**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TIỂU LUẬN MÔN BIG DATA**

**ÁP DỤNG MACHINE LEARNING VÀO PYSPARK**

**ĐỂ PHÂN KHÚC KHÁCH HÀNG**

**Giảng viên giảng dạy :   HỒ KHÔI**

**Sinh viên thực hiện :   VÕ THÁI HIỂN**

**MSSV :   2000005690**

**Lớp :   20DTH1D**

**Chuyên ngành :   Khoa Học Dữ Liệu**

**Môn học :   Big Data**

**Khóa :   2020**

**Tp.HCM, Tháng  9  Năm  2023**

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ …..… NĂM HỌC …….. - ….…** |

**PHIẾU CHẤM THI TIỂU LUẬN/BÁO CÁO**

BM-ChT-11

Môn thi: **BIG DATA** Lớp học phần: **20DTH1D**

Nhóm sinh viên thực hiện : **2**

1………………….**Võ Thái Hiển**………..……..Tham gia đóng góp: **50%**

2……….. **Nguyễn Phùng Tấn Phúc**..…………Tham gia đóng góp: **50%**

3. …………………………………..……………Tham gia đóng góp: ……………………..

4. …………............................................. ………Tham gia đóng góp: ……………………..

5…………………………………….. ………….Tham gia đóng góp: ……………………..

6. Tham gia đóng góp:

Ngày thi: Phòng thi:

Đề tài tiểu luận/báo cáo của sinh viên :

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  |  |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  |  |  |
| * Lập luận |  |  |  |
| * Kết luận |  |  |  |
| Trình bày |  |  |  |
| **TỔNG ĐIỂM** |  |  |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

**Hồ Khôi**

LỜI CẢM ƠN

Lời nói đầu tiên em xin gửi lời cám ơn đến thầy **Hồ Khôi** đã hướng dẫn và giúp đỡ em trong suốt học kỳ này với bộ môn Big data, thầy đã truyền đạt cho em nhiều kiến thức về big data và những kiến thức bổ ích về bộ môn này . Song với đó thầy đã tạo cho em nguồn cảm hứng, sự tìm tòi và học hỏi về môn big data trong KHDL. Đặc biệt trong lúc giảng bài, thầy luôn tích cực giảng rõ từng chi tiết, luôn vui tươi, thoải mái để tạo cho chúng em một bầu không khí dễ chịu dễ tiếp thu được bài học. Vì thế cho đến ngày hôm nay em đã thực sự hiểu rõ hơn về nó và có thêm những kiến thức hay cho ngành mà em đang theo đuổi. Cám ơn thầy đã tận tình hướng dẫn và giải đáp tất cả khúc mắc mỗi khi em gặp khó khăn trong quá trình học và làm bài. Cảm ơn thầy !

Em cũng xin cám ơn các giảng viên khoa Công Nghệ Thông Tin Trường Đại Học Nguyễn Tất Thành đã tạo điều kiện và giúp đỡ em trong quá trình học tập.

Và em cũng gửi lời cảm ơn đến tập thể lớp 20DTH1D đã cùng đồng hành với em trong học kỳ này nói chung và bộ môn này nói riêng.

***Sinh viên thực hiện***

**Võ Thái Hiển**

LỜI MỞ ĐẦU

Công nghệ thông tin là một ngành học được đào tạo để sử dụng máy tính và các phần mềm máy tính để phân phối và xử lý các dữ liệu thông tin, đồng thời dùng để trao đổi, lưu trữ và chuyển đổi các dữ liệu thông tin dưới nhiều hình thức khác nhau.

Sau khi được đào tạo, sinh viên học ngành này sẽ được trang bị kiến thức nền tảng và chuyên sâu về lĩnh vực công nghệ thông tin để nâng cao tay nghề nhằm phát triển khả năng sửa chữa, xây dựng, cài đặt, bảo trì các phần cứng của máy tính cũng như nghiên cứu và phát triển các ứng dụng phần mềm. Ngoài ra cũng được trang bị kiến thức về an toàn và bảo mật thông tin mạng, một trong những lĩnh vực quan trọng được quan tâm hàng đầu trên thế giới hiện nay.

Hiện nay, **Công nghệ thông tin (CNTT)** đang trong giai đoạn phát triển như vũ bão ở mọi lĩnh vực hoạt động khắp nơi trên thế giới.Điều này đã làm cho nghành công nghệ thông tin ngày càng được nhiều người quan tâm đặc biệt là giới trẻ hiện nay, tư tưởng thích cái mới, thích sự tìm tòi sáng tạo là yếu tố quan trọng hấp dẫn các bạn trẻ đến nghành nghề này.

