TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

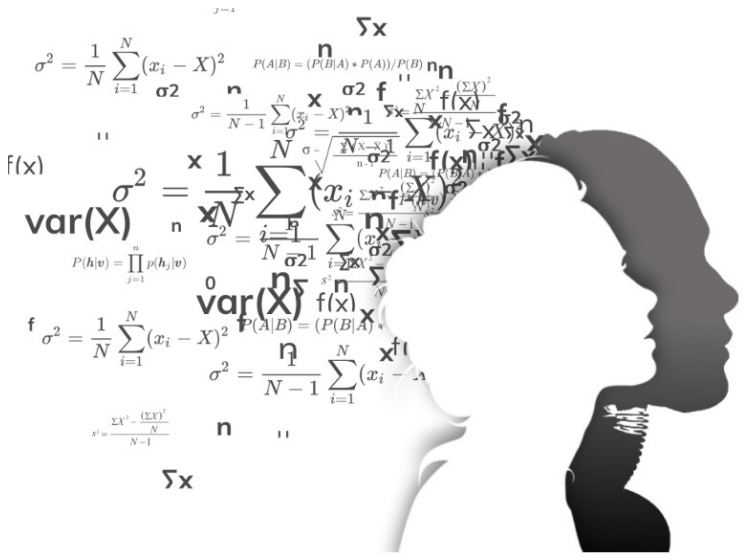
TRUNG TÂM TIN HỌC

Đồ án tốt nghiệp Data Science

Project 2: Recommender System

Tấn Huỳnh

Tuấn Trần



Năm

MỤC LỤC

1 Tổng Quan Bài Toán

2 EDA & Tiền Xử Lý Dữ Liệu

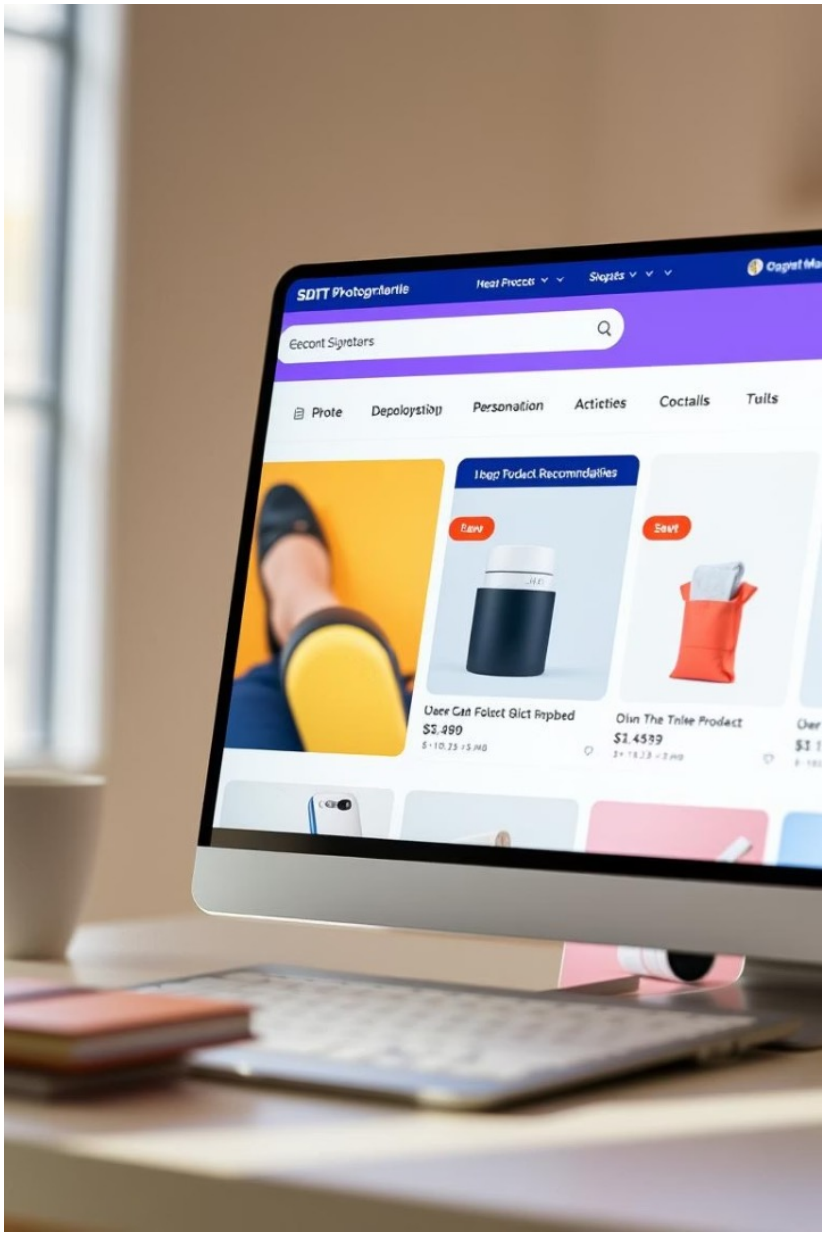
3 Model, Đánh giá và Deployment

Tổng Quan Bài Toán

🎯 **Bài toán: Gợi ý sản phẩm tương tự dựa trên nội dung văn bản – Shopee Recommender System**

📈 **Mục tiêu**

•Phân tích nội dung mô tả sản phẩm.
•Tính toán độ tương đồng giữa các sản phẩm.
•Gợi ý **top N sản phẩm tương tự nhất** với sản phẩm khách đang xem.
•**Tăng tỷ lệ tương tác, giữ chân và chuyển đổi mua hàng.**



Đề Bài

Yêu cầu

