 2 đến dưới 100 được chấp nhận để đảm bảo tính hợp lý. Sau đó, Silhouette Score được sử dụng để đánh giá độ tương đồng giữa các điểm trong cùng cụm và sự khác biệt với cụm khác.
- *Kết quả tối ưu:* Giá trị preference mang lại điểm Silhouette Score cao nhất được chọn làm tham số tối ưu. Điều này đảm bảo số cụm phù hợp với yêu cầu bài toán, đồng thời chất lượng phân cụm đạt mức cao nhất.

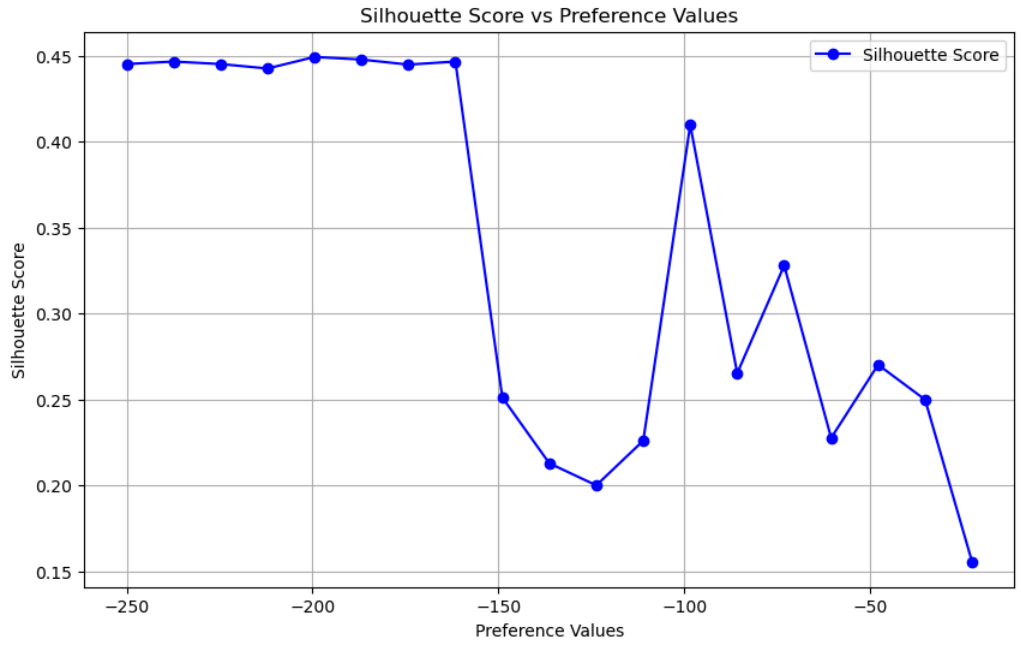
```python
def find_best_preference(scaled_features):
    # Tính ma trận tương đồng (cosine similarity)
    similarity_matrix = cosine_similarity(scaled_features)
    # Tập giá trị preference cần kiểm tra
    preference_values = np.linspace(-250, -10, 20)
    # Biến lưu kết quả tốt nhất
    best_score = -1
    best_preference = None
    best_clusters = None
    best_cluster_centers = None
    preference_list = []
    silhouette_score_list = []
    # Lặp qua các giá trị preference
    for preference in preference_values:
        # Khởi tạo Affinity Propagation
        ap = AffinityPropagation(preference= preference, affinity='precomputed', damping=0.9, random_state=42)
        ap.fit(similarity_matrix)
        # Kiểm tra số cụm hợp lệ
        if len(set(ap.labels_)) > 1 and len(set(ap.labels_)) < 100:
            # Tính Silhouette Score
            score = silhouette_score(scaled_features, ap.labels_)
            preference_list.append(preference)  # Lưu preference
            silhouette_score_list.append(score)  # Lưu Silhouette Score
            #print(f"Preference: {preference}, Silhouette Score: {score}, Number of Clusters: {len(set(ap.labels_))}")
            # Cập nhật giá trị tốt nhất nếu điểm cao hơn
            if score > best_score:
                best_score = score
                best_preference = preference
                best_clusters = ap.labels_
                best_cluster_centers = ap.cluster_centers_indices_

    # Kết quả cuối cùng
    return best_preference, best_score, best_clusters, best_cluster_centers, preference_list, silhouette_score_list
```
✓ 0.0s

