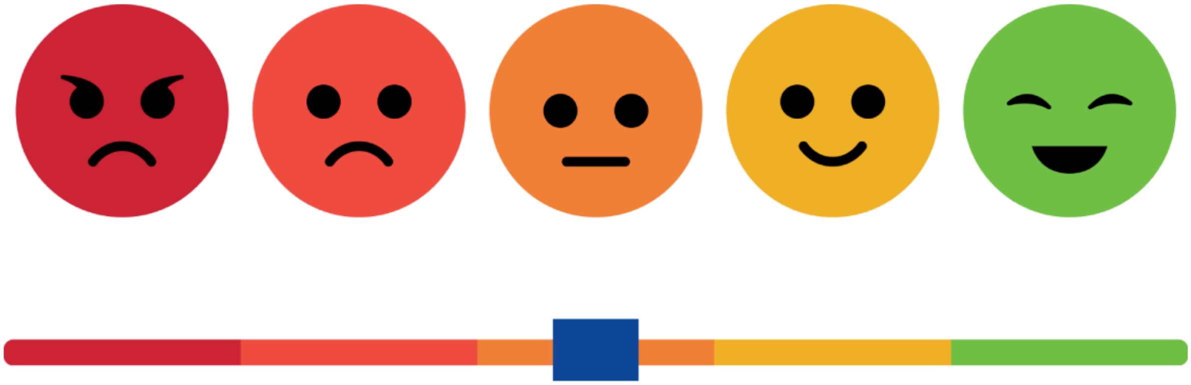


Enhancing Hotel Review Search Engines:

Overview:

In this assessment, we aimed to enhance traditional search engines by moving beyond embedding-based search methods. We explored a dual approach that combines embedding-based search with Aspect-based Sentiment Analysis (ABSA) to improve search relevance and user satisfaction.

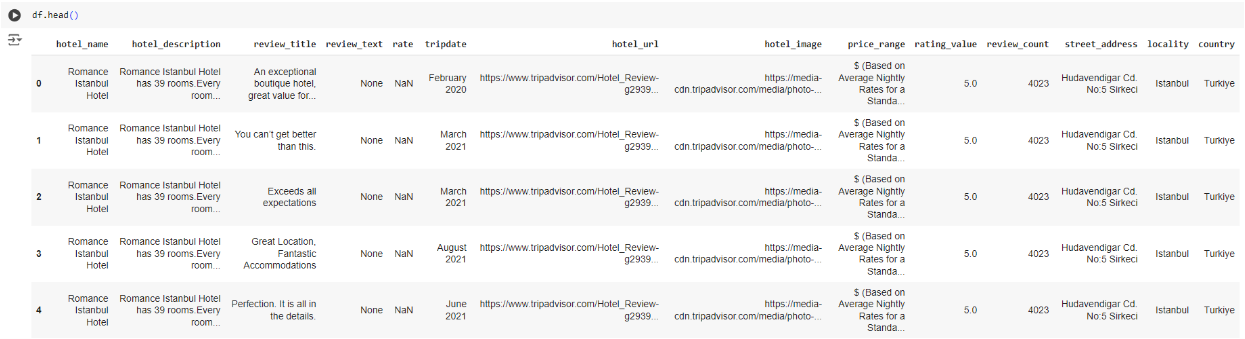
1. Collecting the Data:

The dataset used for this assessment is the **hotel_datasets** from Hugging Face. This dataset contains **5997 rows** and **14 columns** that provide comprehensive information about various hotels in five different cities worldwide.

Dataset Columns:

- **hotel_name**
- **hotel_description**

- **review_title**
- **review_text**
- **rate**
- **tripdate**
- **hotel_url**
- **hotel_image**
- **price_range**
- **rating_value**
- **review_count**
- **street_address**
- **locality**
- **country**



```
df.head()
```

	hotel_name	hotel_description	review_title	review_text	rate	tripdate	hotel_url	hotel_image	price_range	rating_value	review_count	street_address	locality	country
0	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	An exceptional boutique hotel, great value for...	None	NaN	February 2020	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye
1	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	You can't get better than this.	None	NaN	March 2021	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye
2	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	Exceeds all expectations	None	NaN	March 2021	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye
3	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	Great Location, Fantastic Accommodations	None	NaN	August 2021	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye
4	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	Perfection. It is all in the details.	None	NaN	June 2021	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye

2. Search Approach

2.1 Embedding-Based Search

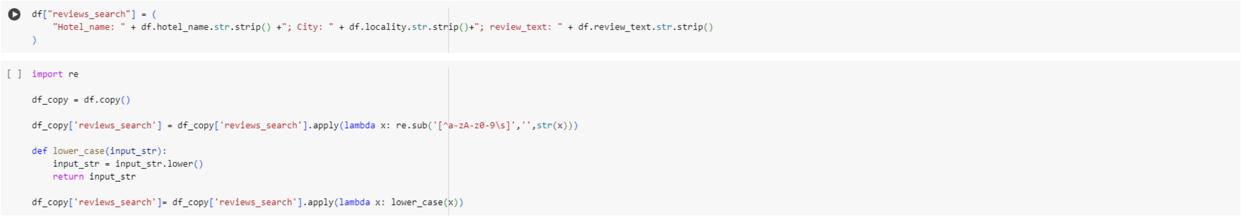
To perform an embedding-based search, we combined data from multiple columns:

- **hotel_name**
- **locality (city)**
- **review_text**

This merged data was stored in a new column called **"reviews_search."**

Data Preprocessing:

- The data in the "reviews_search" column was cleaned to contain only letters and digits (0-9).
- The text was converted to lowercase.
- The resulting text was embedded using a suitable embedding model.
- The generated embeddings were saved in a new column called **"embeddings."**



```
df["reviews_search"] = (
    "Hotel_name: " + df.hotel_name.str.strip() +"; City: " + df.locality.str.strip()+"; review_text: " + df.review_text.str.strip()
)
```

```
import re

df_copy = df.copy()

df_copy['reviews_search'] = df_copy['reviews_search'].apply(lambda x: re.sub('[^a-zA-z0-9\s]','',str(x)))

def lower_case(input_str):
    input_str = input_str.lower()
    return input_str

df_copy['reviews_search']= df_copy['reviews_search'].apply(lambda x: lower_case(x))
```

2.2 Aspect-Based Sentiment Analysis (ABSA)

To expand beyond embedding-based search, we performed Aspect-based Sentiment Analysis (ABSA) on the **review_text** column.

Aspects Considered:

1. Room
2. Service
3. Location
4. Staff
5. Food

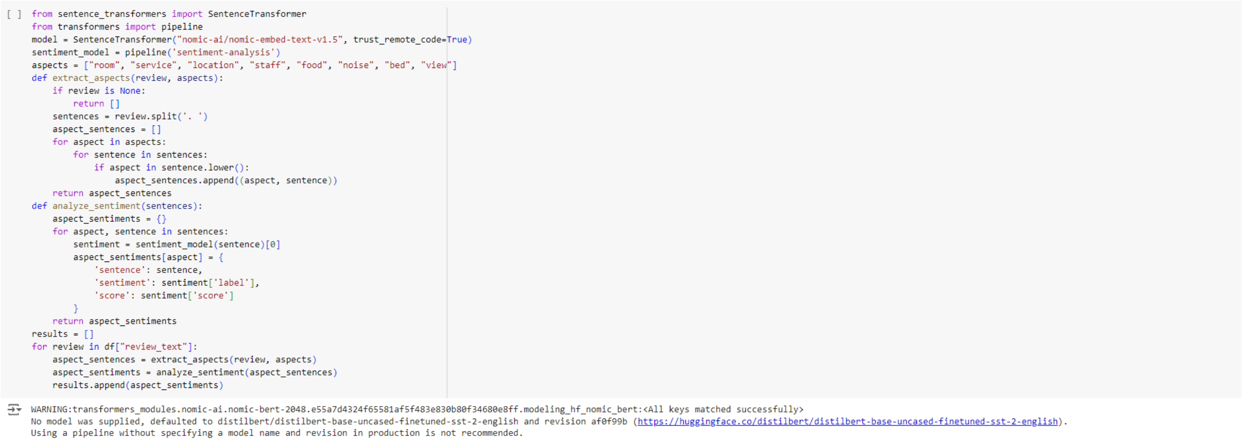
6. Noise
7. Bed
8. View

We used the **nomic-ai/nomic-embed-text-v1.5** model to analyze the sentiments of the sentences in the reviews.

Functions Defined:

1. `extract_aspects(review, aspects)`
   - Splits the review text into sentences.
   - Checks each sentence to see if it mentions any of the defined aspects.
   - Stores the aspect and the sentence as a tuple in a list called `aspect_sentences`.
2. `analyze_sentiment(sentences)`
   - Takes a list of aspect-sentence tuples.
   - Analyzes the sentiment of each sentence using the sentiment model.
   - Records the mentioned aspect, the sentence itself, the sentiment label (Positive or Negative), and the sentiment score.