Xây dựng **hệ thống gợi ý sản phẩm cá nhân hóa** trên nền tảng TMĐT **Shopee**

Phân tích **đặc điểm sản phẩm** và **hành vi người dùng**

Model

Ứng dụng **Content-Based Filtering** (Gensim, Cosine Similarity)

Model

Ứng dụng **Collaborative Filtering** (ALS, Surprise)

Kết quả

Tăng **trải nghiệm cá nhân**, **tỷ lệ chuyển đổi**, và **doanh thu thông qua mô hình gợi ý sản phẩm**

EDA & Tiền Xử Lý Dữ Liệu: Chuẩn Bị Dữ Liệu

1 Kiểm Tra & Xử Lý

Kiểm tra dữ liệu, xử lý giá trị thiếu và ngoại lai.

2 Phân Tích

Thực hiện phân tích đơn biến
Ydata – profiling EDA

3 Chuẩn Hóa

Chuẩn hóa dữ liệu để đảm bảo tính đồng nhất.

EDA data Products_ThoiTrangNam_raw

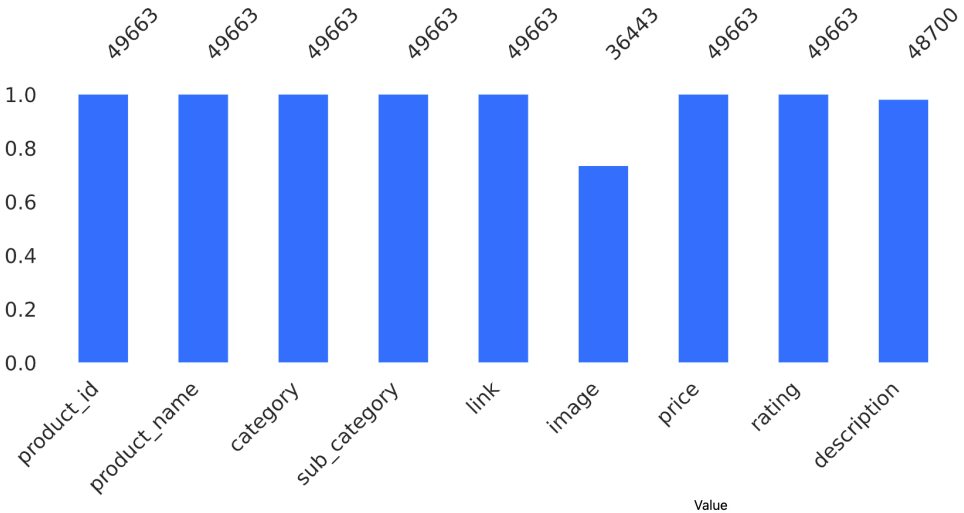


Image: record null => drop (outlier)