Hình 5.1. Thuật toán dùng để xác định preference với Silhouette Score tốt nhất

5.2. Kết Quả Phân Cụm

5.2.1. Phân Tích Preference và Silhouette Score



Hình 5.2. Biểu đồ mối quan hệ giữa Preference và Silhouette Score

Silhouette Score bắt đầu ở mức khá cao khi preference nằm trong khoảng từ -250 đến -200, giữ ổn định ở mức gần 0.45. Điều này cho thấy trong phạm vi này, thuật toán

Affinity Propagation tạo ra các cụm có chất lượng cao, tức các điểm trong cùng một cụm có sự tương đồng cao và rõ ràng giữa các cụm khác nhau.

Khi preference tăng lên và đạt đến giá trị khoảng -160, điểm Silhouette Score bắt đầu giảm mạnh, điều này có thể chỉ ra rằng các cụm trở nên ít rõ ràng và không còn tương đồng tốt giữa các điểm trong cùng cụm. Đặc biệt, điểm Silhouette Score có xu hướng giảm khi giá trị preference từ khoảng -100 trở lên, cho thấy quá trình phân cụm không còn hiệu quả và không thể phân biệt các nhóm khách hàng rõ ràng nữa.

5.2.2. Chọn cụm tốt nhất

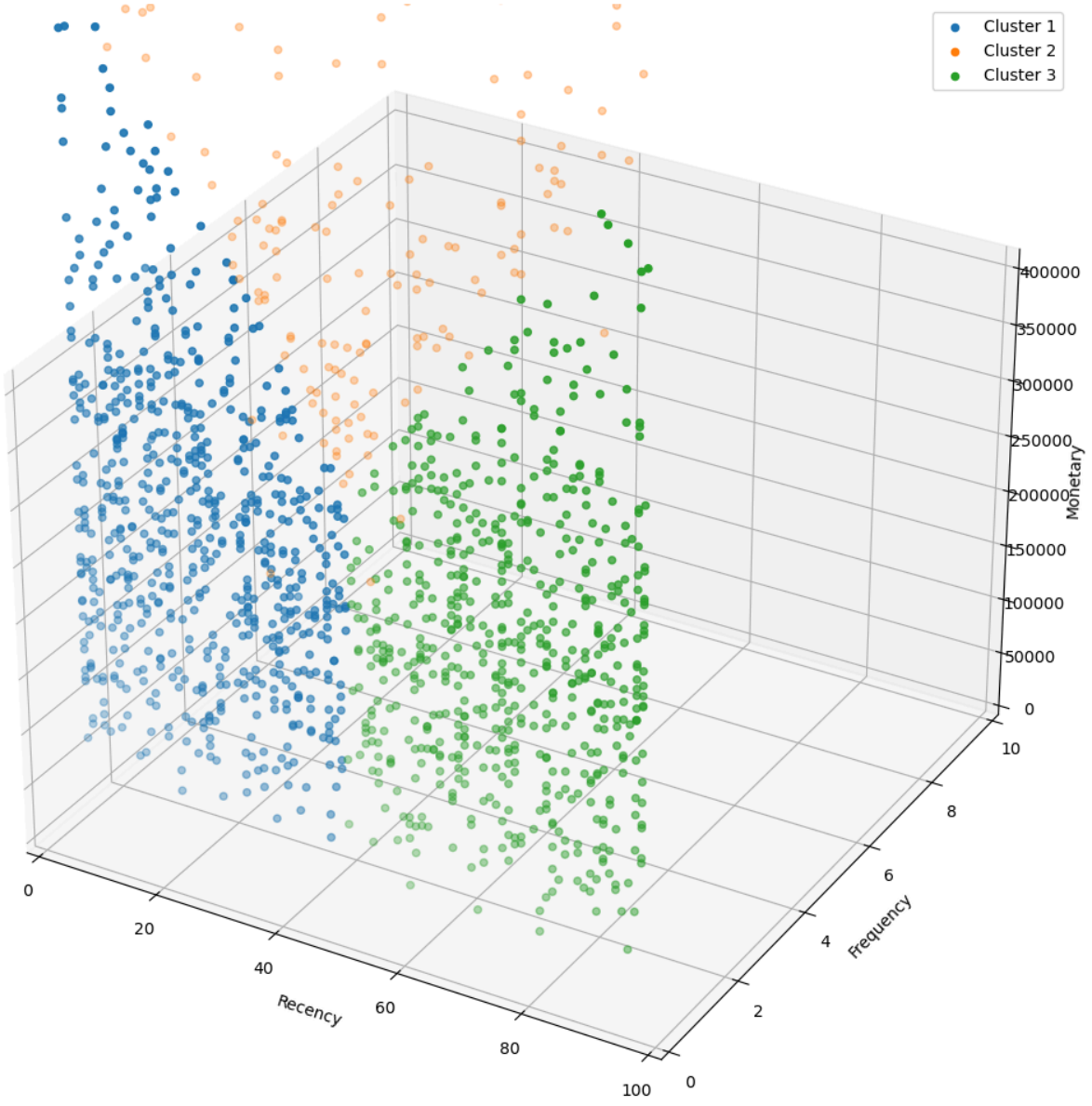
Kết quả khi preference tối ưu đem lại Silhouette Score cao nhất

```
Best Preference: -199.4736842105263, Best Silhouette Score: 0.44930826191355383
Optimal Clusters: [2 2 1 ... 2 0 0]
Cluster Centers: [  38   39 1155]
Number of Clusters: 3
```