Toàn cầu hóa những năm 1990 đã làm xuất hiện khuynh hướng xã hội quan trọng, đó là sự chuyển trạng thái từ xã hội công nghiệp sang xã hội kiến thức và trong đó thông tin giữ vai trò trọng yếu. Sự phát triển và ứng dụng CNTT ngày nay báo trước một thời kỳ mới với những thay đổi xã hội lớn lao. CNTTnhư một công nghệ chung xâm nhập vào mọi lĩnh vực kinh tế xã hội*.*. Là một ngành tổng thể bao gồm nhiều nhánh nhỏ như mạng lưới bưu chính viễn thông, truyền thông đa phương tiện, internet..., chúng ta có thể khẳng định rằng ở Việt Nam đã xây dựng được một cơ cấu hạ tầng có đồng bộ, đầy đủ trong hệ thống ngành công nghệ thông tin. Ở đây chúng ta có thể kể tới một dấu mốc đáng nhớ trong sự phát triển ngành CNTT đó là vào năm 1997, nước ta đã biến *“giấc mơ Internet”* thành hiện thực bằng việc tham gia kết nối vào mạng toàn cầu và tính cho tới thời điểm này, Việt nam đã trở thành quốc gia có tỷ lệ tăng trưởng Internet nhanh nhất trong khu vực và nằm trong số những quốc gia có tỷ lệ trưởng cao nhất thế giới

MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc145158104)

[LỜI MỞ ĐẦU 2](#_Toc145158105)

[MỤC LỤC 3](#_Toc145158106)

[DANH MỤC HÌNH 4](#_Toc145158107)

[CHƯƠNG 1: GIỚI THIỆU 5](#_Toc145158108)

[1.Giới thiệu đề tài 5](#_Toc145158109)

[2.Lý do chọn đề tài 6](#_Toc145158110)

[3.Mục tiêu của đề tài 6](#_Toc145158111)

[4.Phương pháp đề tài 6](#_Toc145158112)

[5.Đối tượng và phạm vi nghiên cứu 6](#_Toc145158113)

[CHƯƠNG 2: ỨNG DỤNG THUẬT TOÁN 7](#_Toc145158114)

[1.Mô tả bài toán 7](#_Toc145158115)

[2. Machine learning, Pyspark, K-means và ứng dụng vào bài toán 8](#_Toc145158116)

[2.1.Machine learning 8](#_Toc145158117)

[2.2.Pyspark 9](#_Toc145158118)

[2.3.K-means 10](#_Toc145158119)

[2.4. Ứng dụng vào bài toán 11](#_Toc145158120)

[CHƯƠNG 3: THỰC NGHIỆM 13](#_Toc145158121)

[1.Xây dựng mô hình 13](#_Toc145158122)

[2.Kết luận 23](#_Toc145158123)

[3.Kết quả đạt được 23](#_Toc145158124)

[4.Điểm còn hạn chế 24](#_Toc145158125)

[5.Hướng phát triển 24](#_Toc145158126)

[LINK SOURCE CODE 24](#_Toc145158127)

[TÀI LIỆU THAM KHẢO 25](#_Toc145158128)

DANH MỤC HÌNH

[Hình 1: Minh họa kinh doanh online 5](#_Toc145157959)

[Hình 2: Minh họa Machine learning 8](#_Toc145157960)

[Hình 3: Minh họa Pypark 9](#_Toc145157961)

[Hình 4: K-means in Machine learning 10](#_Toc145157962)

[Hình 5: Data dùng để thực nghiệm 13](#_Toc145157963)

CHƯƠNG 1: GIỚI THIỆU

1.Giới thiệu đề tài



Hình 1: Minh họa kinh doanh online

Trải qua nhiều thập kỷ, cuộc cách mạng công nghệ đã thay đổi cách chúng ta sống và làm việc. Trong số những thay đổi đáng kể nhất trong cuộc cách mạng này là sự phổ biến ngày càng tăng của việc đặt hàng online. Từ việc mua sắm hàng hóa hàng ngày cho đến dịch vụ cung cấp thực phẩm, tài liệu giáo dục và nhiều lĩnh vực khác, đặt hàng online đã trở thành một phần không thể thiếu của cuộc sống hiện đại.

Tuy nhiên vấn đề đặt ra ở đây đối với các doanh nghiệp, các chủ cửa hàng... là làm thế nào để kiểm soát được lượng khách hàng đặt hàng, lượng hàng trong kho, những khách hàng không nhận hàng,… từ đó có thể quản lý cửa hàng, công ty một cách tuyệt đối nhất và có những chiến lược quảng cáo, chăm sóc khách hàng phù hợp.

Do đó chúng em quyết định sẽ phân khúc khách hàng nhằm mục đích giúp các doanh nghiệp, chủ cửa hàng có thể quản lý dữ liệu một cách dể dàng hơn từ đó có thể giúp học đưa ra những đề xuất chiến lượt để phát triển công ty, cửa hàng.

2.Lý do chọn đề tài

Do chúng em muốn giúp các công ty sử dụng để phân khúc khách hàng của mình, từ đó có những chiến lược quản cáo, chăm sóc khách hàng phù hợp

Ví dụ: Nếu bạn ghé vào cửa hàng quần áo X vào mỗi dịp đầu tháng, có thể bạn được xếp vào nhóm những người tiêu tiền đầu tháng, nghèo cuối tháng, hay lương được trả vào cuối tháng, thích mua sắm hàng tháng, thích mua quần âu,... Và khi đó, công ty X sẽ có những quảng cáo với mác "Chỉ dành riêng cho bạn" hay "Duy nhất trong ngày hôm nay" và những quảng cáo này sẽ xuất hiện cho những người cùng nhóm với bạn, xuất hiện vào đầu tháng, các mặt hàng sale chủ yếu là quần âu, ...