Word Cloud: các cụm từ/ từ xuất hiện quá nhiều

trong description

Value	Count	Frequency (%)
	429662	3.4%
hàng	215832	1.7%
nam	152689	1.2%
sản	145316	1.1%
áo	128330	1.0%
phẩm	121363	1.0%
và	109143	0.9%
không	108652	0.9%
có	106154	0.8%
size	105564	0.8%
Other values (147940)	11015315	87.2%

EDA data Products_ThoiTrangNam_raw

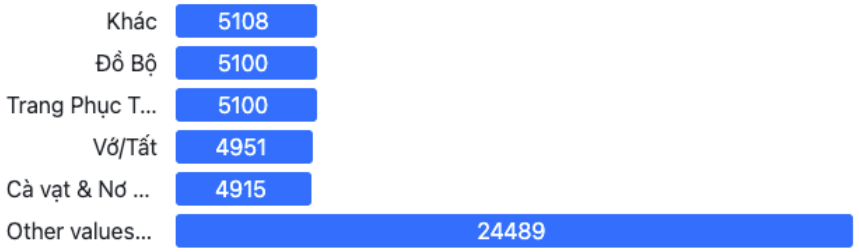
Combine vào Content

sub_category

Categorical

High correlation

Distinct	17
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.6 MiB



product_name

Text

Distinct	47103
Distinct (%)	94.8%
Missing	0
Missing (%)	0.0%
Memory size	13.4 MiB

EDA data Products_ThoiTrangNam_raw

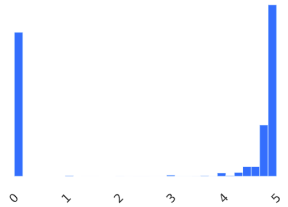
Drop outlier

rating

Real number (ℝ)

Zeros

Distinct	31	**Minimum**	0
Distinct (%)	0.1%	**Maximum**	5
Missing	0	**Zeros**	17964
Missing (%)	0.0%	**Zeros (%)**	36.2%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	3.0854902	**Memory size**	388.1 KiB