Giá trị preference -199.47 là tối ưu, giúp tạo ra số cụm hợp lý và có chất lượng phân cụm tốt nhất, đáp ứng yêu cầu phân tích hành vi khách hàng.

5.3. Nhận Xét và Đánh Giá

5.3.1. Đánh Giá Chất Lượng Phân Cụm



Hình 5.3. Biểu đồ 3D thể hiện các cụm theo 3 trục Recency, Frequency và Monetary

Dựa trên hình ảnh, có thể thấy các điểm dữ liệu được phân thành ba cụm (Cluster 1, Cluster 2 và Cluster 3) với màu sắc khác nhau. Các cụm này phân tách rõ ràng và không có sự chồng chéo giữa chúng, điều này cho thấy phân cụm là tốt, các cụm có sự phân biệt rõ rệt.

5.3.2. Đặc Điểm Từng Nhóm Khách Hàng

Nhóm 1 (màu xanh lam): Khách hàng tiềm năng

Đặc điểm: Những khách hàng này mua gần đây nhất nhưng không thường xuyên, giá trị đơn hàng không cao. Bài toán đặt ra "Làm thế nào để giúp họ hài lòng ở những giao dịch đầu tiên này và quay lại mua hàng nhiều lần hơn, với giá trị đơn hàng lớn hơn"

Nhóm 3 (màu xanh lá): Khách hàng cần giữ chân

Đặc điểm: Đây là nhóm các khách hàng đã lâu chưa quay lại mua hàng, tần suất mua hàng và chi tiêu cũng thấp, không còn tương tác nhiều với doanh nghiệp. Bài toán đặt ra là "Làm thế nào để họ quay lại mua hàng lần nữa và khiến họ mua hàng thường xuyên hơn".

Nhóm 2 (màu cam): Khách hàng tiềm năng và khách hàng không thường xuyên

Đặc điểm: Nhóm khách hàng này bao gồm cả khách hàng mới và khách hàng lâu không giao dịch, với tần suất giao dịch thấp và giá trị đơn hàng phân bố rộng, từ thấp đến cao, có một số ít giao dịch lớn. Nhóm này có cả khách hàng mua sắm nhỏ lẻ và khách hàng chi tiêu cao, nhưng phần lớn chi tiêu ít hơn. Bài toán đặt ra "Làm sao để tăng cường mức độ trung thành, gia tăng tần suất và giá trị giao dịch của nhóm khách hàng này, đồng thời tối ưu hóa các chiến lược tiếp thị để duy trì và phát triển khách hàng tiềm năng."

5.3.3. Đề xuất Chiến Lược

Nhóm 1: Khách hàng tiềm năng

1. Khuyến mãi hấp dẫn cho lần mua tiếp theo:
    - Mục tiêu: Khuyến khích khách hàng quay lại và tăng tần suất mua hàng.

    Thực hiện:

    - Giảm giá theo đơn hàng tiếp theo: Áp dụng khuyến mãi giảm giá cho đơn hàng kế tiếp khi khách hàng quay lại trong thời gian ngắn. Điều này có thể tích hợp bằng mã giảm giá hoặc vào ứng dụng thành viên.
2. Ưu đãi cá nhân hóa:
    - Mục tiêu: Tăng sự hài lòng và tạo ra cảm giác đặc biệt cho khách hàng.