3.Mục tiêu của đề tài

Giúp được các doanh nghiệp, cửa hàng công ty có thể quản lý được dữ liệu một cách dể dàng. Tạo ra chiến lược tiếp thị hiệu quả hơn, chăm sóc khách hàng tốt hơn và tối ưu hóa lợi nhuận.

4.Phương pháp đề tài

Chúng em sẽ áp dụng kiến thức đã học của mình đó là áp Machine Learning vào PySpark và K-means cùng với bộ dữ liệu tìm kiếm được trên mạng để phân khúc khách hàng, giải quyết đề tài

5.Đối tượng và phạm vi nghiên cứu

**Đối tượng**: tất cả các doanh nghiêp, các chủ cửa hàng, những người kinh doanh online

**Phạm vi nghiên cứu:** môi trường kinh doanh trực tuyến, dựa vào internet và các nền tảng kỹ thuật số để mua bán sản phẩm hoặc cung cấp dịch vụ. Đây là một môi trường đa dạng và phức tạp, bao gồm nhiều khía cạnh và yếu tố quan trọng.

CHƯƠNG 2: ỨNG DỤNG THUẬT TOÁN

1.Mô tả bài toán

Bài toán phân khúc khách hàng là một phần quan trọng trong lĩnh vực kinh doanh và marketing. Đây là quá trình phân nhóm khách hàng dựa trên các đặc điểm chung để hiểu rõ hơn về hành vi mua sắm, sở thích và nhu cầu của từng nhóm khách hàng khác nhau. Mục tiêu của việc này là tạo ra chiến lược tiếp thị hiệu quả hơn, chăm sóc khách hàng tốt hơn và tối ưu hóa lợi nhuận.

1. **Thu thập và tiền xử lý dữ liệu**: Đầu tiên, chúng em sẽ thu thập dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu của cửa hàng, hệ thống ghi nhận giao dịch, và dữ liệu khách hàng. Dữ liệu này có thể chứa các thông tin như thời gian mua sắm, số tiền đã tiêu, và danh sách các sản phẩm đã mua. Dữ liệu cần được tiền xử lý để loại bỏ dữ liệu không hợp lệ, thiếu, hoặc nhiễu.
2. **Xây dựng mô hình Machine Learning:** Sử dụng PySpark, mô hình Machine Learning sẽ được xây dựng để phân nhóm khách hàng. Trong trường hợp này, chúng ta sẽ sử dụng thuật toán K-means, một thuật toán phân cụm (clustering) phổ biến, để phân nhóm khách hàng dựa trên các đặc điểm mua sắm của họ.
3. **Chọn số cụm (clusters):** Trước khi áp dụng K-means, công ty cần xác định số lượng cụm (clusters) mà họ muốn phân chia khách hàng thành. Số cụm này có thể được chọn dựa trên kiến thức thực tế hoặc bằng cách sử dụng các phương pháp thống kê như phương pháp Elbow để xác định số tối ưu của cụm.
4. **Áp dụng thuật toán K-means:** Sử dụng thuật toán K-means trong PySpark, công ty sẽ phân nhóm khách hàng thành các cụm dựa trên các đặc điểm mua sắm của họ. K-means sẽ xác định trọng tâm của từng cụm và gán từng khách hàng vào cụm gần nhất dựa trên khoảng cách Euclidean.

Sau khi áp dụng K-means, doanh nghiệp sẽ có các cụm khách hàng. Các nhóm này có thể được mô tả bằng các đặc điểm chung, ví dụ như "những người mua sắm hàng ngày," "những người mua sắm lớn vào các dịp đặc biệt," và "những người mua sắm ít và thường vào cuối tuần." Dựa trên việc phân nhóm khách hàng, doang nghiệp có thể triển khai các chiến lược tiếp thị và quảng cáo riêng biệt cho từng nhóm. Chẳng hạn, họ có thể tạo các ưu đãi hoặc chiến dịch quảng cáo đặc biệt cho từng segment. Công ty sẽ theo dõi hiệu suất của các chiến lược tiếp thị và sử dụng phản hồi từ khách hàng để tối ưu hóa chiến lược theo thời gian.

2. Machine learning, Pyspark, K-means và ứng dụng vào bài toán

2.1.Machine learning

Machine Learning (ML) là một lĩnh vực trong lĩnh vực trí tuệ nhân-made (Artificial Intelligence - AI) tập trung vào việc phát triển các thuật toán và mô hình máy tính có khả năng học hỏi từ dữ liệu và tự động cải thiện hiệu suất hoạt động mà không cần lập trình cụ thể. Trong ML, máy tính được lập trình để phân tích và hiểu dữ liệu, sau đó áp dụng các kiến thức học được từ dữ liệu đó để thực hiện các nhiệm vụ cụ thể mà nó chưa được lập trình trước đó.



