Group 0 Project: Semantic Search engine for Hotels

by: Shayi Aldohaim, Mohammed Alhusini, Abdulaziz Althumairy

Table of Contents

Abstract

Semantic search has improved user experience by showing the most relevant results based on user needs, thus reducing time and increasing hyper personalization

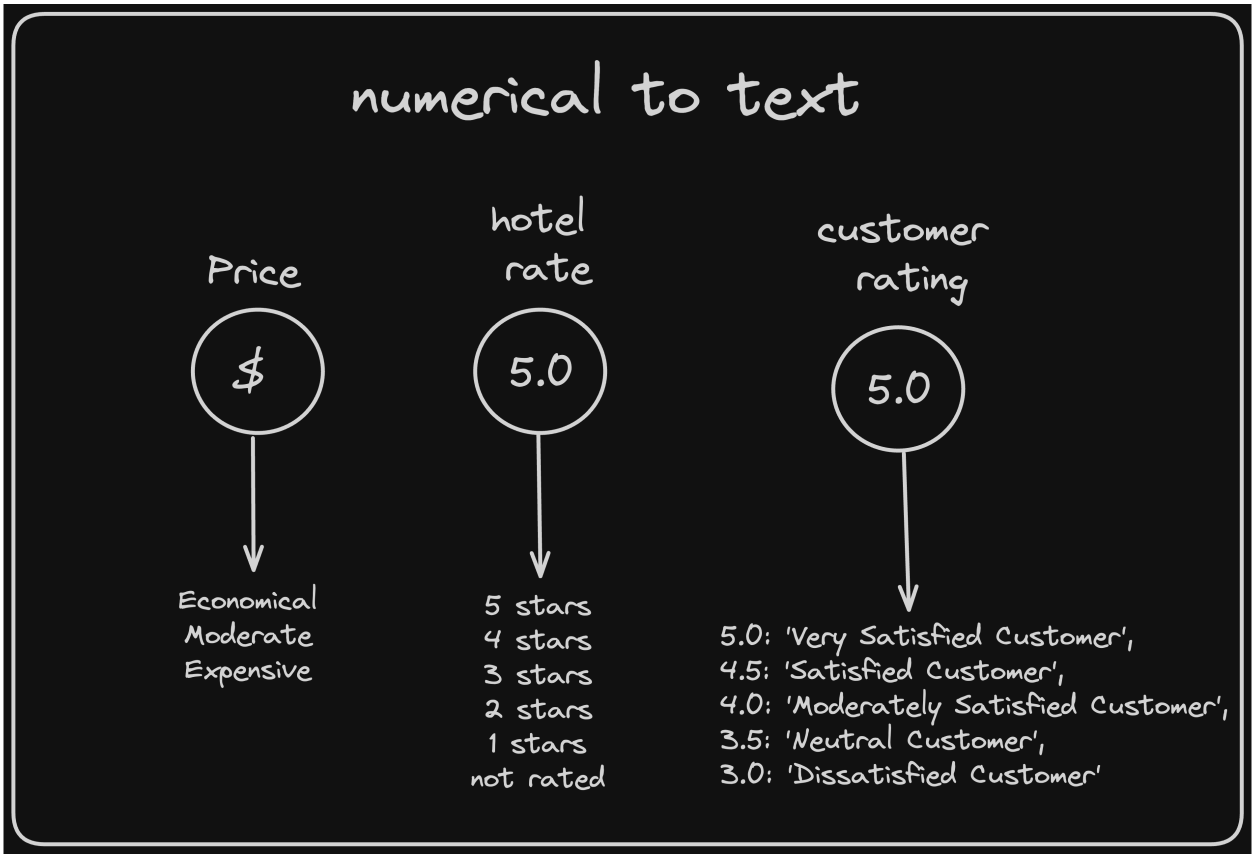
Introduction

The project task is to search in a hotels dataset based on a user query.

