FINAL REPORT: PROJECT_02

AGODA RECOMMENDATION SYSTEM

DL07_K306_PHAM NGOC TRONG_TRAN DINH HUNG

submited: 06/09/2025

AGENDA

01. Business Objective

02. Report Outline

03. Data Acquirement

04. Data Analysis

05. Project Structure

06. Data Workflow - Team task assignment

07. Content-Based Filtering : gensim

08. Content-Based Filtering : cosine-similarity

09. Collaborative Filtering : ALS

10. Hotel Insights & Analytics

11. Comparison

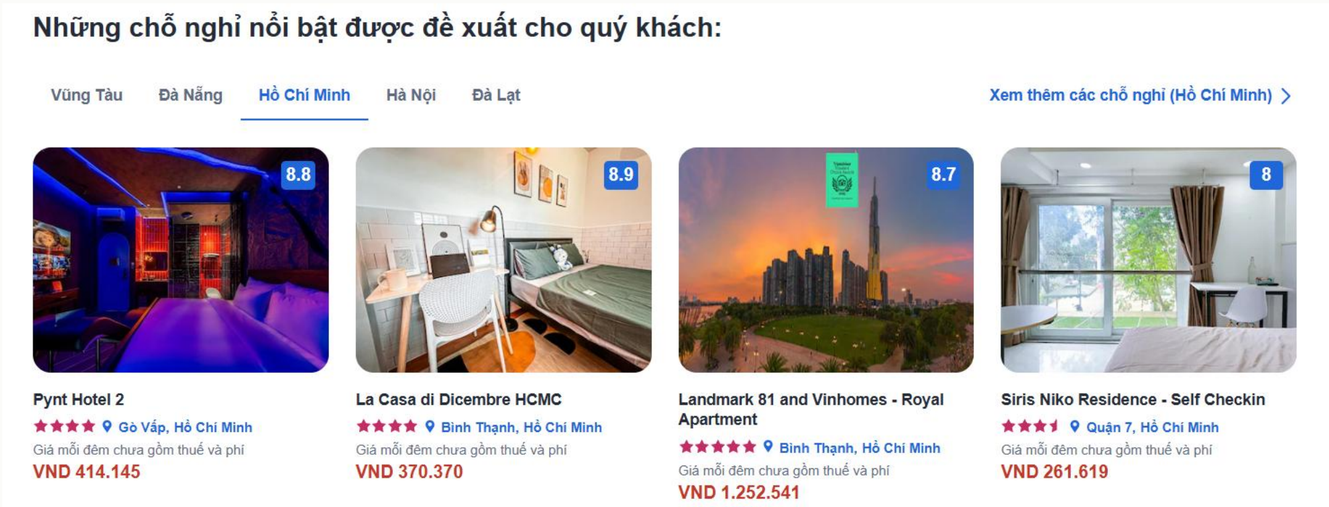
12. Our team

13.

14.

01. Business Objective

Agoda là một trang web đặt phòng trực tuyến có trụ sở tại Singapore, được thành lập vào năm 2005, thuộc sở hữu của Booking Holdings Inc,. Agoda chuyên cung cấp dịch vụ đặt phòng khách sạn, căn hộ, nhà nghỉ và các loại hình lưu trú trên toàn cầu. Trang web này cho phép người dùng tìm kiếm, so sánh và đặt chỗ ở với mức giá ưu đãi.



- Giả sử Agoda chưa triển khai hệ thống Recommender System giúp đề xuất khách sạn / resort phù hợp tới người dùng và bạn được yêu cầu triển khai hệ thống này, bạn sẽ làm gì?
- Chủ khách sạn đặt trên Agoda muốn nắm rõ insight dựa trên thông tin khách hàng, bạn sẽ đem đến cho họ những gì?

02. Recommendation System
Report outline

Content-based Filtering
- Gensim
- Cosine similarity

Collaborative Filtering
- Chuẩn hóa dataset: user-item-weight
- Thuật toán ALS (spark)
- Đánh giá RMSE
- Gợi ý cho khách hàng

1 & 2

3

- **Đề xuất cho người dùng:**
  - sử dụng Content-based Filtering
    - (1) Sử dụng genism
    - (2) Sử dụng coasin similarity
  - sử dụng Collaborative Filtering:
    - (3) sử dụng ALS (spark)
- **Đề xuất các insigh cho chủ khách sạn**

4

Insight cho Chủ khách sạn
- **Thông tin tổng quan khách sạn**
  - Tên, hạng sao, địa chỉ, điểm trung bình
- **Phân tích điểm mạnh & điểm yếu**
  - Dựa trên điểm chi tiết, số lượng & nội dung nhận xét
- **Thống kê khách hàng**
  - Quốc tịch, nhóm khách, xu hướng theo thời gian
- **Phân tích từ khóa trong nhận xét**
  - Từ khóa tích cực/tiêu cực
- **So sánh với trung bình hệ thống**
  - So sánh điểm từng tiêu chí với các khách sạn khác

03. Data Acquirement

hotel_comments.csv

- 80,314 records
- 13 attributes
- 437 hotels
- Date min: 15-11-2009
- Date max: 17-07-2024

	num	Hotel ID	Reviewer ID	Reviewer Name	Nationality	Group Name	Room Type	Stay Details	Score	Score Level	Title	Body	Review Date
0	1	1_1	1_1_1	MARIKO	Nhật Bản	Cặp đôi	Phòng Deluxe 2 Giường đơn Nhìn ra Biển	Đã ở 3 đêm vào Tháng 7 năm 2023	10,0	Trên cả tuyệt vời	Cao nhất!!"	Tôi đã ở cùng chủ nhân trong 4 đêm. Nhân viên ...	Đã nhận xét vào 30 tháng 7 2023
1	2	1_1	1_1_2	Hong	Việt Nam	Đi công tác	Phòng Deluxe 2 Giường đơn Nhìn ra Biển	Đã ở 1 đêm vào Tháng 9 năm 2022	10,0	Trên cả tuyệt vời	Tháng 8"	Lựa chọn Mường Thanh vì giá cả phù hợp. Đặt On...	Đã nhận xét vào 05 tháng 9 2022
2	3	1_1	1_1_3	Guai	Việt Nam	Cặp đôi	Deluxe Hướng biển giường đôi	Đã ở 1 đêm vào Tháng 6 năm 2024	9,2	Trên cả tuyệt vời	Du lịch tại Nha Trang"	Lần này đến với Nha Trang, tôi book phòng tại ...	Đã nhận xét vào 25 tháng 6 2024
3	4	1_1	1_1_4	Nghĩa	Việt Nam	Gia đình có em bé	Deluxe Hướng biển giường đôi	Đã ở 3 đêm vào Tháng 6 năm 2024	8,8	Tuyệt vời	Du lịch Nha Trang tại Mường Thanh"	Hôm đi đến lúc về thì mọi thứ trong Khách sạn ...	Đã nhận xét vào 02 tháng 7 2024
4	5	1_1	1_1_5	Duc	Việt Nam	Cặp đôi	Deluxe 2 giường Hướng phố	Đã ở 1 đêm vào Tháng 6 năm 2024	9,2	Trên cả tuyệt vời	Ks tốt !"	Khách sạn có vị trí trung tâm và sát biển. Nhâ...	Đã nhận xét vào 16 tháng 6 2024

03. Data Acquirement

hotel_info.csv

- 740 records
- 14 attributes
- 740 hotels

	num	Hotel_ID	Hotel_Name	Hotel_Rank	Hotel_Address	Total_Score	Location	Cleanliness	Service	Facilities	Value_for_money	Comfort_and_room_quality	comments_count	Hotel_Description
0	1	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	5 sao trên 5	60 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam	8,8	9,4	8,9	8,9	8,7	8,7	8,3	1269	Khách sạn Mường Thanh Luxury Nha Trang - Nơi l...
1	2	1_2	ALPHA BIRD NHA TRANG	4 sao trên 5	51/19/37 Tue Tinh St, Loc Tho Ward, Nha Trang,...	7,7	7,8	7,6	8,1	7,5	8,1	NaN	337	ALPHA BIRD NHA TRANG - Khách sạn 4.0 sao tại N...
2	3	1_3	Khách sạn Aaron (Aaron Hotel)	3.5 sao trên 5	6Trần Quang Khải, Lộc Thọ, Nha Trang, Việt Nam...	8,5	8,9	8,7	8,8	8,1	8,5	NaN	300	Khách sạn Aaron - Nơi nghỉ dưỡng tuyệt vời tại...
3	4	1_4	Panorama Star Beach Nha Trang	5 sao trên 5	02 Nguyen Thi Minh Khai, Lộc Thọ, Nha Trang, V...	8,8	9,6	8,9	8,9	8,7	9,0	NaN	814	Panorama Star Beach Nha Trang - Một kỳ nghỉ tu...
4	5	1_5	Khách sạn Balcony Nha Trang (Balcony Nha Trang...	4 sao trên 5	98B/13 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam	8,4	8,5	8,7	8,5	8,3	8,6	8,7	294	Khách sạn Balcony Nha Trang - Nơi nghỉ dưỡng t...