Each review in the **review_text** column is processed to extract sentences related to the defined aspects, followed by sentiment analysis.



```
[ ] from sentence_transformers import SentenceTransformer
    from transformers import pipeline
    model = SentenceTransformer("nomic-ai/nomic-embed-text-v1.5", trust_remote_code=True)
    sentiment_model = pipeline('sentiment-analysis')
    aspects = ["room", "service", "location", "staff", "food", "noise", "bed", "view"]
    def extract_aspects(review, aspects):
        if review is None:
            return []
        sentences = review.split('. ')
        aspect_sentences = []
        for aspect in aspects:
            for sentence in sentences:
                if aspect in sentence.lower():
                    aspect_sentences.append((aspect, sentence))
        return aspect_sentences
    def analyze_sentiment(sentences):
        aspect_sentiments = {}
        for aspect, sentence in sentences:
            sentiment = sentiment_model(sentence)[0]
            aspect_sentiments[aspect] = {
                'sentence': sentence,
                'sentiment': sentiment['label'],
                'score': sentiment['score']
            }
        return aspect_sentiments
    results = []
    for review in df["review_text"]:
        aspect_sentences = extract_aspects(review, aspects)
        aspect_sentiments = analyze_sentiment(aspect_sentences)
        results.append(aspect_sentiments)
```

```
WARNING:transformers_modules.nomic-ai.nomic-bert-2048.e55a7d4324f65581af5f483e830b80f34680e8ff.modeling_hf_nomic_bert:<All keys matched successfully>
No model was supplied, defaulted to distilbert/distilbert-base-uncased-finetuned-sst-2-english and revision af0f99b (https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english).
Using a pipeline without specifying a model name and revision in production is not recommended.
```

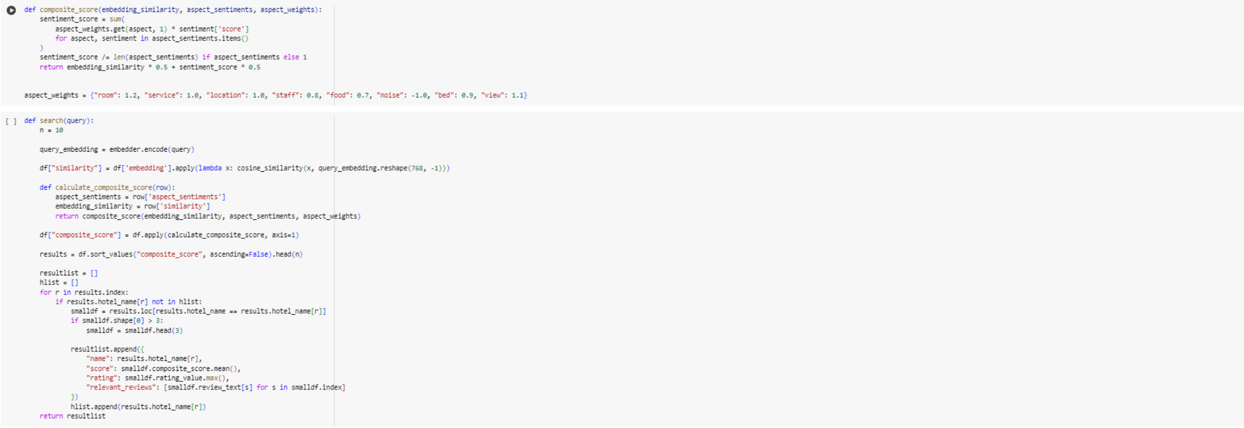
3. Composite Scoring for Enhanced Search

The search engine leverages a function named `composite_score` that integrates both embedding-based similarity and sentiment analysis results.

How It Works:

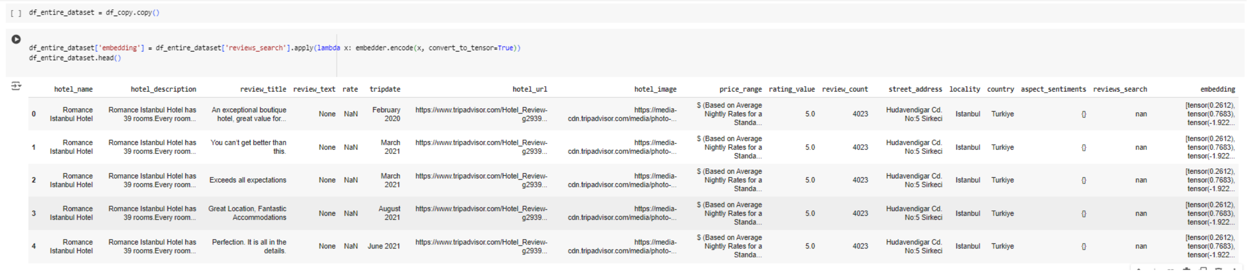
- **Cosine Similarity:** Measures the similarity between the user's search query and the embeddings of the review content.
- **Sentiment Score:** Reflects the sentiment associated with predefined aspects, calculated based on predefined aspect weights.