Data Preprocessing

- The price feature and star rating features have been converted to **text representations** of symbols and numbers (This is semantic mapping for better search):
    - For example, in price ($$$ is converted to expensive).

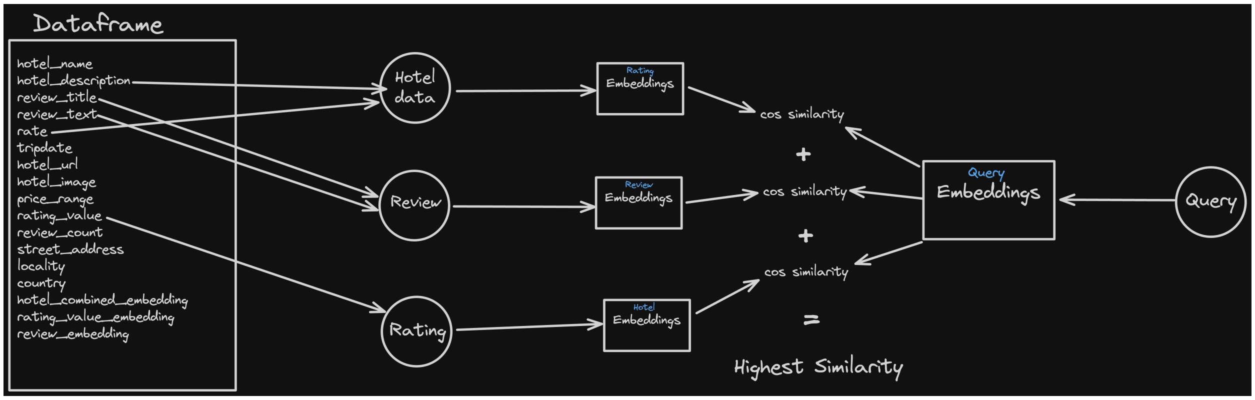
  - For example, in hotel star rate (5.0 is converted to 5 stars).
  - For example, customer rating (5.0 is converted to very satisfied customer).
  - The full mapping is shown in the following graph.



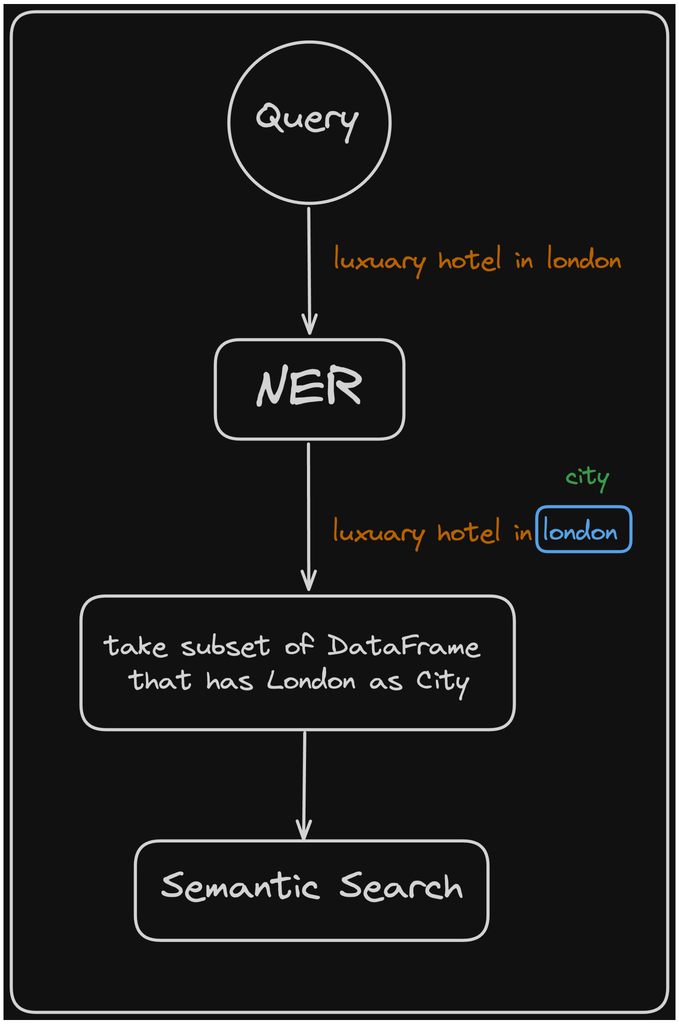
- The review text and review title features are combined.
- The hotel description, star rating, and price features are combined.
- The missing values in the hotel description and star rating features have been addressed by filling them with the correct data. For instance, if a particular hotel has a description in some rows but not in others, the missing rows are updated with the accurate description. The same process is applied to the star rating feature, ensuring that all entries are complete and consistent.