04. Data EDA

** DataFrame after renaming columns, removing 'No information' values, applying astype, and extracting review date/month.

dataframe: 80,314 rows x 14 cols

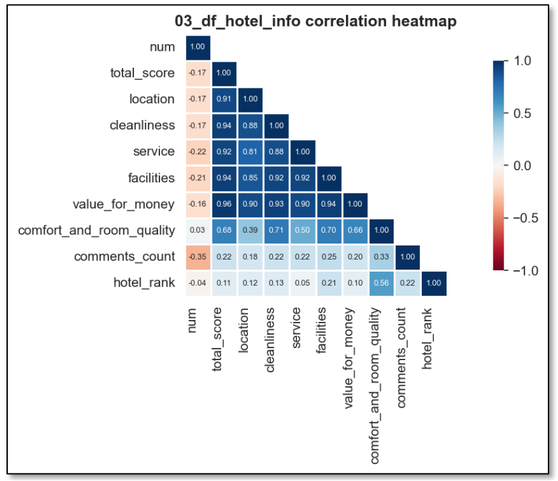
	num	hotel_id	reviewer_id	reviewer_name	nationality	group_name	room_type	score	score_level	title	body	review_date	stay_nights	stay_month
0	1	1_1	1_1_1	MARIKO	Nhật Bản	Cặp đôi	Phòng Deluxe 2 Giường đơn Nhìn ra Biển	10,0	Trên cả tuyệt vời	Cao nhất!!"	Tôi đã ở cùng chủ nhân trong 4 đêm. Nhân viên ...	2023-07-30	3	2023-07-01
1	2	1_1	1_1_2	Hong	Việt Nam	Đi công tác	Phòng Deluxe 2 Giường đơn Nhìn ra Biển	10,0	Trên cả tuyệt vời	Tháng 8"	Lựa chọn Mường Thanh vì giá cả phù hợp. Đặt On...	2022-09-05	1	2022-09-01
2	3	1_1	1_1_3	Guai	Việt Nam	Cặp đôi	Deluxe Hướng biển giường đôi	9,2	Trên cả tuyệt vời	Du lịch tại Nha Trang"	Lần này đến với Nha Trang, tôi book phòng tại ...	2024-06-25	1	2024-06-01

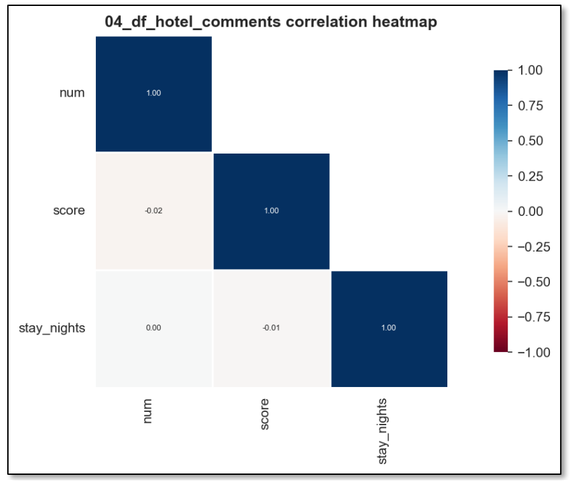
dataframe: 740 rows x 14 cols

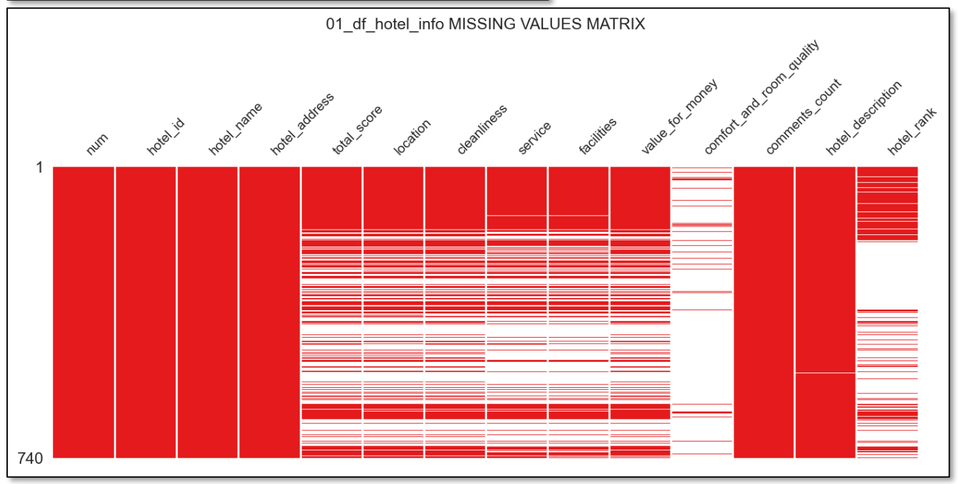
	num	hotel_id	hotel_name	hotel_address	total_score	location	cleanliness	service	facilities	value_for_money	comfort_and_room_quality	comments_count	hotel_description	hotel_rank
0	1	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	60 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam	8,8	9,4	8,9	8,9	8,7	8,7	8,3	1269	Khách sạn Mường Thanh Luxury Nha Trang - Nơi l...	5.0
1	2	1_2	ALPHA BIRD NHA TRANG	51/19/37 Tue Tinh St, Loc Tho Ward, Nha Trang,...	7,7	7,8	7,6	8,1	7,5	8,1	NaN	337	ALPHA BIRD NHA TRANG - Khách sạn 4.0 sao tại N...	4.0
2	3	1_3	Khách sạn Aaron (Aaron Hotel)	6Trần Quang Khải, Lộc Thọ, Nha Trang, Việt Nam...	8,5	8,9	8,7	8,8	8,1	8,5	NaN	300	Khách sạn Aaron - Nơi nghỉ dưỡng tuyệt vời tại...	3.5

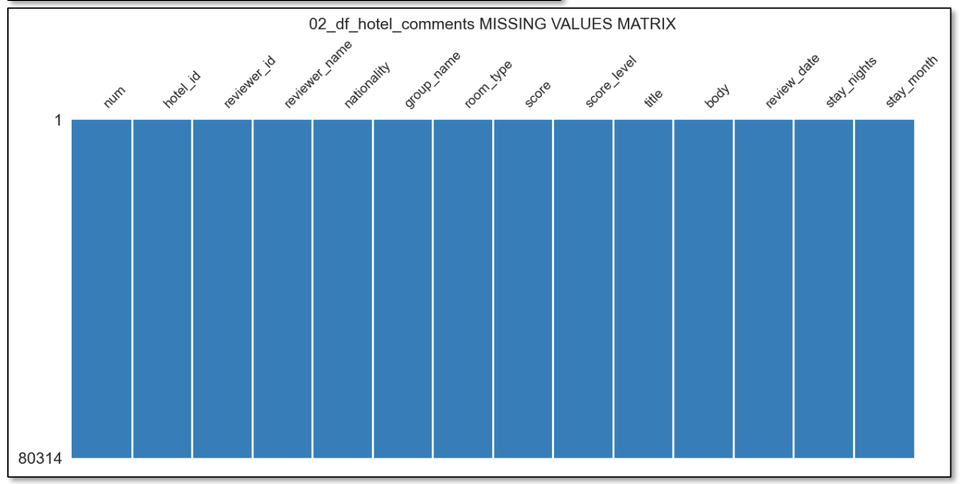
04. Data EDA

** DataFrame after renaming columns, removing 'No information' values, applying astype, and extracting review date/month.

