TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP.HCM

KHOA CÔNG NGHỆ THÔNG TIN

TECHNICAL REPORT

Người thực hiện: Đào Thanh Thiện

1. Basics in Code Practice

Do bản thân chưa từng tìm hiểu về cách Docker hoạt động nên việc tạo một Docker container như yêu cầu là chưa thể thực hiện được.

2. Model Ideation on Jupyter Notebook

2.1 Exploring the data.

Tập dữ liệu đọc từ file ML-technicaltest-ecommerce.csv.

	InvoiceNo	StockCode	Description	Quantity	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	3.39	17850.0	United Kingdom

Dữ liệu có 541909 dòng và 7 cột. Với:

+ InvoiceNo : Mã hóa đơn

+ StockCode : Mã Sản phẩm

+ Description : Mô tả các mặt hàng Sản phẩm

+ Quantity : Số lượng các mặt hàng trong hóa đơn

+ UnitPrice : Đơn giá của các mặt hàng trong hóa đơn

+ CustomerID : ID khách hàng của hóa đơn

+ Country : Quốc gia của khách hàng.

Nhìn sâu hơn vào dữ liệu, ta thấy dữ liệu bị thiếu.

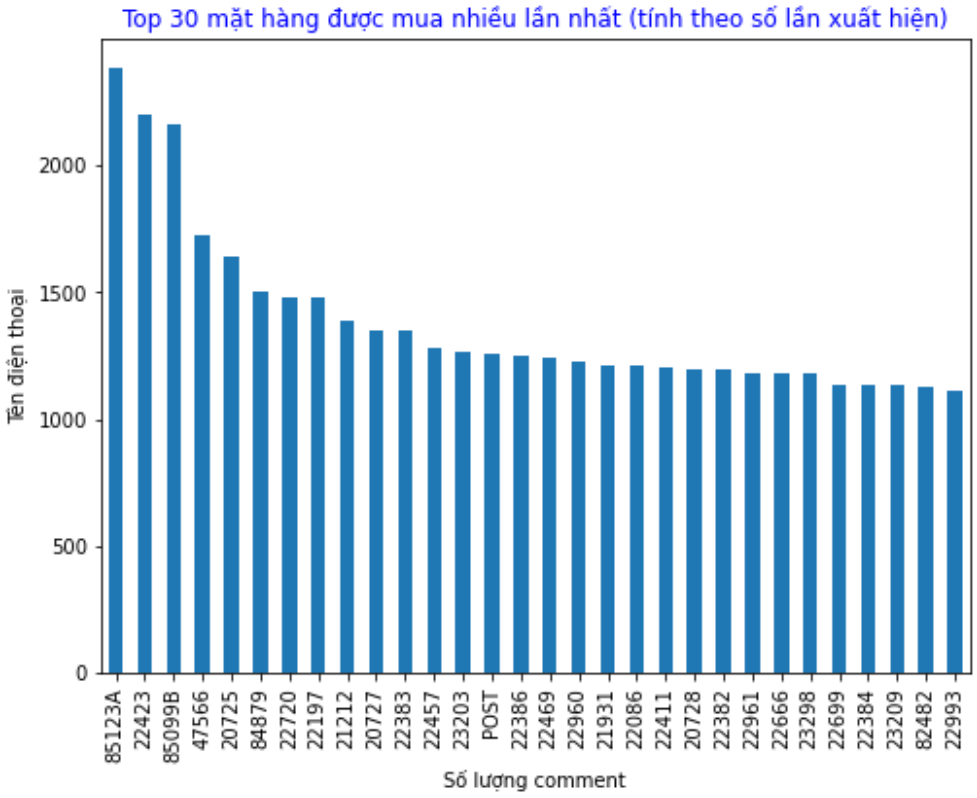
```
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 7 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   UnitPrice    541909 non-null  float64
 5   CustomerID   406829 non-null  float64
 6   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(4)
memory usage: 28.9+ MB

=> Dữ liệu bị thiếu.
```