Hình 2: Minh họa Machine learning

ML chủ yếu được chia thành các loại sau:

* **Học giám sát (Supervised Learning):** Trong loại này, mô hình máy tính được huấn luyện trên một tập dữ liệu đào tạo có nhãn, tức là dữ liệu đã được gán nhãn với kết quả mong muốn. Mục tiêu là để mô hình học cách dự đoán đúng kết quả cho dữ liệu mới mà nó chưa thấy trước đó, dựa trên các mẫu từ tập dữ liệu đào tạo.
* **Học không giám sát (Unsupervised Learning):** Trong loại này, không có nhãn cho dữ liệu đào tạo. Mô hình cố gắng tự động tìm kiếm các cấu trúc, mẫu hoặc nhóm trong dữ liệu mà không cần hướng dẫn. Phân cụm (clustering) và giảm chiều (dimensionality reduction) là ví dụ phổ biến trong học không giám sát.
* **Học tăng cường (Reinforcement Learning):** Trong học tăng cường, mô hình máy tính tương tác với một môi trường và học thông qua việc thử nghiệm và sai lầm. Mô hình được tạo ra để tối ưu hóa một hàm phần thưởng, tức là đạt được kết quả tốt nhất trong môi trường đã cho.

Machine Learning đã trở thành một công cụ quan trọng trong nhiều lĩnh vực doanh nghiệp và nghiên cứu khoa học, và nó ngày càng phát triển và phức tạp với sự xuất hiện của các mô hình sâu (Deep Learning) và nền tảng công nghệ tiên tiến.

2.2.Pyspark

PySpark là một giao diện Python cho Apache Spark, một framework mã nguồn mở mạnh mẽ được thiết kế để xử lý dữ liệu lớn và tính toán phân tán. PySpark cho phép bạn sử dụng Python để tương tác và phát triển ứng dụng trên nền Spark.



Hình 3: Minh họa Pypark

Apache Spark là một hệ thống xử lý dữ liệu phân tán, có khả năng xử lý và phân tích dữ liệu với tốc độ nhanh hơn so với nhiều giải pháp truyền thống. Nó hỗ trợ nhiều loại công việc, bao gồm xử lý dữ liệu lớn, phân tích dữ liệu, học máy, xử lý luồng dữ liệu (streaming), và nhiều ứng dụng khác.

PySpark cho phép bạn sử dụng Python để tận dụng sức mạnh của Apache Spark, giúp dễ dàng viết và thực thi các tác vụ xử lý dữ liệu phân tán bằng Python. Điều này làm cho việc phát triển ứng dụng và thực hiện các nhiệm vụ phức tạp liên quan đến xử lý dữ liệu trở nên thuận tiện hơn cho những người làm việc trong lĩnh vực dữ liệu lớn và khoa học dữ liệu.

2.3.K-means

K-Means là một thuật toán phân cụm (clustering) trong machine learning và thống kê. Nó được sử dụng để phân chia một tập hợp dữ liệu thành các nhóm con (clusters) dựa trên sự tương tự giữa các điểm dữ liệu. Mục tiêu của K-Means là tìm ra các clusters sao cho các điểm dữ liệu trong cùng một cluster có sự tương tự lớn nhất với nhau, trong khi các điểm ở các clusters khác nhau có sự tương tự nhỏ nhất.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, Đồ họa

Mô tả được tạo tự động

Hình 4: K-means in Machine learning

Thuật toán K-Means hoạt động như sau:

* **Chọn số lượng clusters (K):** Trước khi bắt đầu, bạn phải xác định số lượng clusters mà bạn muốn tạo. Điều này có thể được xác định trước dựa trên kiến thức về dữ liệu hoặc bằng cách sử dụng các phương pháp như phân tích biểu đồ Elbow.
* **Chọn điểm ban đầu cho các centroids:** Mỗi cluster được đại diện bởi một điểm trung tâm gọi là centroid. Ban đầu, bạn cần chọn K điểm dữ liệu từ tập dữ liệu làm các centroid ban đầu. Cách chọn ban đầu này có thể ảnh hưởng đến kết quả cuối cùng.
* **Phân loại các điểm dữ liệu vào clusters**: Dựa trên sự tương tự giữa các điểm dữ liệu và các centroids, mỗi điểm dữ liệu sẽ được gán vào cluster có centroid gần nhất.
* **Cập nhật centroids**: Sau khi phân loại các điểm dữ liệu, centroids của các clusters được cập nhật bằng cách tính trung bình của tất cả các điểm thuộc về cluster đó.
* **Lặp lại quá trình 3 và 4:** Quá trình phân loại và cập nhật centroids được lặp lại cho đến khi không có sự thay đổi đáng kể nào trong việc phân loại hoặc khi đạt đến số lần lặp tối đa được xác định trước (hoặc một điều kiện dừng khác).
* **Kết quả:** Kết quả cuối cùng của K-Means là các clusters và centroids tương ứng.

K-Means thường được sử dụng trong nhiều lĩnh vực, chẳng hạn như phân tích dữ liệu, nhóm khách hàng, phát hiện bất thường, và nhiều ứng dụng khác trong machine learning và khoa học dữ liệu.

2.4. Ứng dụng vào bài toán

**1.Thu thập và Tiền xử lý dữ liệu:**

- Thu thập dữ liệu về hành vi mua sắm và hành vi truy cập trang web của khách hàng.

- Tiền xử lý dữ liệu để loại bỏ dữ liệu không cần thiết, xử lý giá trị thiếu, và chuẩn hóa dữ liệu nếu cần.

**2.Sử dụng PySpark:**

- Sử dụng PySpark để xử lý dữ liệu lớn và tính toán phân tán.