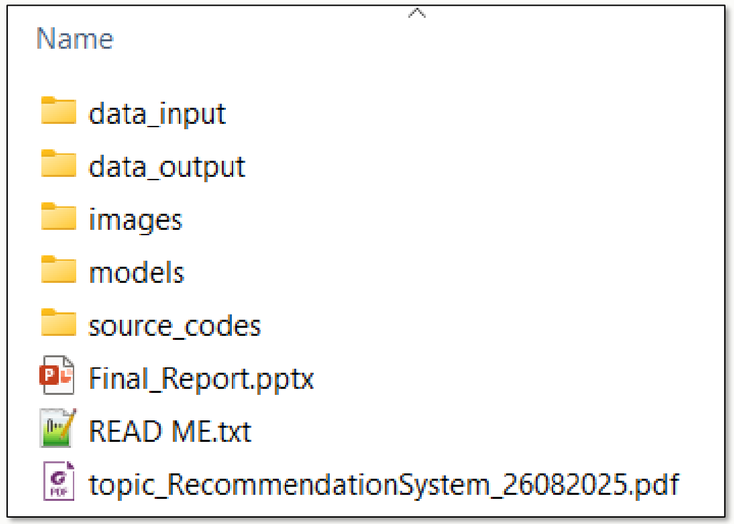


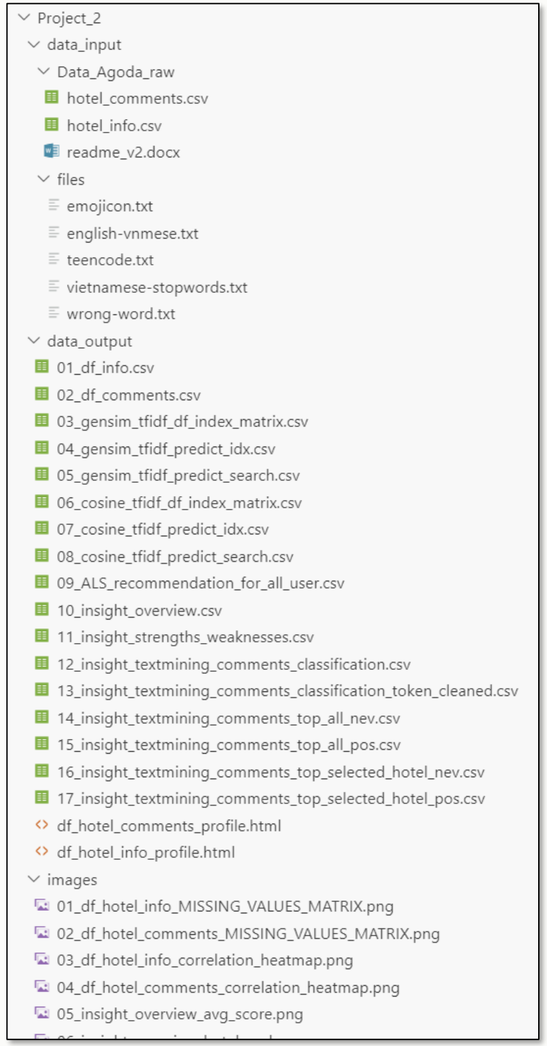


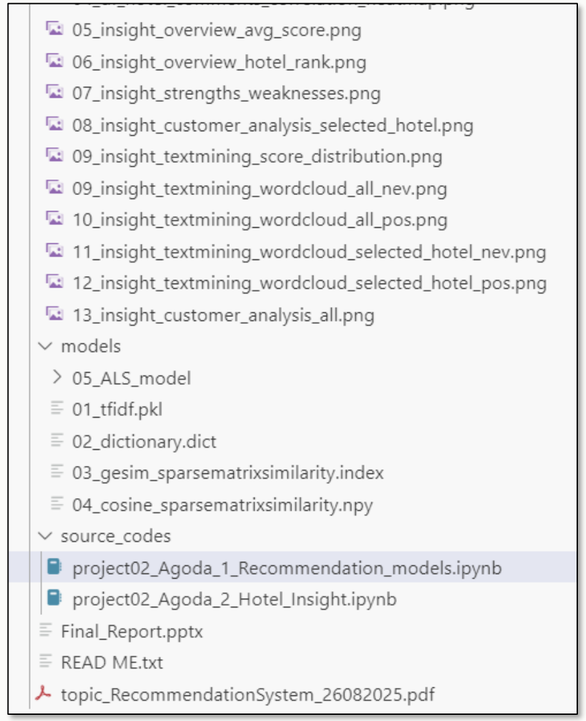




05. Project Structure







06. Data Workflow

	Collaborative filtering	Content-based filtering		Hotel Insights & Analytics
ETL • Rename columns • Check for Null / NaN values • Extract data (date, numeric values) • Validate records • Missing scores • Score distribution	**Preprocessing for ALS** • Data Preparation (User-Item Matrix)	**Preprocessing for gensim & TF-IDF** • Vectorization (TF-IDF / Gensim)	**Preprocessing for Cosine-similarity** • Similarity Computation (Cosine Similarity)	**Analytics** • Hotel Overview • Name, Star Rating, Location, Average Score • Strengths & Weaknesses Analysis • Based on detailed ratings & reviews • Customer Analysis • Nationality, Customer Segments, Temporal Trends • Text Mining on Reviews • Positive / Negative Keywords • Comparison with system-wide averages • Additional Insights & Recommendations
	ALS Modeling (spark) & Evaluate /w RMSE • Metrics: RMSE	**Modeling** • Implementation & Evaluation	**Modeling** • Implementation & Evaluation	
	Predit • Recommendations for all users	**Predit** Given hotel_id / hotel_idx: - Perform recommendation Given a search keyword: - Perform recommendation	**Predit** Given hotel_id / hotel_idx: - Perform recommendation Given a search keyword: - Perform recommendation	**Hotel insight**
	Compare • Comparison of Content-Based vs Collaborative Filtering			

06. Team task assignment

Phạm Ngọc Trọng
Project Lead

Trần Đình Hùng
Commercial Business Domain Advisor

config

```python
content_gem_cleaned = fn_clean_tokens(
    tokens=content_gem,
    dict_list=[emoji, teencode, engvie],
    stopword=stopword_vie,
    wrongword=wrongword,
    remove_number=False,
    remove_punctuation=True,
    remove_vie_tone=False,
    lower=True,
)
```

```python
top_N: int = 3 # top 3 khách sạn tương đồng nhất
info_cols_toshow = [
    "hotel_id",
    "hotel_name",
    "hotel_address",
    "total_score",
    "location",
    "cleanliness",
    "service",
    "facilities",
    "value_for_money",
    "comfort_and_room_quality",
    "comments_count",
    "hotel_description",
    "hotel_rank",
]
```

Chọn 'Khách sạn' trên index

```python
# seed
np.random.seed(random_state)

# Giả sử chọn 1 khách sạn:
search_hotel_idx = np.random.randint(0, len(hotel_lst))
search_hotel_id = hotel_lst.iloc[search_hotel_idx]["hotel_id"]
search_hotel_name = hotel_lst.iloc[search_hotel_idx]["hotel_name"]

print(f"hotel_idx  : {search_hotel_idx}")
print(f"hotel_id   : {search_hotel_id}")
print(f"hotel_name : {search_hotel_name}")
```

✓ 0.0s

```
hotel_idx  : 102
hotel_id   : 5_6
hotel_name : Khách sạn và Spa Florida Nha Trang (Florida Nha Trang Hotel And Spa)
```

Chọn 'Khách sạn' dựa trên từ khóa tìm kiếm

```python
# Giả sử nhập tìm kiếm như sau
search_input = "Khách sạn mới, phòng ngủ rộng, gần biển và phù hợp với nhu cầu du lịch cho gia đình"
```

✓ 0.0s

07. gensim

data_output/

03_gensim_tfidf_df_index_matrix.csv

	0	1	2	3	4	5	6	7	8	9	...	730	731	732	733	734	735	736	737	738	
0	1.000001	0.717552	0.667657	0.608054	0.678060	0.660585	0.593786	0.567430	0.542096	0.225440	...	0.149598	0.578201	0.172201	0.465260	0.138731	0.130643	0.060227	0.062263	0.171666	0.5
1	0.717552	1.000000	0.729511	0.656515	0.705651	0.670575	0.637486	0.598590	0.578906	0.240527	...	0.147532	0.622692	0.158801	0.494004	0.143008	0.146073	0.064915	0.071014	0.184164	0.5
2	0.667657	0.729511	1.000000	0.633033	0.702865	0.694385	0.592685	0.594441	0.573340	0.222699	...	0.157252	0.616460	0.170359	0.500444	0.145348	0.136219	0.061876	0.084173	0.189228	0.5
3	0.608054	0.656515	0.633033	1.000000	0.644842	0.646200	0.521426	0.508086	0.524830	0.194606	...	0.117412	0.558084	0.153658	0.447650	0.129045	0.150817	0.063732	0.072496	0.159725	0.5
4	0.678060	0.705651	0.702865	0.644842	1.000000	0.656860	0.570696	0.570811	0.558600	0.226511	...	0.148136	0.560243	0.170531	0.453021	0.140242	0.153757	0.067606	0.089793	0.206425	0.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
735	0.130643	0.146073	0.136219	0.150817	0.153757	0.118529	0.099580	0.125982	0.123746	0.275415	...	0.440905	0.121638	0.310945	0.124536	0.249698	1.000000	0.239623	0.252172	0.283176	0.1
736	0.060227	0.064915	0.061876	0.063732	0.067606	0.052402	0.053265	0.049562	0.050451	0.151760	...	0.164424	0.057865	0.174235	0.074846	0.153125	0.239623	1.000000	0.196072	0.224557	0.0
737	0.062263	0.071014	0.084173	0.072496	0.089793	0.065935	0.070539	0.049631	0.066621	0.243010	...	0.183539	0.075824	0.304928	0.053376	0.206151	0.252172	0.196072	1.000000	0.306382	0.0
738	0.171666	0.184164	0.189228	0.159725	0.206425	0.149769	0.149768	0.151456	0.159341	0.382742	...	0.272559	0.147373	0.409776	0.145553	0.298313	0.283176	0.224557	0.306382	1.000000	0.1
739	0.553756	0.580339	0.571479	0.521817	0.530957	0.534759	0.513152	0.501152	0.503360	0.202424	...	0.135650	0.538165	0.117360	0.521727	0.114696	0.157067	0.057689	0.066434	0.142095	1.0