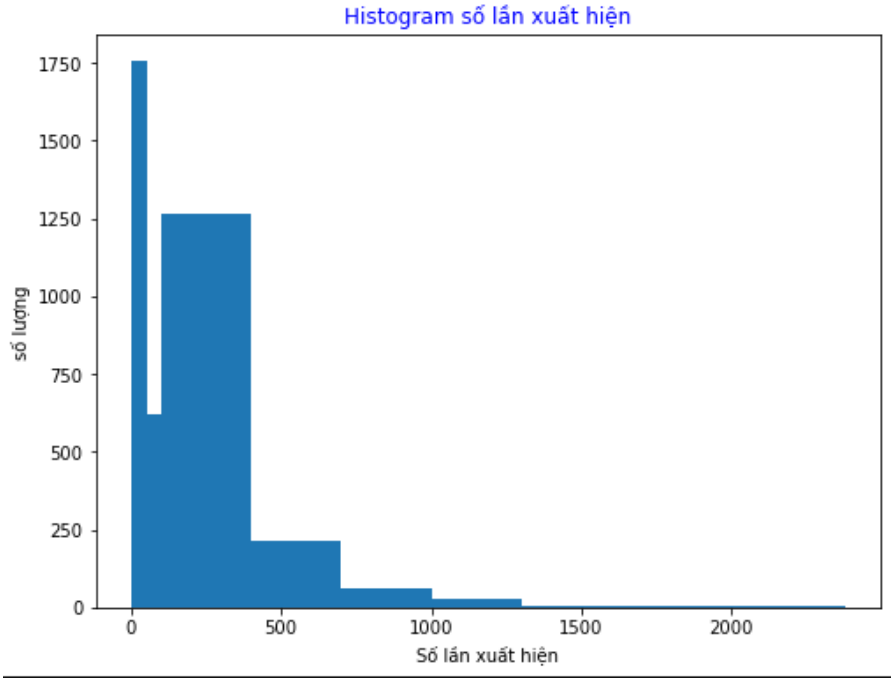
Theo bảng thông kế trên, ta thấy dữ liệu dữ liệu bị thiếu ở hai cột Description và CustomerID. Với:

+ Description có 1454 dòng null (chiếm 0.26% bộ dữ liệu)

+ CustomerID có 135080 dòng null (chiếm 24.92% bộ dữ liệu)



Biểu này thể hiện số lần xuất hiện của top 30 mặt hàng tính trên số lần xuất hiện trong hóa đơn. Nhưng khi nhìn vào bảng dưới.



Các sản phẩm có số lần xuất hiện trong hóa đơn là ở top đầu có giá trị lớn nhưng số lượng là khá nhỏ, các sản phẩm chỉ xuất hiện một vài lần có nhiều hơn xuất hiện nhiều hơn nhưng so với tổng thể bộ dữ liệu thì không quá ảnh hưởng.

2.2 Cleaning data steps

Các bước tiền xữ lý bao gồm:

+ chuẩn hóa dữ liệu chuỗi ở hai cột StockCode và Description về dạng in hoa để tránh xãy ra sai sót ở các bước sau.

+ Chuyển đổi các chuỗi Description có giá trị NULL về chuỗi rỗng “” để tránh lỗi trong quá trình huấn luyện

+ Loại bỏ các dòng có giá trị có giá trị không hợp lí (tạm thời bỏ qua không xữ lý) như các dòng có UnitPrice hoặc Quantity nhỏ hơn hoặc bằng giá trị 0.

+ Loại bỏ các dòng có các giá trị StockCode trùng lặp. Sau khi thực hiện, mỗi sản phẩm chỉ còn xuất hiện trên 1 dòng duy nhất.

Sau khi tiền xữ lý, ta thu được tập dữ liệu như sau.