Methodology

- **Three embeddings to represents the text data:**
    - Review text and review title
    - Hotel description, price range, and star ratings
    - Customer rating

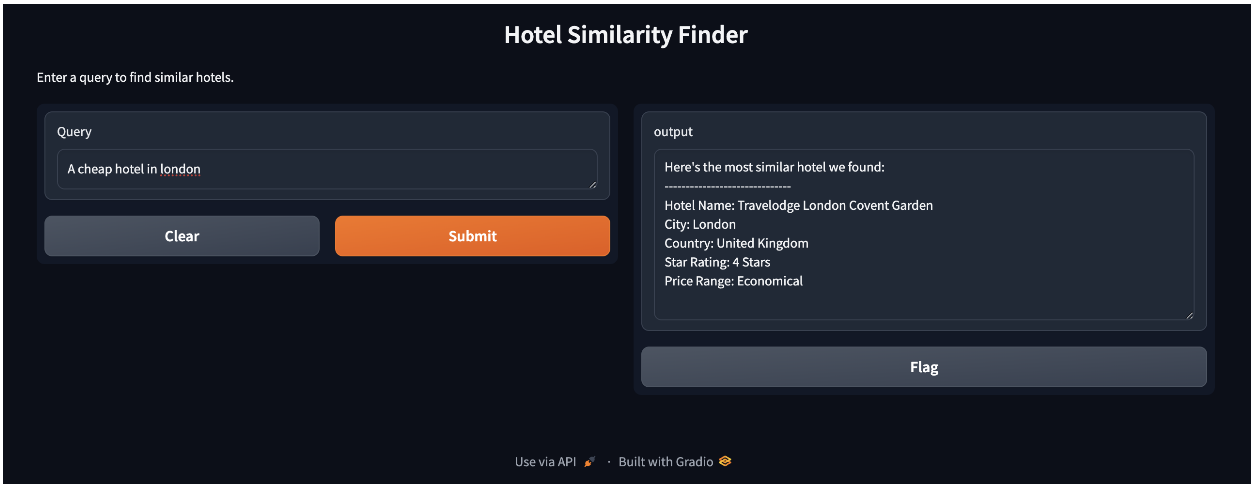


- **Steps:**
    1. The user will input the query.
    2. Use NER (name entity recognition) to recognize the desired city.
    3. The engine will make a new data frame for the desired city. (see graph below)
    4. Encode the query to get its embedding.
    5. Calculate the cosine similarities between the query and the three embeddings in the city data frame.
    6. Sum the three similarities for every row.
    7. Show the top 3 results (as cosine similarities) to the user.



Model Deployment

we used **Gradio** for deploying the model



A cheap hotel in london

```
Here's the most similar hotel we found:
------------------------------
Hotel Name: Travelodge London Covent Garden
City: London
Country: United Kingdom
Star Rating: 4 Stars
Price Range: Economical
```

Conclusion

The results we have found are very relevant as expected based on engine testing and the examples shown previously in this page.

Limitation and future work

The engine could be improved to be faster.

If the dataset has more values, some technique might be needed. Because the dataset in this project has only 6000 rows.