740 rows × 740 columns

07. gensim

7.1.1 Search bằng index

```python
# top N khách sạn tương đồng với Khách sạn đang chọn
top_indices_c1 = df_index_matrix.loc[search_hotel_idx].drop(search_hotel_idx, errors='ignore').nlargest(top_N).index.tolist()
print(f"top indices idx: {top_indices_c1}")

# predict
df_predict_c1 = df_hotel_info.loc[df_hotel_info.index[top_indices_c1], info_cols_toshow]
fn_show(df=df_predict_c1, n=None)

# save csv
fn_save_csv(df=df_predict_c1, folder_path = path + out_path, file_name = '04_gensim_tfidf_predict_idx.csv', header=True, index=True)
```

✓ 0.0s Python

```
top indices idx: [29, 624, 1]
dataframe: 3 rows x 13 cols
```

	hotel_id	hotel_name	hotel_address	total_score	location	cleanliness	service	facilities	value_for_money	comfort_and_room_quality	comments_count	hotel_description	hotel_rank
29	1_30	Areca Hotel Nha Trang	46A đường Hoàng Văn Thụ, phường Vạn Thạnh...	8.9	9.0	9.2	9.1	8.9	9.2	NaN	160	Khám phá Areca Hotel Nha Trang - Khách sạn 4.0...	4.0
624	4_23	Khách sạn Havana Nha Trang (Havana Nha Trang ...	38 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam, 650000	9.0	9.5	9.1	9.2	9.1	9.2	8.8	1338	Khách sạn Havana Nha Trang - Nơi lưu trú hoàn...	5.0
1	1_2	ALPHA BIRD NHA TRANG	51/19/37 Tue Tinh St, Loc Tho Ward, Nha Trang,...	7.7	7.8	7.6	8.1	7.5	8.1	NaN	337	ALPHA BIRD NHA TRANG - Khách sạn 4.0 sao tại N...	4.0

```
data_output/
------------------------------
04_gensim_tfidf_predict_idx.csv
```

07. gensim

7.1.2 Search bằng từ khóa

✓ `# từ khóa` ···

```
search: 'Khách sạn mới, phòng ngủ rộng, gần biển và phù hợp với nhu cầu du lịch cho gia đình'
```

✓ `# xử lý token của search_input` ···

```
top indices idx: [326, 718, 159]
dataframe: 3 rows x 13 cols
```

	hotel_id	hotel_name	hotel_address	total_score	location	cleanliness	service	facilities	value_for_money	comfort_and_room_quality	comments_count	hotel_description	hotel_rank
29	1_30	Areca Hotel Nha Trang	46A đường Hoàng Văn Thụ, phường Vạn Thạnh...	8.9	9.0	9.2	9.1	8.9	9.2	NaN	160	Khám phá Areca Hotel Nha Trang - Khách sạn 4.0...	4.0
624	4_23	Khách sạn Havana Nha Trang (Havana Nha Trang ...	38 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam, 650000	9.0	9.5	9.1	9.2	9.1	9.2	8.8	1338	Khách sạn Havana Nha Trang - Nơi lưu trú hoàn...	5.0
1	1_2	ALPHA BIRD NHA TRANG	51/19/37 Tue Tinh St, Loc Tho Ward, Nha Trang,...	7.7	7.8	7.6	8.1	7.5	8.1	NaN	337	ALPHA BIRD NHA TRANG - Khách sạn 4.0 sao tại N...	4.0

```
data_output/
------------------------------
05_gensim_tfidf_predict_search.csv
```

08. consine

```
models/
------------------------------
04_cosine_sparsematrixsimilarity.npy
data_output/
------------------------------
06_cosine_tfidf_df_index_matrix.csv
```

```
1  df_index_matrix_cosine
```

✓ 0.0s Open 'df_index_matrix_cosine' in Data Wrangler Python

	0	1	2	3	4	5	6	7	8	9	...	730	731	732	733	734	735	736	737	738	
0	1.000000	0.839117	0.800056	0.765528	0.817198	0.775473	0.781804	0.761340	0.726494	0.403454	...	0.248487	0.729465	0.296210	0.687484	0.313315	0.189167	0.123167	0.127557	0.339244	0.70
1	0.839117	1.000000	0.827591	0.790088	0.822653	0.790618	0.799456	0.765421	0.748388	0.407882	...	0.241897	0.759351	0.280746	0.695304	0.308539	0.192132	0.126951	0.138115	0.345909	0.71
2	0.800056	0.827591	1.000000	0.784834	0.824884	0.814482	0.763680	0.765700	0.741231	0.384377	...	0.255998	0.758478	0.286450	0.693168	0.309565	0.185908	0.115669	0.154124	0.352752	0.70
3	0.765528	0.790088	0.784834	1.000000	0.793094	0.793801	0.720629	0.703060	0.718386	0.368283	...	0.224532	0.717590	0.274103	0.656268	0.290838	0.208265	0.119591	0.155801	0.326945	0.69
4	0.817198	0.822653	0.824884	0.793094	1.000000	0.784874	0.757880	0.746481	0.737853	0.394788	...	0.254339	0.718640	0.292636	0.670822	0.308213	0.208086	0.125076	0.160063	0.374791	0.67
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
735	0.189167	0.192132	0.185908	0.208265	0.208086	0.166372	0.163382	0.180244	0.180801	0.333763	...	0.550029	0.153528	0.349782	0.201276	0.331375	1.000000	0.324734	0.228884	0.328037	0.19
736	0.123167	0.126951	0.115669	0.119591	0.125076	0.106792	0.118203	0.103065	0.106930	0.237563	...	0.204015	0.091943	0.232896	0.139591	0.253644	0.324734	1.000000	0.179991	0.273637	0.11
737	0.127557	0.138115	0.154124	0.155801	0.160063	0.146024	0.129029	0.114430	0.143531	0.324726	...	0.234625	0.140164	0.356554	0.127130	0.326276	0.228884	0.179991	1.000000	0.368414	0.14
738	0.339244	0.345909	0.352752	0.326945	0.374791	0.305058	0.318961	0.315057	0.323776	0.595826	...	0.422223	0.287234	0.624414	0.324218	0.538594	0.328037	0.273637	0.368414	1.000000	0.29
739	0.702653	0.715042	0.705002	0.693814	0.679683	0.682650	0.696497	0.673769	0.674098	0.347926	...	0.242413	0.693881	0.228050	0.699891	0.262790	0.198917	0.111886	0.140454	0.297507	1.00

740 rows × 740 columns

08. consine

7.2.1 Search bằng index

```python
# top N khách sạn tương đồng với Khách sạn đang chọn
top_indices_c2 = df_index_matrix_cosine.loc[search_hotel_idx].drop(search_hotel_idx, errors='ignore').nlargest(top_N).index.tolist()
print(f"top indices idx: {top_indices_c2}")

# predict
df_predict_c2 = df_hotel_info.loc[df_hotel_info.index[top_indices_c2], info_cols_toshow]
fn_show(df=df_predict_c2, n=None)

# save csv
fn_save_csv(df=df_predict_c2, folder_path = path + out_path, file_name = '07_cosine_tfidf_predict_idx.csv', header=True, index=True)
```