- Gửi các ưu đãi dựa trên sản phẩm mà họ đã mua gần đây. Ví dụ: nếu họ thích pizza hải sản, gửi mã giảm giá cho các loại pizza hải sản.

3. Chương trình khách hàng thân thiết:
    - Mục tiêu: Xây dựng mối quan hệ dài hạn với khách hàng, tăng tần suất mua hàng và giá trị đơn hàng trung bình.

Thực hiện:

- Tích điểm cho mỗi giao dịch: Giới thiệu chương trình khách hàng thân thiết nơi khách hàng tích điểm cho mỗi giao dịch. Khi tích đủ điểm, họ có thể nhận được phần quà hoặc giảm giá. Đảm bảo rằng phần thưởng đủ hấp dẫn để khuyến khích khách hàng tiếp tục mua sắm
- Ưu đãi đặc biệt cho thành viên: Tạo ra các ưu đãi chỉ dành riêng cho thành viên của chương trình khách hàng thân thiết, như giảm giá vào ngày sinh nhật hoặc các sự kiện đặc biệt. Ví dụ: "Là thành viên thân thiết, bạn sẽ nhận được giảm giá 20% vào ngày sinh nhật và các sự kiện đặc biệt khác!"

Nhóm 3: Khách hàng cần giữ chân

1. Chiến dịch tái kích hoạt:
    - Mục tiêu: Thu hút sự chú ý của khách hàng và khuyến khích họ quay lại mua hàng.

Thực hiện:

- Email hoặc tin nhắn với ưu đãi hấp dẫn: Gửi email hoặc tin nhắn cá nhân hóa với thông điệp hấp dẫn để tái kích hoạt sự quan tâm của khách hàng. Ưu đãi nên đủ lớn để tạo động lực quay lại.
- Ưu đãi thời hạn ngắn: Đặt thời hạn cho ưu đãi để tạo cảm giác khẩn cấp, khuyến khích khách hàng hành động nhanh chóng.Ví dụ: "Ưu đãi chỉ kéo dài trong 7 ngày! Đừng bỏ lỡ cơ hội nhận giảm giá 50%."

2. Khuyến mãi đặc biệt:
    - Mục tiêu: Cung cấp các ưu đãi đặc biệt để khách hàng cảm thấy được đánh giá cao và quay lại.

Thực hiện:

- Ưu đãi lớn hơn nhóm khách hàng khác: Cung cấp các ưu đãi mà nhóm khách hàng khác không có để làm cho khách hàng này cảm thấy đặc biệt.Ví dụ: “Mua 1 tặng 1 cho tất cả các loại pizza chỉ dành cho bạn! Ưu đãi có hạn, hãy đặt hàng ngay hôm nay.”
- Gói combo đặc biệt: Tạo ra các gói combo đặc biệt với giá hấp dẫn để thu hút khách hàng quay lại. Ví dụ: “Combo đặc biệt chỉ 199.000 VND: 1 pizza lớn, 1 phần khoai tây chiên và 1 lon nước ngọt.”

3. Nhắc nhở và cập nhật:

- Mục tiêu: Giữ liên lạc và thu hút sự chú ý của khách hàng bằng cách cung cấp thông tin mới và khuyến mãi hiện tại.

Thực hiện:

- Gửi thông tin về sản phẩm mới: Cập nhật khách hàng về các sản phẩm mới hoặc cải tiến để khơi gợi sự tò mò và hứng thú. Ví dụ: “Thử ngay món pizza mới của chúng tôi! Đặc biệt dành cho bạn với giá ưu đãi 20%.”
- Thông báo chương trình khuyến mãi hiện tại: Gửi thông báo về các chương trình khuyến mãi và sự kiện hiện tại để khuyến khích khách hàng quay lại. Ví dụ: “Khám phá chương trình khuyến mãi tháng này! Giảm giá 30% cho tất cả các món ăn kèm khi mua bất kỳ pizza nào.”
- Sử dụng các kênh khác nhau: Sử dụng nhiều kênh liên lạc như email, tin nhắn SMS, và mạng xã hội để đảm bảo thông điệp đến được với khách hàng. Ví dụ: Gửi email chi tiết về sản phẩm mới, đồng thời đăng bài trên Facebook và Instagram để tăng khả năng tiếp cận.