EDA data Products_ThoiTrangNam_rating_raw

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1024482 entries, 0 to 1024481
Data columns (total 4 columns):
 #   Column      Non-Null Count    Dtype
---  ------      --------------    -----
 0   product_id  1024482 non-null  int64
 1   user_id     1024482 non-null  int64
 2   user        1024482 non-null  object
 3   rating      1024482 non-null  int64
dtypes: int64(3), object(1)
memory usage: 31.3+ MB
```

	product_id	user_id	user	rating
0	190	1	karmakyun2nd	5
1	190	2	tranquangvinh_vv	5
2	190	3	nguyenquoctoan2005	5
3	190	4	nguyenthuyhavi	5
4	190	5	luonganh5595	5

	product_id	product_name	category	sub_category
0	190	Áo ba lỗ thun gân ,form body tôn dáng	Thời Trang Nam	Áo Ba Lỗ
1	191	Áo Ba Lỗ Nam Trắng Chất Cotton Siêu Mát, Siêu Đẹp	Thời Trang Nam	Áo Ba Lỗ
2	192	Áo Ba Lỗ Nam Tyasuo chất vải co dãn mát, không...	Thời Trang Nam	Áo Ba Lỗ
3	193	ÁO BA LỖ HÀNG VIỆT NAM 100% COTTON	Thời Trang Nam	Áo Ba Lỗ
4	194	Áo Thun Nam Thể Thao Ba Lỗ Mẫu Mới Siêu Đẹp (B...	Thời Trang Nam	Áo Ba Lỗ

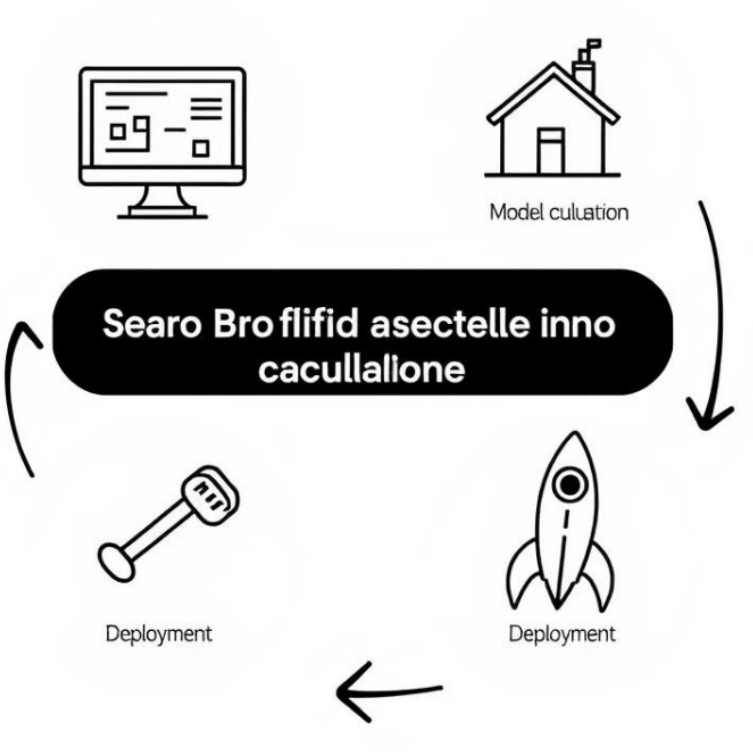
⇒ Merge 2 df để tạo df đủ thông tin cho mô hình Collaborative

Modeling & Evaluation: Xây Dựng Mô Hình

Chọn thuật toán

Xác định cụm

Đánh giá



<u>Thuật toán:</u>

🔷 **Content-Based Filtering**
→ Áp dụng **Cosine Similarity** để đo độ tương đồng giữa các sản phẩm
→ Sử dụng thư viện **Gensim** để xây dựng vector đặc trưng từ dữ liệu mô tả sản phẩm

🔷 **Collaborative Filtering**
→ Áp dụng **thuật toán ALS** từ thư viện pyspark.ml.recommendation
→ Kết hợp thư viện **SurPRISE** để xây dựng hệ thống gợi ý dựa trên hành vi người dùng

Content-Based Filtering

Cosine Similarity vs Gensim Similarity

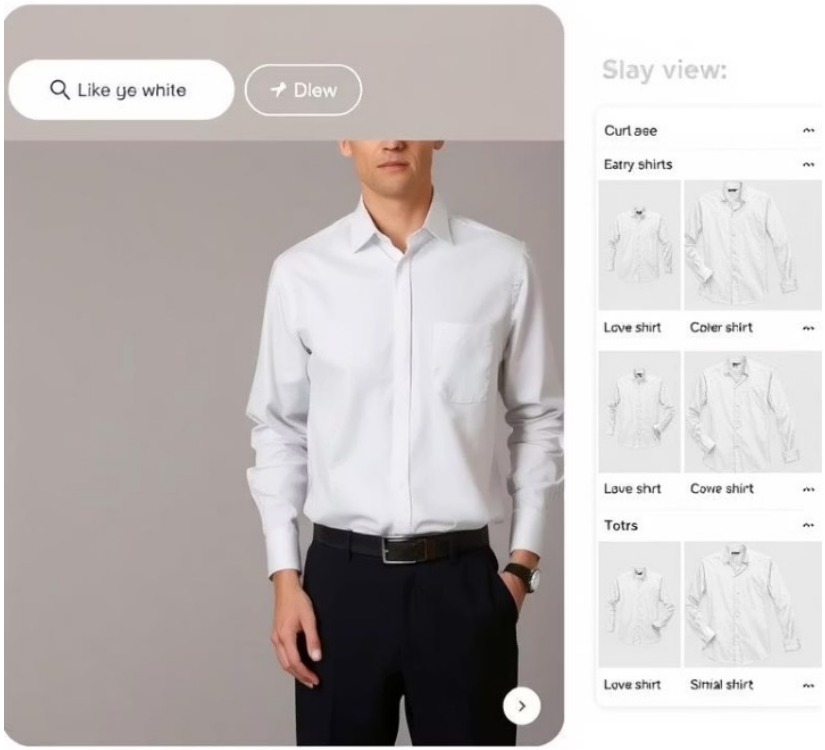
🔷 Content-Based Filtering

1 Khái niệm

Gợi ý sản phẩm tương tự dựa trên đặc điểm sản phẩm.

2 Ví dụ

Người dùng thích áo sơ mi trắng, gợi ý áo sơ mi cùng màu.



🔷 Content-Based Filtering >>

Gensim Similarity – Result

	0	1	2	3	4	5	6	7	8	9	...	1047	1048	1049	1050	1051	1052	105
0	1.000000	0.047288	0.017517	0.043782	0.017684	0.049587	0.071200	0.050366	0.059055	0.014825	...	0.091048	0.010314	0.031619	0.030166	0.020144	0.055378	0.0271
1	0.047288	1.000000	0.011097	0.027563	0.017957	0.237665	0.069957	0.065409	0.032102	0.100745	...	0.099126	0.010785	0.055959	0.019678	0.037477	0.056177	0.0308
2	0.017517	0.011097	1.000000	0.004053	0.006774	0.012774	0.012792	0.003434	0.025766	0.014499	...	0.029443	0.002316	0.010159	0.004499	0.011086	0.004571	0.0056
3	0.043782	0.027563	0.004053	1.000000	0.012169	0.007234	0.021855	0.019324	0.007869	0.031784	...	0.045934	0.014902	0.015152	0.002801	0.014391	0.014180	0.0194
4	0.017684	0.017957	0.006774	0.012169	1.000000	0.015166	0.004984	0.014783	0.015278	0.016852	...	0.030700	0.001134	0.005139	0.085302	0.073955	0.021549	0.0825
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1052	0.055378	0.056177	0.004571	0.014180	0.021549	0.049331	0.026057	0.050040	0.167881	0.064904	...	0.037890	0.009640	0.023259	0.023424	0.052293	1.000000	0.0423