✓ 0.0s Python

```
top indices idx: [29, 0, 34]
dataframe: 3 rows x 13 cols
```

	hotel_id	hotel_name	hotel_address	total_score	location	cleanliness	service	facilities	value_for_money	comfort_and_room_quality	comments_count	hotel_description	hotel_rank
29	1_30	Areca Hotel Nha Trang	46A đường Hoàng Văn Thụ, phường Vạn Thạnh...	8.9	9.0	9.2	9.1	8.9	9.2	NaN	160	Khám phá Areca Hotel Nha Trang - Khách sạn 4.0...	4.0
0	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	60 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam	8.8	9.4	8.9	8.9	8.7	8.7	8.3	1269	Khách sạn Mường Thanh Luxury Nha Trang - Nơi l...	5.0
34	2_5	Khách sạn Boutique Seasing (Seasing Boutique H...	Số 1 Củ Chi, Vĩnh Hải, Nha Trang, Việt Nam	8.9	8.5	9.2	9.2	8.9	9.1	9.2	139	Khách sạn Boutique Seasing - Nơi lý tưởng để k...	4.0

```
data_output/
------------------------------
07_cosine_tfidf_predict_idx.csv
```

08. consine

7.2.2 Search bằng từ khóa

```
1  # từ khóa
2  print(f"search: '{search_input}'")
```

✓ 0.0s Python

```
search: 'Khách sạn mới, phòng ngủ rộng, gần biển và phù hợp với nhu cầu du lịch cho gia đình'
```

✓ # Khởi tạo lại vectorizer và fit trên toàn bộ dữ liệu để đảm bảo tính nhất quán khi transform search_input ···

```
top indices idx: [641, 330, 683]
dataframe: 3 rows x 13 cols
```

	hotel_id	hotel_name	hotel_address	total_score	location	cleanliness	service	facilities	value_for_money	comfort_and_room_quality	comments_count	hotel_description	hotel_rank
641	17_14	Muong Thanh Vien Trieu Apartment Review Nha Trang	3 Phạm Văn Đồng, Mường Thanh Viễn Triều, Vĩnh ...	5.2	2.0	2.0	NaN	NaN	2.0	NaN	0	Mường Thanh Viễn Triều là khu tổ hợp gồm 4 tòa...	NaN
330	38_10	Căn hộ 70 m² 2 phòng ngủ, 2 phòng tắm riêng ở ...	Xương Huân, Nha Trang, Việt Nam	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0	CĂN HỘ ĐA TIỆN ÍCH PHÙ HỢP CHO GIA ĐÌNH & NHÓM...	NaN
683	11_3	Nhà riêng 400 m² 4 phòng ngủ, 5 phòng tắm riên...	Vạn Ninh, Nha Trang, Việt Nam	9.7	8.7	10.0	NaN	NaN	10.0	NaN	0	- Với không gian sân vườn rộng có chổ đậu xe ô...	3.0

```
data_output/
------------------------------
08_cosine_tfidf_predict_search.csv
```

09. ALS

```python
# data input ALS
# sử dụng "score" làm cột "rating"
df_hotel_comments_ALS = spark.createDataFrame(df_hotel_comments[df_hotel_comments["score"].notna()][["hotel_id", "reviewer_name", "score"]])

# show sample
print(f"Hotel comments original data: {df_hotel_comments.shape[0]:,} rows x {df_hotel_comments.shape[1]} cols")
print(f"ALS                  input data: {df_hotel_comments_ALS.count():,} rows x {len(df_hotel_comments_ALS.columns)} cols\n")
fn_show_spark(df=df_hotel_comments_ALS, n=n_rows, printSchema=True, truncate=False)
```

✓ 15.7s

```
Hotel comments original data: 80,314 rows x 14 cols
ALS                  input data: 80,314 rows x 3 cols

dataframe: 80,314 rows x 3 cols

root
 |-- hotel_id: string (nullable = true)
 |-- reviewer_name: string (nullable = true)
 |-- score: double (nullable = true)

+--------+-------------+-----+
|hotel_id|reviewer_name|score|
+--------+-------------+-----+
|1_1     |MARIKO       |10.0 |
|1_1     |Hong         |10.0 |
|1_1     |Guai         |9.2  |
+--------+-------------+-----+
```

09. ALS

✓ # distinct reviewers and hotels ···

```
Total numerator (ratings): 80,314
Total users              : 8,191
Total hotels             : 473
```

✓ # denominator ···

```
Total denominator (ratings matrix): 3,874,343
Sparsity: 97.93%
```

```python
# string indexer
indexer = StringIndexer(inputCol="hotel_id", outputCol="hotel_id_idx")
indexer_model = indexer.fit(df_hotel_comments_ALS)
data_indexed = indexer_model.transform(df_hotel_comments_ALS)

indexer_1 = StringIndexer(inputCol="reviewer_name", outputCol="reviewer_name_idx")
indexer_1_model = indexer_1.fit(data_indexed)
data_indexed = indexer_1_model.transform(data_indexed)

fn_show_spark(df=data_indexed, n=n_rows, truncate=False)
```

✓ 8.4s

```
dataframe: 80,314 rows x 5 cols

+--------+-------------+-----+------------+-----------------+
|hotel_id|reviewer_name|score|hotel_id_idx|reviewer_name_idx|
+--------+-------------+-----+------------+-----------------+
|1_1     |MARIKO       |10.0 |15.0        |1682.0           |
|1_1     |Hong         |10.0 |15.0        |34.0             |
|1_1     |Guai         |9.2  |15.0        |1929.0           |
+--------+-------------+-----+------------+-----------------+
only showing top 3 rows
```

09. ALS

```
# # ALS model
# als = ALS(
#     userCol="reviewer_name_idx",
#     itemCol="hotel_id_idx",
#     ratingCol="score",
#     coldStartStrategy="drop",
#     nonnegative=True
# )

# # evaluator
# evaluator = RegressionEvaluator(metricName="rmse", labelCol="score", predictionCol="prediction")

# # grid search
# paramGrid = (
#     ParamGridBuilder()
#     .addGrid(als.rank, [10, 25, 50])
#     .addGrid(als.regParam, [0.01, 0.05, 0.1])
#     .addGrid(als.maxIter, [10, 20])
#     .build()
# )

# # cross validation
# cv = CrossValidator(
#     estimator=als,
#     estimatorParamMaps=paramGrid,
#     evaluator=evaluator,
#     numFolds=3,             # 3-fold CV
#     parallelism=4           # chạy song song
# )

# # fit model
# cvModel = cv.fit(df_ALS_train)

# # best model
# bestModel = cvModel.bestModel
# print("Best rank:", bestModel.rank)
# print("Best regParam:", bestModel._java_obj.parent().getRegParam())
# print("Best maxIter:", bestModel._java_obj.parent().getMaxIter())
```

✓ 0.0s

```
# Best rank: 50
# Best regParam: 0.05
# Best maxIter: 20
```

✓ 0.0s

09. ALS

```python
# predictions
ALS_predictions = ALS_model.transform(df_ALS_test)
ALS_predictions.select(["hotel_id_idx", "reviewer_name_idx", "score", "prediction"]).show(n=n_rows)
```

✓ 4.1s Python

```
+------------+-----------------+-----+----------+
|hotel_id_idx|reviewer_name_idx|score|prediction|
+------------+-----------------+-----+----------+
|        27.0|           1580.0|  9.6|  9.577587|
|        63.0|            463.0| 10.0|  9.888889|
|         6.0|            496.0| 10.0|  9.949855|
+------------+-----------------+-----+----------+
only showing top 3 rows
```

```python
# evaluate
rmse = RegressionEvaluator(metricName="rmse", labelCol="score", predictionCol="prediction").evaluate(ALS_predictions)
print(f"Root Mean Squared Error (RMSE) = {rmse:.2f}")
```

✓ 3.3s Python

```
Root Mean Squared Error (RMSE) = 0.62
```

Nhận xét: RMSE

Chỉ số	Giá trị	Ý nghĩa
RMSE `sau tuning`	**0.62**	- Cải thiện đáng kể, mức khá mạnh cho collaborative filtering - Model dự đoán khá gần với hành vi user

09. ALS

Providing Recommendations: for all users

```
1  user_recs = ALS_model.recommendForAllUsers(numItems=top_N)
2  user_recs.show(n=n_rows, truncate=False)
```

1] ✓ 15.6s Python

```
+-----------------+------------------------------------------------------+
|reviewer_name_idx|recommendations                                       |
+-----------------+------------------------------------------------------+
|0                |[{427, 11.360981}, {435, 11.257358}, {433, 11.257358}] |
|1                |[{427, 11.238722}, {335, 11.0712595}, {323, 11.061553}]|
|2                |[{427, 11.312377}, {323, 11.049965}, {335, 11.036599}] |
+-----------------+------------------------------------------------------+
only showing top 3 rows
```

✓ `import pyspark.sql.functions as F ···`

```
1  print(f"new_user_recs: {new_user_recs_joined.count():,} rows x {len(new_user_recs_joined.columns)} cols\n")
2  new_user_recs_joined.printSchema()
3  new_user_recs_joined.select("reviewer_name", "recommendations").show(3, truncate=False)
```