- Tạo một phiên làm việc Spark để làm việc với dữ liệu lớn và tận dụng tính hiệu suất của Spark.

**3.Chọn số lượng clusters (K):**

- Sử dụng PySpark và các công cụ trực quan để xác định số lượng clusters phù hợp. Phương pháp phổ biến là sử dụng biểu đồ Elbow để xác định K tối ưu.

**4.Áp dụng thuật toán K-Means:**

- Sử dụng thuật toán K-Means trong PySpark để phân khúc khách hàng thành các clusters dựa trên các đặc trưng đã chuẩn bị trước đó. PySpark cho phép bạn thực hiện việc này trên dữ liệu lớn một cách hiệu quả.

**5.Phân tích và đánh giá kết quả:**

- Sau khi áp dụng K-Means, phân tích kết quả của việc phân khúc khách hàng. Điều này bao gồm việc xem xét đặc điểm của từng cluster và hiểu rõ hành vi của từng nhóm khách hàng.

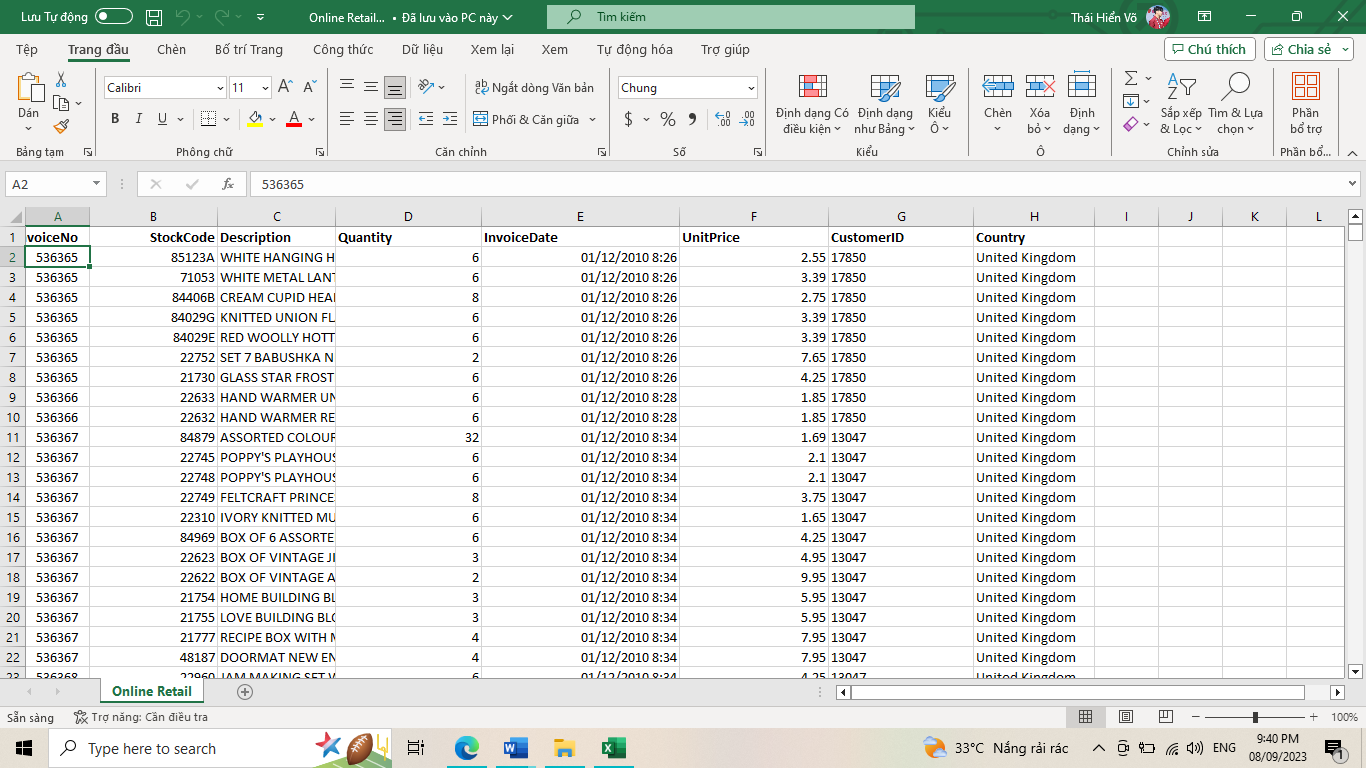
- Đánh giá chất lượng của mô hình K-Means bằng cách sử dụng các chỉ số định lượng như Silhouette Score hoặc phân tích trực quan.