1053	0.027176	0.030888	0.005657	0.019416	0.082512	0.030056	0.014224	0.033809	0.082365	0.018511	...	0.060653	0.020274	0.004486	0.072147	0.127225	0.042320	1.0000
1054	0.028078	0.031788	0.007855	0.022371	0.018898	0.088192	0.042177	0.028539	0.049632	0.028076	...	0.020847	0.048324	0.004350	0.026349	0.055931	0.060285	0.0634
1055	0.068813	0.056108	0.006669	0.012164	0.130988	0.011598	0.036674	0.041451	0.117765	0.016998	...	0.078367	0.021343	0.006831	0.097765	0.152373	0.055363	0.1569
1056	0.074931	0.025793	0.000022	0.080770	0.007633	0.000029	0.058632	0.001606	0.000093	0.031520	...	0.010828	0.000000	0.030418	0.014800	0.011460	0.029933	0.0511

Cosine – Result

	0	1	2	3	4	5	6	7	8	9	...	1047	1048	1049	1050	1051	1052	105
0	1.000000	0.085622	0.034910	0.075532	0.042669	0.082981	0.101466	0.092633	0.109675	0.032599	...	0.135925	0.012951	0.042997	0.068477	0.055191	0.111421	0.0580
1	0.085622	1.000000	0.029439	0.049175	0.046919	0.289348	0.101868	0.131898	0.060423	0.162810	...	0.158314	0.012735	0.106456	0.039943	0.081229	0.090638	0.0596
2	0.034910	0.029439	1.000000	0.014258	0.021472	0.025108	0.027877	0.013904	0.042231	0.032504	...	0.056566	0.007156	0.028727	0.013590	0.024970	0.018390	0.0189
3	0.075532	0.049175	0.014258	1.000000	0.027860	0.023146	0.038201	0.042851	0.040876	0.035239	...	0.061263	0.018721	0.026546	0.017826	0.032231	0.044933	0.0460
4	0.042669	0.046919	0.021472	0.027860	1.000000	0.034959	0.020596	0.034262	0.042874	0.036368	...	0.056159	0.003477	0.017075	0.119048	0.132395	0.049857	0.1271
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1052	0.111421	0.090638	0.018390	0.044933	0.049857	0.091316	0.053019	0.098423	0.272324	0.091671	...	0.074244	0.018085	0.043301	0.055165	0.106987	1.000000	0.0881
1053	0.058094	0.059633	0.018928	0.046025	0.127160	0.060561	0.034256	0.063523	0.132303	0.038207	...	0.104135	0.022450	0.015394	0.108792	0.195662	0.088169	1.0000
1054	0.058251	0.066854	0.021542	0.050965	0.038003	0.128366	0.059585	0.057878	0.091437	0.038090	...	0.053537	0.062596	0.016392	0.048803	0.094219	0.099565	0.1139
1055	0.108066	0.095814	0.022964	0.026775	0.186697	0.037630	0.066073	0.071324	0.153207	0.039601	...	0.120524	0.021602	0.024487	0.177569	0.218706	0.097947	0.2466
1056	0.096273	0.042995	0.007325	0.105998	0.018355	0.009224	0.073669	0.009942	0.010567	0.043487	...	0.034106	0.000000	0.047710	0.030816	0.027933	0.047470	0.0712

🔷 **Content-Based Filtering >>**

Gensim Similarity – Result

	product_id	product_name	price	rating	category	sub_category	similarity_score
45317	17605	Combo 20 Đôi Tất Nam Cổ Dài Uni Nhật Bản Giá Rẻ | Vớ Nam Uni	209000.0	4.9	Thời Trang Nam	Vớ/Tất	0.378132
47599	173004	[được chọn màu] tất nam cổ ngắn hàng xuất Nhật	8500.0	5.0	Thời Trang Nam	Vớ/Tất	0.336519
45931	171219	Combo 5 đôi Tất Nhật Nam Cao cấp, Khử Mùi, Siêu bền, Co giãn tốt không bai xù - SetT	75000.0	4.6	Thời Trang Nam	Vớ/Tất	0.239429

Cosine – Result