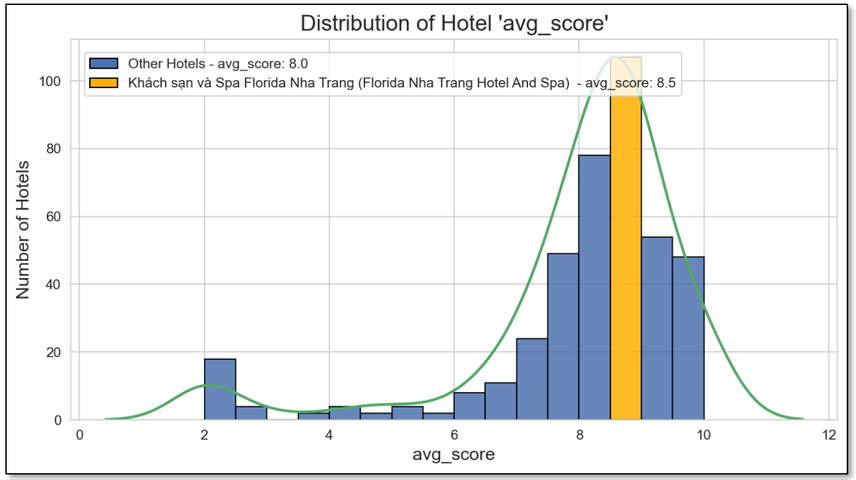
3] ✓ 33.1s Python

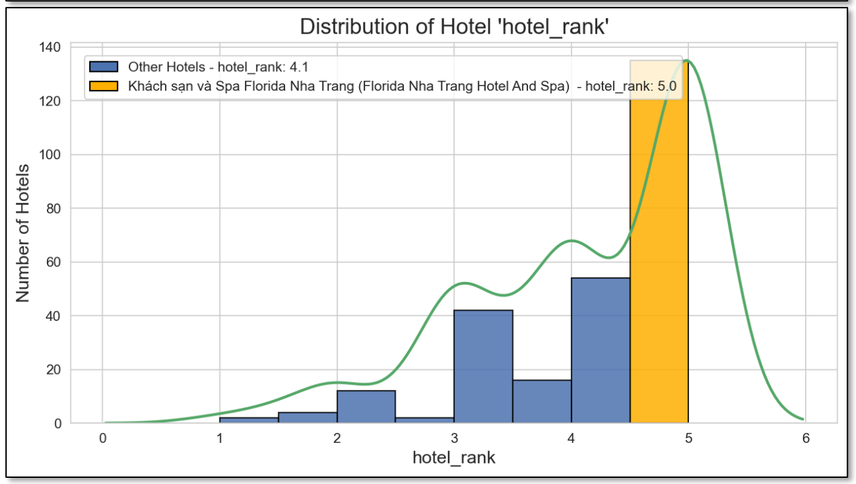
```
new_user_recs: 7,645 rows x 3 cols

root
 |-- reviewer_name_idx: integer (nullable = false)
 |-- recommendations: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- hotel_id: string (nullable = true)
 |    |    |-- rating: float (nullable = true)
 |-- reviewer_name: string (nullable = true)

+-------------+------------------------------------------------------------+
|reviewer_name|recommendations                                             |
+-------------+------------------------------------------------------------+
|Nguyễn       |[{18_30, 11.360981}, {26_11, 11.257358}, {25_16, 11.257358}]|
|Nguyen       |[{18_30, 11.238722}, {7_9, 11.0712595}, {12_13, 11.061553}] |
|Thanh        |[{39_17, 10.900191}, {18_30, 10.742319}, {16_14, 10.684663}]|
+-------------+------------------------------------------------------------+
```

10. Hotel Insights – Over view

Hotel Overview

```
{'hotel_id': '5_6',
 'hotel_name': 'Khách sạn và Spa Florida Nha Trang (Florida Nha Trang Hotel And Spa)',
 'hotel_address': '66 Quang Trung, Lộc Thọ, Nha Trang, Việt Nam, 650000',
 'hotel_rank': 5.0,
 'avg_score': 8.5}

data_output/
------------------------------
10_insight_overview.csv
```





10. Hotel Insights – strengths & weaknesses

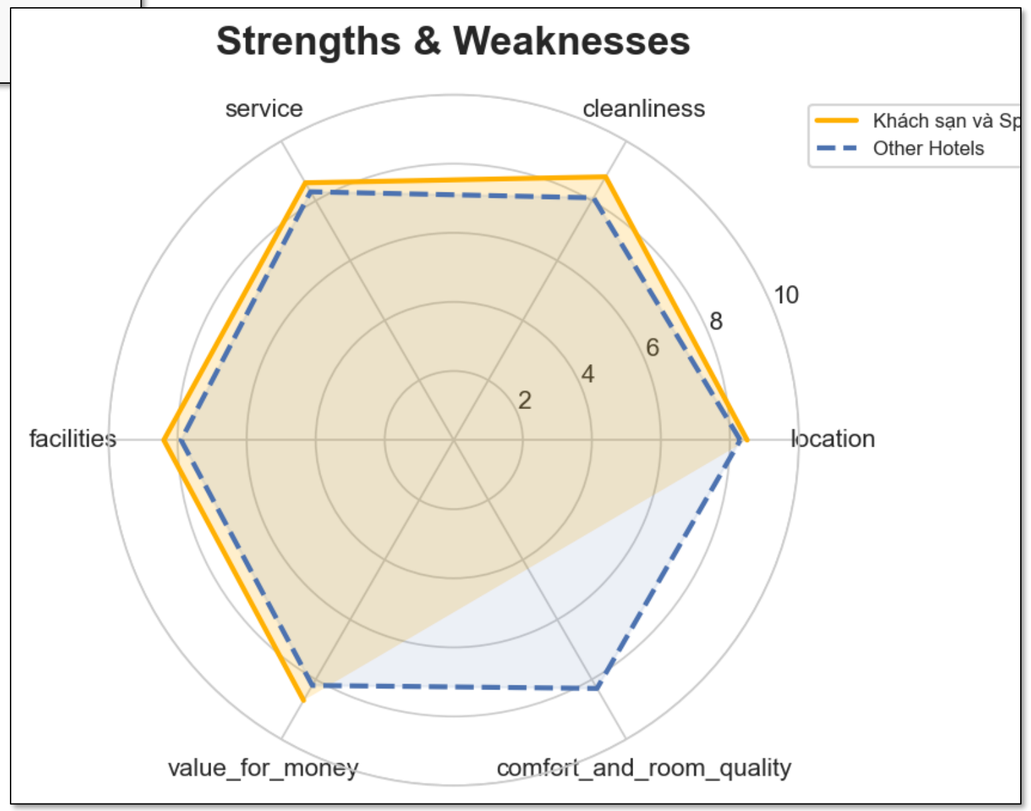
```
rating_cols = ["location","cleanliness","service","facilities","value_for_money","comfort_and_room_quality"]
score_classify_dict = {
    "Strength": 8.5,
    "Neutral": 7.5,
}
```

Strengths & Weaknesses | Khách sạn và Spa Florida Nha Trang (Florida Nha Trang Hotel And Spa)