Nhóm 2: Khách hàng tiềm năng và khách hàng không thường xuyên

1. Mục tiêu chiến lược

- Tăng cường mức độ trung thành: Khuyến khích khách hàng quay lại và gia tăng tần suất mua hàng.
- Gia tăng giá trị giao dịch: Thúc đẩy khách hàng chi tiêu nhiều hơn trong các lần giao dịch.
- Duy trì và phát triển khách hàng tiềm năng: Tạo ra các chiến lược tiếp thị phù hợp để giữ chân và thu hút thêm khách hàng từ nhóm này.

2. Cấu trúc chương trình khách hàng thân thiết

a. Các cấp độ thành viên (Tiered Loyalty Program)

Cấp độ cơ bản (Nhóm khách hàng mới và không thường xuyên): Khách hàng tham gia lần đầu hoặc có ít hơn 3 giao dịch trong một quý.

Ưu đãi:

- Giảm giá 5-10% cho đơn hàng tiếp theo.
- Mã quà tặng có giá trị nhỏ cho lần mua sau (ví dụ: 50.000 VNĐ).
- Ưu đãi đặc biệt cho lần mua hàng tiếp theo (ví dụ: mua 1 tặng 1 sản phẩm giá trị thấp).

Cấp độ vàng (Khách hàng có tần suất giao dịch ổn định nhưng chưa thường xuyên): Khách hàng đã thực hiện từ 3 đến 5 giao dịch trong vòng 3 tháng hoặc có tổng giá trị mua hàng đạt mức tối thiểu.

Ưu đãi:

- Giảm giá 10-15% cho các lần mua sau.
- Các phần thưởng đặc biệt cho khách hàng vàng như quà tặng miễn phí (ví dụ: sản phẩm nhỏ, voucher trị giá 100.000 VNĐ).
- Tham gia các chương trình khuyến mãi đặc biệt chỉ dành cho khách hàng vàng.

Cấp độ kim cương (Khách hàng tiềm năng, có tần suất và chi tiêu cao hơn): Khách hàng thực hiện từ 6 giao dịch trở lên hoặc đạt tổng chi tiêu trên một mức tối thiểu trong 3 tháng (ví dụ: 5 triệu VNĐ).

Ưu đãi:

- Giảm giá 20-25% cho các lần mua tiếp theo.
- Quà tặng giá trị cao hoặc sản phẩm miễn phí khi đạt mức chi tiêu nhất định.
- Quà tặng sinh nhật hoặc quà tặng trong các dịp đặc biệt.
- Cung cấp quyền truy cập sớm vào các chương trình khuyến mãi đặc biệt, đợt giảm giá lớn.

b. Hệ thống tích điểm và thưởng cho các giao dịch

Tích điểm theo giá trị giao dịch: Mỗi giao dịch của khách hàng sẽ được quy đổi thành điểm thưởng. Điểm có thể được sử dụng để đổi quà tặng, giảm giá cho các lần mua hàng tiếp theo, khách hàng có thể tích điểm để lên cấp độ cao hơn và nhận các ưu đãi đặc biệt.

c. Quà tặng và khuyến mãi định kỳ

Quà tặng định kỳ: Cung cấp quà tặng miễn phí cho khách hàng sau mỗi số lần giao dịch nhất định. Quà tặng có thể là sản phẩm hoặc voucher giảm giá cho lần mua tiếp theo. Khuyến mãi trong các dịp đặc biệt: tổ chức các chương trình khuyến mãi vào các dịp lễ, Tết, hoặc sinh nhật khách hàng, cung cấp ưu đãi đặc biệt cho khách hàng thân thiết.

d. Chương trình giới thiệu bạn bè

Khuyến khích khách hàng giới thiệu bạn bè: Tạo cơ hội cho khách hàng hiện tại giới thiệu bạn bè tham gia chương trình khách hàng thân thiết.