The `composite_score` function averages these two scores, providing a balanced evaluation of content relevance and sentiment, which is then used to rank the search results.



```
def composite_score(embedding_similarity, aspect_sentiments, aspect_weights):
    sentiment_score = sum(
        aspect_weights.get(aspect, 1) * sentiment['score']
        for aspect, sentiment in aspect_sentiments.items()
    )
    sentiment_score /= len(aspect_sentiments) if aspect_sentiments else 1
    return embedding_similarity * 0.5 + sentiment_score * 0.5


aspect_weights = {"room": 1.2, "service": 1.0, "location": 1.0, "staff": 0.8, "food": 0.7, "noise": -1.0, "bed": 0.9, "view": 1.1}
```

```
def search(query):
    n = 10

    query_embedding = embedder.encode(query)

    df["similarity"] = df['embedding'].apply(lambda x: cosine_similarity(x, query_embedding.reshape(768, -1)))

    def calculate_composite_score(row):
        aspect_sentiments = row['aspect_sentiments']
        embedding_similarity = row['similarity']
        return composite_score(embedding_similarity, aspect_sentiments, aspect_weights)

    df["composite_score"] = df.apply(calculate_composite_score, axis=1)

    results = df.sort_values("composite_score", ascending=False).head(n)

    resultlist = []
    hlist = []
    for r in results.index:
        if results.hotel_name[r] not in hlist:
            smalldf = results.loc[results.hotel_name == results.hotel_name[r]]
            if smalldf.shape[0] > 3:
                smalldf = smalldf.head(3)

            resultlist.append({
                "name": results.hotel_name[r],
                "score": smalldf.composite_score.mean(),
                "rating": smalldf.rating_value.max(),
                "relevant_reviews": [smalldf.review_text[s] for s in smalldf.index]
            })
            hlist.append(results.hotel_name[r])
    return resultlist
```



```
df_entire_dataset = df_copy.copy()
```

```
df_entire_dataset['embedding'] = df_entire_dataset['reviews_search'].apply(lambda x: embedder.encode(x, convert_to_tensor=True))
df_entire_dataset.head()
```

	hotel_name	hotel_description	review_title	review_text	rate	tripdate	hotel_url	hotel_image	price_range	rating_value	review_count	street_address	locality	country	aspect_sentiments	reviews_search	embedding
0	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	An exceptional boutique hotel, great value for...	None	NaN	February 2020	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye	{}	nan	[tensor(0.2612), tensor(0.7683), tensor(-1.922...
1	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	You can't get better than this.	None	NaN	March 2021	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye	{}	nan	[tensor(0.2612), tensor(0.7683), tensor(-1.922...
2	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	Exceeds all expectations	None	NaN	March 2021	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye	{}	nan	[tensor(0.2612), tensor(0.7683), tensor(-1.922...
3	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	Great Location, Fantastic Accommodations	None	NaN	August 2021	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye	{}	nan	[tensor(0.2612), tensor(0.7683), tensor(-1.922...
4	Romance Istanbul Hotel	Romance Istanbul Hotel has 39 rooms.Every room...	Perfection. It is all in the details.	None	NaN	June 2021	https://www.tripadvisor.com/Hotel_Review-g2939...	https://media-cdn.tripadvisor.com/media/photo-...	$ (Based on Average Nightly Rates for a Standa...	5.0	4023	Hudavendigar Cd. No:5 Sirkeci	Istanbul	Turkiye	{}	nan	[tensor(0.2612), tensor(0.7683), tensor(-1.922...

4. Conclusion

By combining embedding-based search with Aspect-based Sentiment Analysis, we have enhanced the search engine's ability to deliver more relevant and sentiment-aware results. This dual approach ensures that both the content's relevance and the user's emotional response to key aspects are considered, leading to improved user satisfaction.

Notebook(Demo):

Task2-Hotel Review Search Engines

Team member:

Sara Almutairi

Hadeel Alnasiri

Sarah Almohammedsaleh