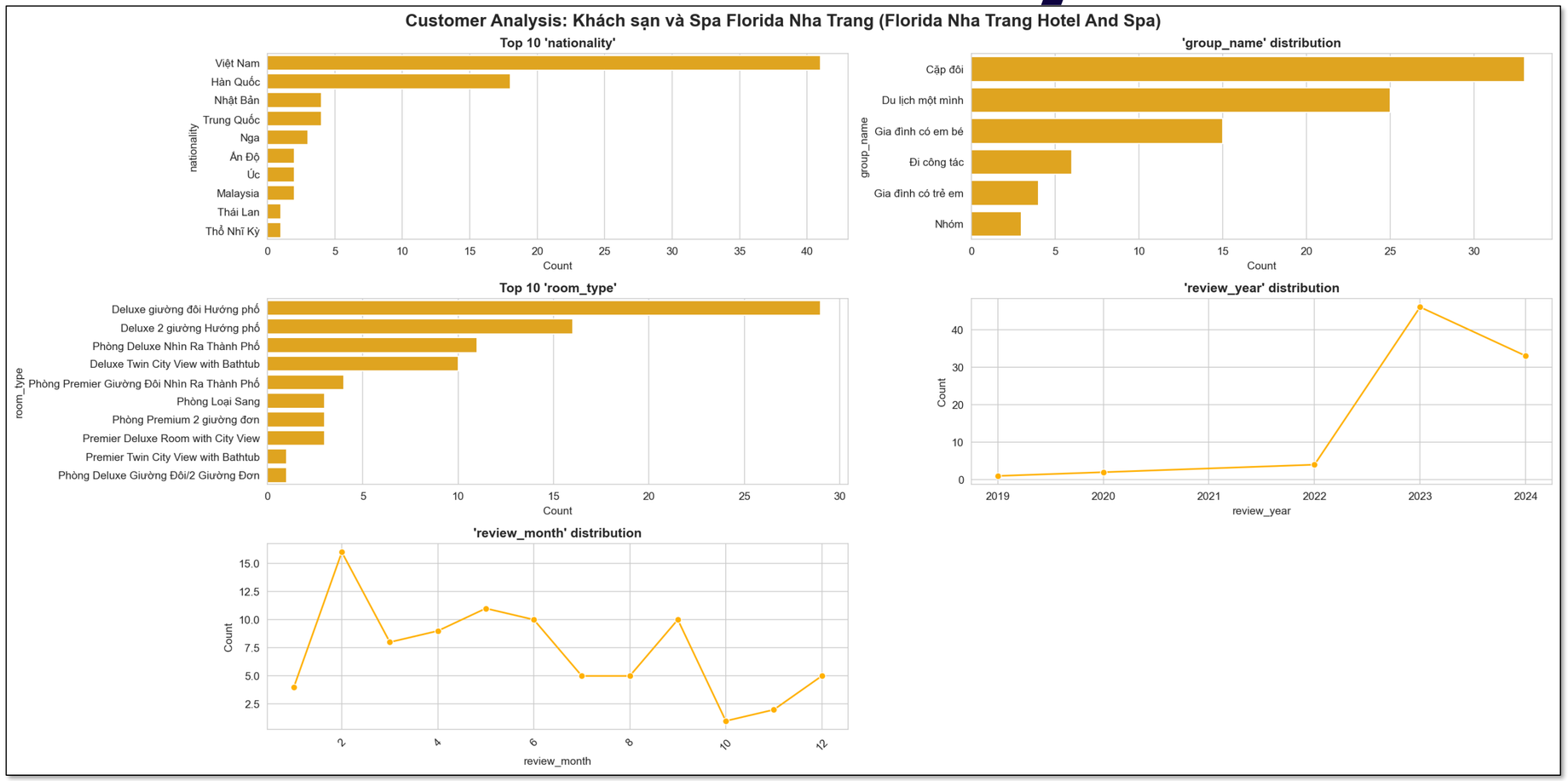
dataframe: 6 rows x 6 cols

	attr	selected_hotel	all_mean	diff	top_percent	score_classify
0	location	8.5	8.3	0.2	63.6	Strength
1	cleanliness	8.8	8.1	0.7	44.8	Strength
2	service	8.6	8.3	0.3	59.2	Strength
3	facilities	8.4	7.9	0.5	45.9	Neutral
4	value_for_money	8.7	8.2	0.5	49.8	Strength
5	comfort_and_room_quality	NaN	8.3	NaN	NaN	Missing rating

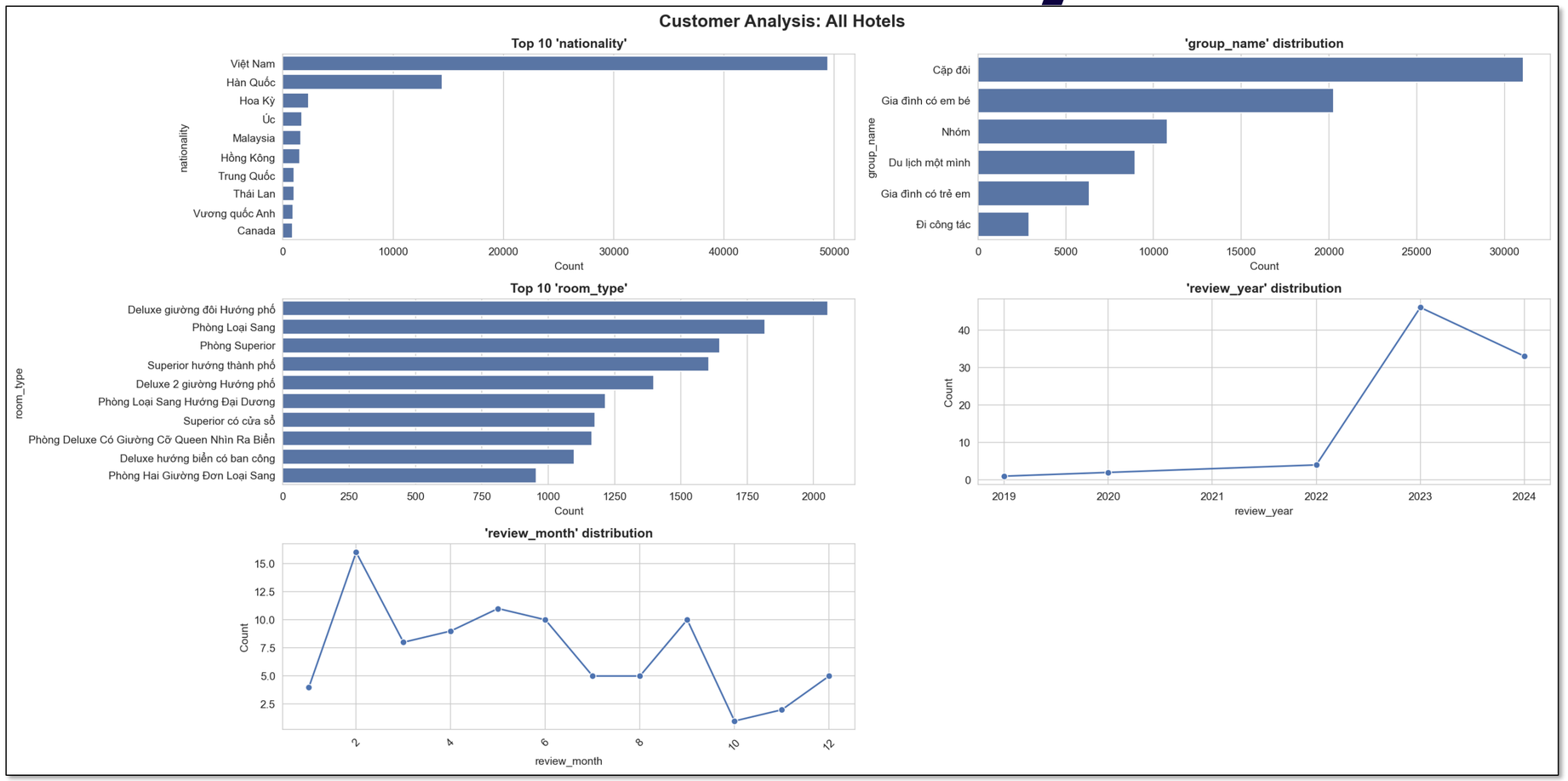
```
Strength:
        + location
        + cleanliness
        + service
        + value_for_money
Neutral:
        + facilities
Weakness:
Missing rating:
        + comfort_and_room_quality


data_output/
------------------------------
11_insight_strengths_weaknesses.csv
```



10. Hotel Insights – customer analysis



10. Hotel Insights – customer analysis



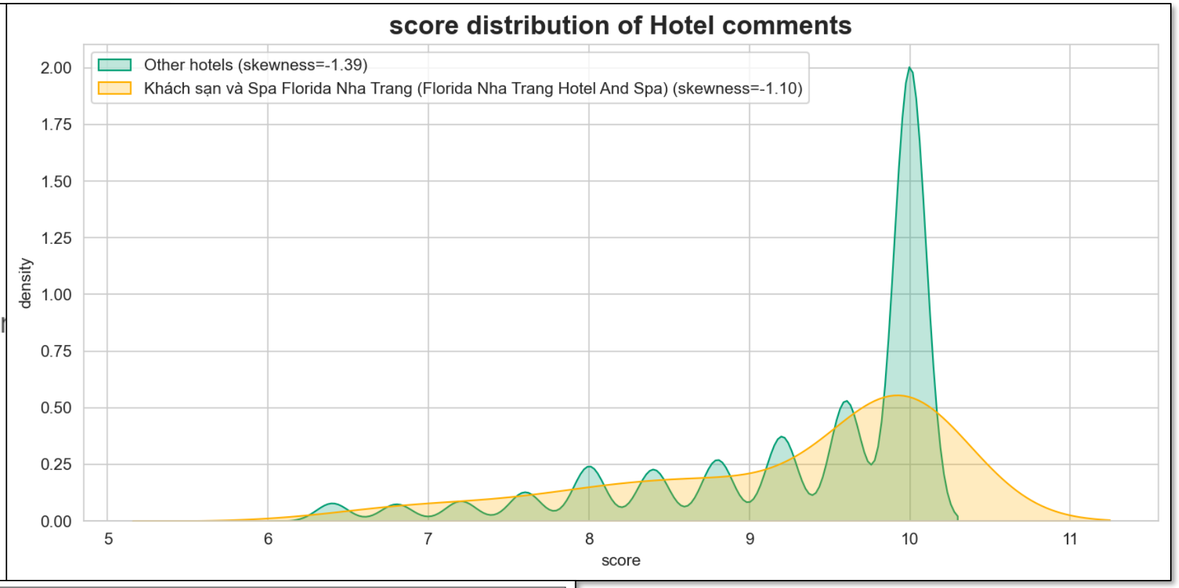
10. Hotel Insights – text-mining on reviews