CHƯƠNG 3: THỰC NGHIỆM

1.Xây dựng mô hình

Dữ liệu dùng để thực hiện bài toán gồm có:

* InvoiceNo: ID của đơn hàng, nếu ID bắt đầu bằng chữ "c" thể hiện đơn hàng đó bị hủy (Cancel)
* StockCode: Mã sản phẩm
* Description: Tên sản phẩm
* Quantity: Số lượng sản phẩm trên đơn đặt hàng
* InvoiceDate: Ngày và giờ khi đơn hàng được tạo
* UnitPrice: Giá sản phẩm trên mỗi đơn vị, tính bằng pound
* CustomerID: ID của khách hàng
* Country: Quốc gia nơi khách hàng cư trú



Hình 5: Data dùng để thực nghiệm

**Thực hiện phân tích dữ liệu cơ bản:**

Để bắt đầu các thao tác dữ liệu với PySpark, chúng ta cần khởi tạo session với SparkSession. Nó xây dựng một khung dữ liệu trong PysPark để chúng ta có thể sử dụng các chức năng của PysPark lên dữ liệu của mình

# Init SparkSession

spark = SparkSession.builder.appName("Pyspark Tutorial").config("spark.memory.offHeap.enabled","true").config("spark.memory.offHeap.size","10g").getOrCreate()

Cùng đọc dữ liệu trong file CSV vừa convert bằng PySpark

# Read data frame

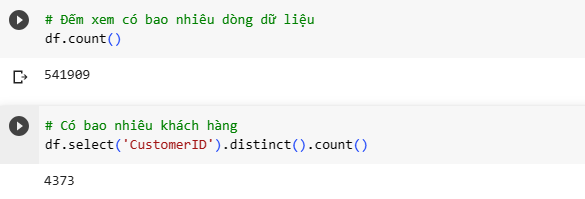
df = spark.read.csv('/content/drive/MyDrive/Đồ án/Big data/Online Retail.csv',header=True,escape="\"")

df.show(5)

**Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, tài liệu

Mô tả được tạo tự động**

Một số truy vấn cơ bản rất dể hiểu bằng pyspark



**Tiền xử lý dữ liệu**

Quay lại bài toán phân cụm, có thể thấy trong file dữ liệu có tới 7 trường dữ liệu, không phải trường dữ liệu nào cũng có ích cho phân cụm, và cũng không phải sẽ dùng được dữ liệu luôn.

Loại bỏ những đơn hàng bị hủy

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Mô tả được tạo tự động

Định dạng lại trường dữ liệu ngày tháng

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

**RFM** (bao gồm Recency (Tần suất mua hàng gần đây), Frequency (Tần suất mua hàng), và Monetary (Giá trị đặt hàng)) là một phương pháp phân tích khách hàng được sử dụng rộng rãi trong lĩnh vực tiếp thị để đánh giá giá trị của khách hàng dựa trên hành vi mua hàng của họ.

* **Recency:** đo thời điểm mà khách hàng đã mua hàng lần cuối. Khách hàng mới mua hàng gần đây được xem là có giá trị cao hơn so với khách hàng mua hàng lâu đến không mua hàng nữa.
* **Frequency:** đo tần suất mà khách hàng mua hàng trong một khoảng thời gian nhất định. Khách hàng mua hàng thường xuyên được xem là có giá trị hơn so với những khách hàng mua hàng ít lần.
* **Monetary:** đo giá trị đặt hàng của khách hàng. Khách hàng đặt hàng có giá trị cao hơn được xem là có giá trị cao hơn so với những khách hàng đặt hàng có giá trị thấp.

**Recency**

Trong phần này, chúng em sẽ có mục tiêu tính toán ra một giá trị đại diện cho việc thời điểm khách hàng mua lần cuối so với 1 mốc 0 nhất định (ở đây mình chọn mốc 0 là thời gian đầu tiên có đơn hàng đã tính ra bên trên). Sau đó chỉ cần lấy thời gian gần nhất khách hàng đặt đơn trừ đi mốc thời gian đó, ta sẽ có 1 giá trị đại diện cho Recency. Rõ ràng giá trị này càng lớn chứng tỏ khách hàng càng mua gần đây. Việc tính toán này sẽ thông qua 1 số bước như sau

* Tạo 1 cột mới, đặt giá trị của tất cả cột đó là ngày đầu tiên có đơn hàng. Cột này có tên "from\_date"

df = df.withColumn("from\_date", lit("2010-12-01 08:26:00"))

* Lấy giá trị thời gian mua của từng đơn hàng trừ đi from\_date, ta sẽ biết đơn hàng đo được đặt cách mốc thời gian 0 là bao nhiêu (theo đơn vị timestamp), giá trị sẽ được lưu tại cột 'recency'

df = df.withColumn('from\_date',to\_timestamp("from\_date", 'yy-MM-dd HH:mm'))

df2=df.withColumn('from\_date',to\_timestamp(col('from\_date'))).withColumn('recency',col("date").cast("long") - col('from\_date').cast("long"))

* Mỗi khách hàng có thể mua nhiều lần vào nhiều mốc thời gian khác nhau, chúng ta chỉ quan tâm lần cuối cùng họ mua, vì vậy cần xử lý lại cột 'recency'

df2=df2.join(df2.groupBy('CustomerID').agg(max('recency').alias('recency')),on='recency',how='leftsemi')

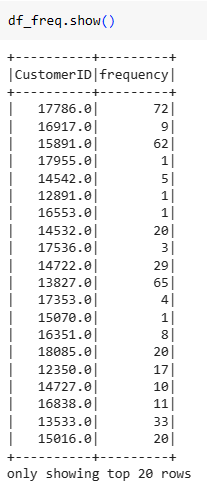
**Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, số