	StockCode	Description	UnitPrice	Country
0	85123A	WHITE HANGING HEART T-LIGHT HOLDER	2.55	United Kingdom
1	71053	WHITE METAL LANTERN	3.39	United Kingdom
2	84406B	CREAM CUPID HEARTS COAT HANGER	2.75	United Kingdom
3	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	3.39	United Kingdom
4	84029E	RED WOOLLY HOTTIE WHITE HEART.	3.39	United Kingdom

2.3 Model Design

2.3.1. What is the raw input of the model?

Raw input của model là Description của các sản phẩm.

2.3.2. Preprocessing and any feature engineering steps

+ Loại bỏ các stop word.

+ Lập ma trận TF-IDF

+ Tính tính ma trận similarity bằng công thức cosine

2.3.3. What are inputs/outputs in predicting?

- input: a StockCode

- output: top 5 stockCode tương tự với input

2.3.4. ML Algorithms you choose, and why?

- Mô hình được sữ dụng: content-based recommender system

- Mô tả: dựa trên Description của sản phẩm tìm điểm chung của các sản phẩm với nhau, sau đó recommend cho người dùng.

- Cách tính độ lỗi mô hình: chưa xác định được.

- Lý do chọn mô hình: Lựa chọn mô hình content-based do không tìm được các giá trị để liên kết các sản phẩm với sản phẩm, hoặc sản phẩm với người dùng. Nếu chỉ dùng các giá trị số lượng sản phẩm mua thì cảm thấy không đủ tin tưởng khi gợi ý các sản phẩm khác.

- Điểm mạnh mô hình: lựa chọn được các sản phẩm liên quan đến sản phẩm được chọn nếu có các từ liên quan trong description.

- Điểm yếu mô hình:

+ Không thể đề xuất các sản phẩm mới

+ Không thể đề xuất các sản phẩm không chứa Description

- Cách mở rộng mô hình: nếu có thể tính được độ hài lòng của khách hàng đối với mỗi sản phẩm được mua (rating) hoặc phân loại các sản phẩm theo từng mục từ lớn đến nhỏ thì có thể cải thiện được mô hình phân lớp từ đó cải thiện được mô hình dự đoán.

3. Scale & Productize

3.1 Design Blueprints for ML Pipelines

- Data ingestion: Dùng Apache Beam vì nó hỗ trợ tốt cho nhiều nền tảng và nhiều loại dữ liệu (cả pack và stream)

- Training: Có 3 nền tảng train dữ liệu nổi tiếng là TensorFlow, Pytorch và scikit-learn. Cá nhân cảm nhận thấy scikit-learn dễ tiếp cận hơn nhưng để đạt hiệu quả hơn thì nên dùng Pytorch.

- Evaluation: Hai hàm lỗi để đánh giá mô hình được dùng nhiều nhất là Mean-Squared Error và Cross-validation.

- Serving: Tiêu biểu nhất hiện nay ta có Docker để quản lí các phiên bản và hổ trợ khá tốt trong việc mở rộng mô hình.

3.2 Write code for ML Pipelines

Code dùng Apache beam để tiền xữ lý dữ liệu. Ở đây dùng để gom nhóm dữ liệu sản phẩm và tìm ra top những người mua sản phẩm đó.

Code chạy trên comment line với công thức

python -m <module path> --input <input file path> --output <output file path> --ntop <number>

Với module path là

- <input file path> là đường dẫn đến file dữ liệu dầu vào.

- <output file path> là đường dẫn đến file dữ liệu kết quả.

- <number> là top về số lượng các người mua sản phẩm.

VD: python -m task3 --input test.csv --output output.txt --ntop 20

Do quá trình code chưa hoàn thiện, các hàm tiền xữ lí chưa hoạt động được như yêu cầu nên code này không thể hoạt động với tập dữ liệu có các giá trị thiếu.

Em có gữi một tập dữ liệu test.csv, trích tập dữ gốc gồm 1000 dòng có đầy đủ các giá trị để chạy test.