Score distribution | All hotel

	score_level	count	count_pct	mean	min	q1	q2	q3	max
0	Trên cả tuyệt vời	58051	72.3	9.823586	9.0	9.6	10.0	10.0	10.0
1	Tuyệt vời	14814	18.4	8.414999	8.0	8.0	8.4	8.8	8.8
2	Rất tốt	4356	5.4	7.437006	7.0	7.2	7.6	7.6	7.7
3	Hài Lòng	3093	3.9	6.593113	6.3	6.4	6.4	6.8	6.8

Score distribution | Khách sạn và Spa Florida Nha Trang (Florida Nha Trang Hotel Ar

	score_level	count	count_pct	mean	min	q1	q2	q3	max
0	Trên cả tuyệt vời	56	65.1	9.871429	9.2	9.9	10.0	10.0	10.0
1	Tuyệt vời	20	23.3	8.440000	8.0	8.0	8.4	8.8	8.8
2	Rất tốt	6	7.0	7.400000	7.2	7.2	7.4	7.6	7.6
3	Hài Lòng	4	4.7	6.700000	6.4	6.7	6.8	6.8	6.8



Quan sát	Nhận xét
Phân bố review score cực kỳ lệch về phía cao	- Phần lớn đánh giá có điểm rất cao, cho thấy phân phối lệch về phía tích cực (thiên lệch chọn mẫu). - Peak rõ ràng tại điểm 10, cho thấy đa số khách cho điểm cao, review tích cực, bias nặng, khó khai thác insight từ review tiêu cực.
Hành vi đánh giá	Khách hàng chủ yếu để lại đánh giá khi hài lòng, dẫn đến tỷ lệ điểm cao bị đại diện quá mức.

Phân loại score_level	score_level được xác định như sau:
	`Trên cả tuyệt vời`: score `≥ 9`
	`Tuyệt vời`: score `≥ 8`
	`Rất tốt`: score `≥ 7`
	`Hài lòng`: score `≥ 6`
	Và hiện tại chưa có đánh giá nào có score < 6.0

10. Hotel Insights – text-mining on reviews

Score distribution | All hotel

	score_level	count	count_pct	min
0	Trên cả tuyệt vời	58051	72.3	9.0
1	Tuyệt vời	14814	18.4	8.0
2	Rất tốt	4356	5.4	7.0
3	Hài Lòng	3093	3.9	6.3

Gom nhóm	score threshold mới	pct	classification
Trên cả tuyệt vời + Tuyệt vời	≥ 8.0	90.7%	Positive
Rất tốt + Hài Lòng	< 8.0	9.3%	Negative

Lý do phân nhóm:
Để giảm thiểu ảnh hưởng của việc mất cân bằng dữ liệu (imbalance):
- Nhóm `Negative` được chọn sao cho chiếm khoảng 10% tổng số đánh giá, gồm các mức `Rất tốt` và `Hài Lòng`.
- Nhóm `Positive` gồm `Trên cả tuyệt vời` và `Tuyệt vời`, chiếm khoảng 90%.

10. Hotel Insights – text-mining on reviews

```python
# xử lý token
df_nev_pos_comments['body_new_clean'] = df_nev_pos_comments['body_new'].apply(
    lambda x: fn_clean_tokens(
        tokens=[t.replace(' ', '_') for t in tokenize(x).split()],
        dict_list=[emoji, teencode, engvie],
        stopword=stopword_vie,
        wrongword=wrongword,
        remove_number=True,
        remove_punctuation=True,
        remove_vie_tone=False,
        lower=True,
    )
)
```

✓ 4m 28.0s

✓ `# sau xử lý token` ···

dataframe: 80,314 rows x 7 cols

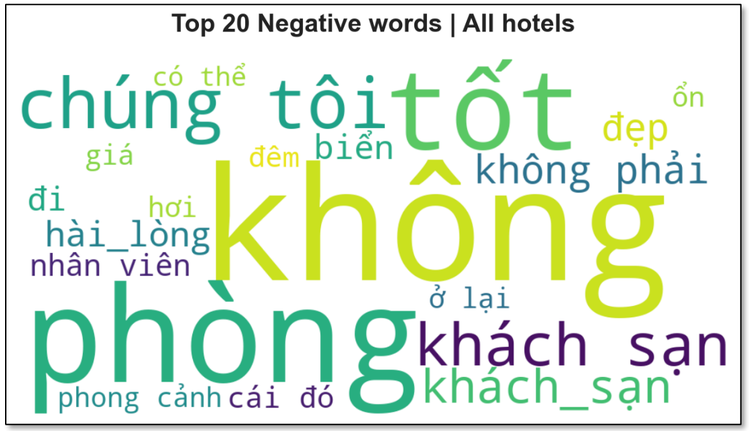
	num	hotel_id	score	score_level	classify	body_new	body_new_clean
0	1	1_1	10.0	Trên cả tuyệt vời	pos	Cao nhất!!" Tôi đã ở cùng chủ nhân trong 4 đêm...	[!!, , chủ_nhân, đêm, thân_thiện, tầm, phòng, ...
1	2	1_1	10.0	Trên cả tuyệt vời	pos	Tháng 8" Lựa chọn Mường Thanh vì giá cả phù hợ...	[lựa_chọn, mường_thanh, giá_cả, online, ưu_đãi...
2	3	1_1	9.2	Trên cả tuyệt vời	pos	Du lịch tại Nha Trang" Lần này đến với Nha Tra...	[du_lịch, nha_trang, nha_trang, sách, phòng, k...

10. Hotel Insights – text-mining on reviews

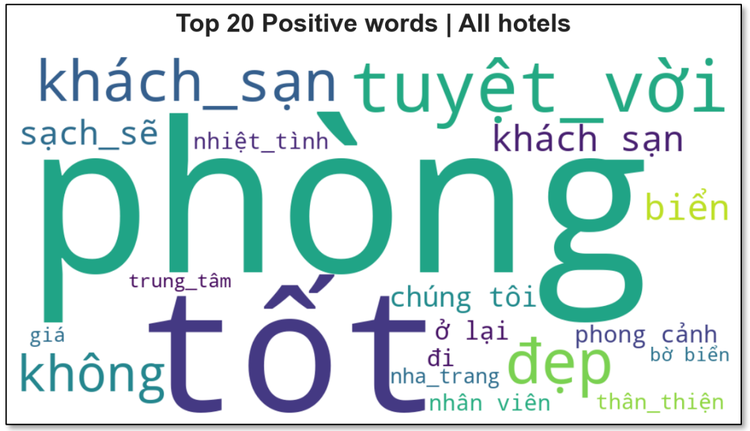
Top 20 Negative words | Khách sạn và Spa Florida Nha Trang (Florida Nha Trang Hotel And Spa)

Top 20 Positive words | Khách sạn và Spa Florida Nha Trang (Florida Nha Trang Hotel And Spa)

Top 20 Negative words | All hotels



Top 20 Positive words | All hotels



11. Comparison

Phương pháp	Ưu điểm	Nhược điểm	Hiệu quả / Ứng dụng chính
Content-based (Gensim embedding)	- Hiểu ngữ nghĩa tốt - Giảm cold-start - Hỗ trợ query từ khóa linh hoạt	- Phụ thuộc chất lượng embedding - Khó khai thác sở thích ẩn	Phù hợp search-based recommendation, metadata rich
Content-based (Cosine similarity)	- Đơn giản, dễ implement - Tính toán nhanh với tập nhỏ - Minh bạch, dễ giải thích	- Không hiểu ngữ nghĩa sâu - Khó xử lý synonym - Pairwise cosine chậm nếu dataset lớn	Baseline, ổn cho dữ liệu nhỏ/gọn, văn bản ngắn
Collaborative Filtering (ALS)	- Khai thác hành vi tập thể - Cho gợi ý bất ngờ (serendipity) - Hiệu quả khi data lớn	- Cold-start user/item mới - Cần dữ liệu hành vi dày - Khó giải thích latent factor	Mạnh khi có nhiều log hành vi (booking, rating)

OUR TEAM

Trần Đình Hùng

Phạm Ngọc Trọng