- Ưu đãi cho người giới thiệu: Cung cấp điểm thưởng, giảm giá hoặc quà tặng cho mỗi người bạn được giới thiệu thành công.
- Ưu đãi cho người được giới thiệu: Khách hàng mới sẽ nhận được ưu đãi giảm giá đặc biệt cho lần mua đầu tiên (ví dụ: giảm 10-15% hoặc voucher trị giá 50.000 VNĐ).

CHƯƠNG 6. TỔNG KẾT ĐỀ TÀI

6.1. Kết Luận

Thông qua việc sử dụng thuật toán Affinity Propagation để phân cụm khách hàng từ bộ dữ liệu của Pizza Hut theo mô hình RFM (Recency - Frequency - Monetary), nhóm tác giả đã xác định được ba nhóm khách hàng chủ yếu, bao gồm:

- **Nhóm khách hàng tiềm năng:** Những khách hàng này mua gần đây nhất nhưng không thường xuyên, giá trị đơn hàng không cao.
- **Nhóm khách hàng cần giữ chân:** Đây là nhóm các khách hàng đã lâu chưa quay lại mua hàng, tần suất mua hàng và chi tiêu cũng thấp, không còn tương tác nhiều với doanh nghiệp.
- **Nhóm khách hàng tiềm năng và khách hàng không thường xuyên**: Nhóm khách hàng này bao gồm cả khách hàng mới và khách hàng lâu không giao dịch, với tần suất giao dịch thấp và giá trị đơn hàng phân bố rộng, từ thấp đến cao, có một số ít giao dịch lớn. Nhóm này có cả khách hàng mua sắm nhỏ lẻ và khách hàng chi tiêu cao, nhưng phần lớn chi tiêu ít hơn.

Kết quả phân cụm giúp doanh nghiệp có cái nhìn rõ ràng về hành vi của khách hàng trong từng nhóm. Việc phân cụm khách hàng tạo cơ sở giúp doanh nghiệp:

- Tăng cường hiệu quả marketing: Các chiến lược truyền thông và khuyến mãi có thể được thiết kế riêng biệt cho từng nhóm khách hàng, giúp tăng tỷ lệ chuyển đổi và sự hài lòng của khách hàng.
- Chăm sóc khách hàng: Xây dựng các chương trình chăm sóc khách hàng và chương trình khách hàng thân thiết để duy trì mối quan hệ thân thiết, củng cố mối quan hệ lâu dài với khách hàng trung thành, đồng thời kích thích khách hàng tiềm năng quay lại mua sắm và gia tăng tần suất mua sắm.
- Đánh giá và cải tiến liên tục: Đánh giá hiệu quả các chiến dịch marketing và chăm sóc khách hàng định kỳ. Điều chỉnh và cải thiện kịp thời các chiến lược marketing dựa trên kết quả phân cụm để phù hợp với sự thay đổi trong hành vi và nhu cầu

của khách hàng, đáp ứng nhu cầu cụ thể của từng nhóm, tối đa hóa hiệu quả của các chiến dịch.

6.2. Những Hạn Chế

- Về dữ liệu

Đề tài phân nhóm khách hàng theo mô hình RFM dựa trên bộ dữ liệu của Pizza Hut, các chỉ số RFM dễ thu thập và tính toán nhưng chúng không phản ánh đầy đủ bức tranh toàn cảnh về hành vi và nhu cầu của khách hàng. Việc chỉ sử dụng các chỉ số này khiến dữ liệu trở nên khá đơn giản và không đủ để phản ánh hết các khía cạnh quan trọng trong hành vi khách hàng, như các yếu tố nhân khẩu học, tâm lý, thói quen tiêu dùng, hay tác động của các yếu tố bên ngoài. Do đó, việc sử dụng bộ dữ liệu này có thể hạn chế khả năng phân tích và đánh giá chính xác các nhóm khách hàng, đồng thời chưa phản ánh đúng tiềm năng và hiệu quả của thuật toán Affinity Propagation

- Về kết quả

Kết quả áp dụng thuật toán Affinity Propagation tạo ra một cụm không rõ ràng gồm hai nhóm Khách hàng tiềm năng và Khách hàng không thường xuyên. Điều này gây khó khăn trong việc hiểu rõ về nhu cầu của khách hàng và xác định chiến lược marketing phù hợp. Lý do cho sự mơ hồ này có thể xuất phát từ tính đơn giản của dữ liệu hoặc từ sự thiếu đa dạng trong dữ liệu khách hàng, khiến thuật toán không thể phân biệt chính xác các nhóm.