Mô tả được tạo tự động**

**Frequency**

Phần này thì chúng ta sẽ tính tần suất một khách hàng mua một đồ gì đó. Chúng ta chỉ cần nhóm theo từng ID khách hàng và đếm số mặt hàng họ đã mua

df\_freq=df2.groupBy('CustomerID').agg(count('InvoiceNo').alias('frequency'))



Lúc này frequency dataframe của chúng ta chỉ có 2 cột đứng riêng lẻ, chúng ta sẽ nối nó vào dataframe chúng ta đang làm việc để thống nhất, cũng như check xem mình truy xuất có đúng không

# join data

df3 = df2.join(df\_freq,on='CustomerID',how='inner')

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, số

Mô tả được tạo tự động

**Monetary**

Phần này tính xem mỗi khách hàng đã chi bao nhiêu tiền để mua sắm, phần này sẽ chia làm 2 bước

* Tính số lượng và đơn giá của một lần mua hàng

m\_val = df3.withColumn('TotalAmount',col("Quantity") \* col("UnitPrice"))

* Tính tổng số tiền mà khách hàng đã chi

m\_val=m\_val.groupBy('CustomerID').agg(sum('TotalAmount').alias('monetary\_value'))

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

* Join dữ liệu vào dataframe đang tổng hợp

# join

final\_df = m\_val.join(df3,on='CustomerID',how='inner')

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, tài liệu

Mô tả được tạo tự động

Kiểm tra những dữ liệu cần thiết cho phần sau đã đủ và đúng hay chưa thôi nào. Từ đây chúng ta cũng chỉ cần quan tâm 4 trường dữ liệu này

final\_df=final\_df.select(['recency','frequency','monetary\_value','CustomerID']).distinct()

final\_df.show()

Ảnh có chứa văn bản, thực đơn, ảnh chụp màn hình, tài liệu

Mô tả được tạo tự động

**Chuẩn hóa dữ liệu**

Đây là một phần vô cùng quan trọng và không thể thiếu trong các bài toán về dữ liệu. Mỗi trường dữ liệu có một đơn vị khác nhau, nếu không chuẩn hóa thì chắc chắn sẽ nảy sinh nhiều vấn đề về sau.

from pyspark.ml.feature import VectorAssembler

from pyspark.ml.feature import StandardScaler

assemble=VectorAssembler(inputCols=[

    'recency','frequency','monetary\_value'

], outputCol='features')

assembled\_data=assemble.transform(final\_df)

scale=StandardScaler(inputCol='features',outputCol='standardized')

data\_scale=scale.fit(assembled\_data)

data\_scale\_output=data\_scale.transform(assembled\_data)

Thử xem dữ liếu sau khi chuẩn hóa của chúng ta sẽ trông thế nào

**Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động**

**Triển khai học máy**

**Tìm số cụm k**

Một trong những chiến lược đơn giản nhất để quyết định số cụm trong K-means đó là phương pháp khủy tay ([elbow-method](https://towardsdatascience.com/elbow-method-is-not-sufficient-to-find-best-k-in-k-means-clustering-fc820da0631d#:~:text=The%20elbow%20method%20is%20a,cluster%20and%20the%20cluster%20centroid.)). Tức ta sẻ thử chạy K-means với nhiều cụm và trực quan hóa kết quả, tìm ra điểm uốn giống như khủy tay và lựa chọn điểm này

from pyspark.ml.clustering import KMeans

from pyspark.ml.evaluation import ClusteringEvaluator

import numpy as np

from tqdm import tqdm

cost = np.zeros(10)

evaluator = ClusteringEvaluator(predictionCol='prediction', featuresCol='standardized',metricName='silhouette', distanceMeasure='squaredEuclidean')

for i in tqdm(range(2,10)):

    KMeans\_algo=KMeans(featuresCol='standardized', k=i)

    KMeans\_fit=KMeans\_algo.fit(data\_scale\_output)

    output=KMeans\_fit.transform(data\_scale\_output)

    cost[i] = KMeans\_fit.summary.trainingCost

Lúc này các giá trị cost đã được lưu lại thành 1 mảng, plot lên chúng ta sẽ thấy được điểm k cần tìm (ở đây là 3, ưu tiên k càng nhỏ càng tốt)

import pandas as pd

import pylab as pl

df\_cost = pd.DataFrame(cost[2:])

df\_cost.columns = ["cost"]

new\_col = range(2,10)

df\_cost.insert(0, 'cluster', new\_col)

pl.plot(df\_cost.cluster, df\_cost.cost)

pl.xlabel('Number of Clusters')

pl.ylabel('Score')

pl.title('Elbow Curve')

pl.show()

Ảnh có chứa văn bản, hàng, biểu đồ, Sơ đồ

Mô tả được tạo tự động

K bằng 4 là hợp lý nhất đối với dữ liệu

**Trực quan hóa**

Phần này chúng ta sẽ sử dụng matplotlib để trực quan hóa phân khúc khách hàng

import matplotlib.pyplot as plt

import seaborn as sns

df\_viz=preds.select('recency','frequency','monetary\_value','prediction')

df\_viz = df\_viz.toPandas()

avg\_df = df\_viz.groupby(['prediction'], as\_index=False).mean()

list\_value = ['recency','frequency','monetary\_value']

for i in list\_value:

    sns.barplot(x='prediction',y=str(i),data=avg\_df)

    plt.show()

kết quả:

Ảnh có chứa văn bản, ảnh chụp màn hình, Hình chữ nhật, biểu đồ

Mô tả được tạo tự động Ảnh có chứa văn bản, ảnh chụp màn hình, Hình chữ nhật, biểu đồ

Mô tả được tạo tự động

Ảnh có chứa văn bản, ảnh chụp màn hình, Hình chữ nhật, màn hình

Mô tả được tạo tự động

Hình vẽ tương đối rõ ràng về các đặc trưng, chúng ta có thể rút ra được một số nhận xét tổng quan như sau:

* *Nhóm 0*: Nhóm này có tần suất mua hàng tương đối cao, cao trội hơn 3 nhóm còn lại, lần truy cập gần nhất cũng tương đối cao, giá trị tiền mua hàng tương đối nhỏ, cho thấy là một đối tượng đa số là cá nhân, hướng tới các sản phẩm giá rẻ
* *Nhóm 1*: Nhóm này có cả 3 chỉ số lần truy cập gần nhất, tấn suất mua hàng và tổng tiền mua hàng rất thấp, không có quá nhiều hi vọng là khách hàng tiềm năng, khả năng cao sẽ ngừng mua hàng trong thời gian tới
* *Nhóm 2*: Tần suất đặt hàng rất ít, tuy nhiên gần đây lại đặt phổ biến, lượng tiền mua hàng cao vượt trội nhiều lần so với các nhóm khác. Nhóm này khả năng là các doanh nghiệp có xu hướng mua các loại hàng có giá trị cao hoặc mua có số lượng lớn
* *Nhóm 3*: Nhóm này có tổng tiền mua hàng rất thấp, tần suất mua hàng không quá nhiều, và truy cập gần đây tương đối cao. Nhóm này có khả năng mua theo đợt, hoặc là những người mới tham gia sàn thương mại.

2.Kết luận

Như vậy chúng em đã hoàn thành xây dựng mô hình phân khúc khách hàng. Mặc dù mô hình thoạt nhìn rất đơn giản, tuy nhiên để xây dựng được một mô hình hoàn chỉnh phải bỏ ra công sức nghiên cứu lựa chọn mô hình phù hợp cũng như tối ưu tốc độ xử lý và độ chính xác cho chương trình

3.Kết quả đạt được

- Xây dựng được mô hình phân khúc khách hàng cho bài toán đặt ra

- Trong quá trình thực hiện tiểu luận, chúng em đã tìm hiểu được cách xây dựng mô hình Machine learning trong pyspark và thuật toán k-means và áp dụng vào bài toán và sử dụng ngôn ngữ python để thực thi bài toán

- Sử dụng các thuật toán trong thư viện

- Kết quả chạy chương trình tương đối tốt so với yêu cầu dự đoán

4.Điểm còn hạn chế

**-** Do thời gian thực hiện và kiến thức có hạn nên mô hình chưa thực sự tối ưu nhất

- Độ chính xác còn sai số so với dự tính

5.Hướng phát triển

**-** Trong tương lai tôi sẽ cải thiện hơn về độ chính xác của mô hình

- Cải thiện tốc độ tính toán cho thuật toán nhằm đáp ứng cho các bài toán đòi hỏi thời gian thực

- Nâng cao tính thực dụng của chương trình đối với đời thực

LINK SOURCE CODE

<https://colab.research.google.com/drive/1i8kBe8J5n7HWJ9XxfNKg6IWIldxLfqSf?usp=sharing>

TÀI LIỆU THAM KHẢO

[1] **"Learning PySpark" của Tomasz Drabas:** Cuốn sách này cung cấp hướng dẫn cụ thể về cách sử dụng PySpark và Machine Learning trong ngữ cảnh Apache Spark.

Link: [Learning PySpark](https://www.packtpub.com/product/learning-pyspark/9781786463708)

[2] **"Advanced Analytics with Spark" của Sandy Ryza, Uri Laserson, Sean Owen và Josh Wills:** Cuốn sách này giới thiệu cách sử dụng Spark cho Machine Learning và phân tích dữ liệu nâng cao.

Link: [Advanced Analytics with Spark](https://www.oreilly.com/library/view/advanced-analytics-with/9781491912768/)

[3] **"Apache Spark Machine Learning Blueprints" của Alex Liu và Jason Bell:** Cuốn sách này cung cấp các bản mẫu và hướng dẫn về việc triển khai các dự án Machine Learning bằng Apache Spark.

Link: [Apache Spark Machine Learning Blueprints](https://www.packtpub.com/product/apache-spark-machine-learning-blueprints/9781785880392)

[4] **"Spark: The Definitive Guide" của Bill Chambers và Matei Zaharia:** Cuốn sách này không chỉ cung cấp thông tin về PySpark và Machine Learning mà còn giới thiệu nhiều khía cạnh khác của Apache Spark.

Link: [Spark: The Definitive Guide](https://www.oreilly.com/library/view/spark-the-definitive/9781491912201/)

[5] **Bài viết trực tuyến "Introduction to K-Means Clustering with Apache Spark" của Databricks:** Bài viết này giới thiệu cách triển khai K-Means Clustering bằng PySpark và cung cấp ví dụ cụ thể.