```
23749                                                  Hộp 4 Sịp Đùi Boxer Thông Hơi Cao Cấp Dành Cho Nam
23031                                        Quần lót ARISTINO-boxer- cotton thấm hút mồ hôi (abx037- 036-1603)
22565    Quần Lót Nam Thoáng Mát Nam Tính Quần Boxer Cotton Thấm Hút Mồ Hôi Quần Sịp Nam Thun Lạnh Co Giãn 4 Chiều Cao Cấp GALIO
Name: product_name, dtype: object
```

🔷 **Content-Based Filtering >>**

	Cosine Top Products	Gensim Top Products
0	Bộ Đồ Ngủ 2 Món Xinh Xắn Thời Trang Cho Cặp Đôi	Bộ Đồ Ngủ 2 Món Xinh Xắn Thời Trang Cho Cặp Đôi
1	Bộ Đồ Cộc Pyjama Nam Nữ Vải Lụa , Bộ đồ ngủ nam nữ họa tiết cao cấp 2 mẫu HULikKing4119	Bộ Đồ Cộc Pyjama Nam Nữ Vải Lụa , Bộ đồ ngủ nam nữ họa tiết cao cấp 2 mẫu HULikKing4119
2	Bộ đồ ngủ Pijama lụa Satin sang trọng cho các cặp đôi - Bộ đồ đôi nam nữ (hàng có sẵn)	Bộ Ngủ pijama Nam cao cấp thời thượng
3	Bộ Ngủ pijama Nam cao cấp thời thượng	Bộ đồ ngủ Pijama lụa Satin sang trọng cho các cặp đôi - Bộ đồ đôi nam nữ (hàng có sẵn)
4	Áo đôi áo cặp - CAO CẤP - Đồ đôi nam nữ đẹp Set váy sơ mi đôi phong cách Hàn Quốc	Áo Khoác Chống Nắng Adidas Chất Lượng Cao Cho Cặp Đôi

Sample_index trực quan => Cosine

Collaborative Filtering

SVD vs ALS

🔷 Collaborative Filtering

1 Dựa trên

Hành vi người dùng.

2 User-based CF

Gợi ý dựa trên người dùng tương tự.

3 Item-based CF

Gợi ý dựa trên sản phẩm tương tự.

Collaborative Filtering >> SVD model