- Về đánh giá

Đề tài chưa thực hiện việc đánh giá kết quả phân cụm bằng cách kết hợp nhiều mô hình học máy khác nhau. Việc chỉ sử dụng Affinity Propagation để phân cụm mà không thực hiện sự so sánh với các thuật toán phân cụm khác khiến việc đánh giá hiệu quả của thuật toán này trở nên một chiều và không phản ánh được đầy đủ các ưu điểm và hạn chế của nó.

TÀI LIỆU THAM KHẢO

Pizza Hut Việt Nam. (2023). Bộ dữ liệu đề thi RMIT Business Analytics Champion (RBAC) 2023. Dữ liệu nội bộ.

Asriny, N. I., Muhajir, M., & Andrian, D. (2021). K-affinity propagation clustering algorithm for the classification of part-time workers using the internet. *Indonesian Journal of Electrical Engineering and Computer Science, 24*(1), 464–472. https://doi.org/10.11591/ijeecs.v24.i1.pp464-472

A'yuni, T. Q., Febriati, B. N., Effendie, L. I., & Yotenka, R. (2023). MSME sales clustering based on business aid distribution priority using K-affinity propagation. *Enthusiastic International Journal of Applied Statistics and Data Science, 3*(1), Article 10. https://doi.org/10.20885/enthusiastic.vol3.iss1.art10

Dueck, D., & Frey, B. J. (2007). Non-metric affinity propagation for unsupervised image categorization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2007)*, 1–8. https://doi.org/10.1109/ICCV.2007.4408853

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972–976. https://doi.org/10.1126/science.1136800

Gan, G., & Ng, M. K.-P. (2015). Subspace clustering using affinity propagation. *Pattern Recognition*, 48(5), 1594-1607. https://doi.org/10.1016/j.patcog.2014.11.003

GeeksforGeeks.(2024, May 22). Affinity propagation. GeeksforGeeks. https://www.geeksforgeeks.org/affinity-propagation/

Karegar, P. A. (2018). Wireless fingerprinting indoor positioning using affinity propagation clustering methods. *Wireless Networks, 24*(3), 1–9. https://doi.org/10.1007/s11276-017-1507-0

Kazantseva, A., & Szpakowicz, S. (2011). Linear text segmentation using affinity propagation. *Proceedings of the 2011 Conference on Empirical Methods in Natural*

Language Processing (EMNLP 2011) (pp. 27–31). John McIntyre Conference Centre, Edinburgh, UK. https://aclanthology.org/D11-1026

Leone, M., Sumedha, & Weigt, M. (2007). Clustering by soft-constraint affinity propagation: Applications to gene-expression data. *Bioinformatics, 23*(20), 2708–2715. https://doi.org/10.1093/bioinformatics/btm414

Liu, Y., Liu, J., Jin, Y., & Zheng, T. (2020). An affinity propagation clustering based particle swarm optimizer for dynamic optimization. *Knowledge-Based Systems, 195*, 105711. https://doi.org/10.1016/j.knosys.2020.105711

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (pp. 281-297). University of California Press. http://projecteuclid.org/euclid.bsmsp/1200512992

Makhija, P. (2018). RFM Analysis for Customer Segmentation | CleverTap. CleverTap. https://clevertap.com/blog/rfm-analysis/

Qian, Y., Yao, F., & Jia, S. (2010). Band selection for hyperspectral imagery using affinity propagation. *IET Computer Vision, 3*(4), 213–222. https://doi.org/10.1049/iet-cvi.2009.0034

Santana, R., McGarry, L. M., Bielza, C., & Yuste, R. (2013). Classification of neocortical interneurons using affinity propagation. *Frontiers in Neural Circuits, 7*(185), 185. https://doi.org/10.3389/fncir.2013.00185

Tomorrow Marketers. (2023, April 8). Phân khúc khách hàng là gì và các bước phân tích khách hàng theo RFM. Tomorrow Marketers. https://blog.tomorrowmarketers.org/phan-tich-rfm-la-gi/

Wang, J., Gao, Y., Wang, K., & Lim, S.-J. (2019). An affinity propagation-based self-adaptive clustering method for wireless sensor networks. *Sensors, 19*(11), 2579. https://doi.org/10.3390/s19112579