```
model = SVD(n_factors=25, n_epochs=20, lr_all=0.025, reg_all=0.16)
```

Result:

```
RMSE: 0.8664
```

	product_id	product_name	sub_category	predicted_rating
0	1957	Áo ba lỗ đi biển đi du lịch màu sắc tươi tắn, ...	Áo Ba Lỗ	5
1	1958	Áo đông xuân nam loại 1, áo đông xuân nam hàng...	Áo Ba Lỗ	5
2	1970	Áo ba lỗ nam, áo tanktop sát nách in chữ RUNNI...	Áo Ba Lỗ	5
3	1992	Áo ba lỗ nam **LION**, chất liệu thoáng mát ...	Áo Ba Lỗ	5
4	19103	Áo ba lỗ thể thao sát nách nam tanktop giá rẻ ...	Áo Ba Lỗ	5
5	19117	Áo tanktop nam nữ in hình City Cycle - áo ba l...	Áo Ba Lỗ	5
6	19118	Áo nam After All Tanktop destroy, đục lỗ, màu ...	Áo Ba Lỗ	5
7	19129	Áo thun 3 lỗ nam chất đẹp co giãn 4 chiều,Áo t...	Áo Ba Lỗ	5
8	19135	Áo thun ba lỗ cho nam	Áo Ba Lỗ	5
9	19149	Áo ba lỗ nam tập gym, chơi thể thao, chạy bộ...	Áo Ba Lỗ	5

Deployment & Ứng Dụng: Triển Khai Giải Pháp

1

Công cụ model

- Content based: Cosine
- Collaborative: Surprise SVD

2

Phát triển giao diện người dùng GUI

- Content based: Cosine

Content-based

Tiêu chí	Cosine	Gensim
Độ rõ ràng	✅ Rất cao (phù hợp khi so sánh top-n)	Tốt nhưng phân tán hơn
Hiệu suất xử lý	✅ Nhanh	Chậm hơn nếu dữ liệu lớn
Mở rộng hệ thống	Trung bình	✅ Linh hoạt hơn với mô hình lớn
Mức độ liên kết ngữ nghĩa	Chặt hơn	✅ Phản ánh tốt hơn

Triển khai

- Sử dụng Consine triển khai GUI

Sơ Đồ Hệ Thống Gợi Ý GUI

1

Input

Người dùng + Dữ liệu lịch sử.

2

Mô hình

Content-Based Filtering, Collaborative Filtering.

3

Output

Danh sách sản phẩm gợi ý.

Kết Luận và Kinh nghiệm

Chúng ta đã đi qua quy trình Data Science trong bài toán hệ thống gợi ý

🔍 Giá trị đạt được:

Hiểu rõ quy trình xây dựng hệ thống gợi ý

Thành thạo đánh giá mô hình

Rút ra ưu nhược điểm từng mô hình và ứng dụng phù hợp

Nắm được cách thức triển khai đến người dùng cuối

🤝 Kỹ năng phát triển:

Làm việc nhóm hiệu quả, phân tích dữ liệu thực tế

Trình bày rõ ràng kết quả bằng hình ảnh và biểu đồ

🎓 **Cảm ơn Cô Phương và các bạn đã lắng nghe!**