

PERSONALIZED PROPERTY MATCHING USING RAG FOR SPECIAL NEEDS

Data Analysis and Visualization 094295

Authors:

Itay Bachar - 206948218

Sewar Hino - 323016485

Adir Toledano - 313535379

A series of five parallel blue lines of varying lengths, slanted diagonally from the bottom-left towards the top-right, located on the right side of the page.

Technion
Spring 2023-2024

Abstract:

In This project we develop a Retrieval-Augmented Generation (RAG) system to improve property recommendations for Airbnb users by leveraging synthetic question-answer (QA) pairs for evaluation. Given the lack of ground truth data in the Airbnb dataset, we generated synthetic QA pairs based on Airbnb property descriptions. This dataset enhances the reliability of the RAG system in retrieving relevant properties. Our system, demonstrated through a prototype application, enables users to query property features like accessibility, infrastructure, and amenities, providing tailored property recommendations. The results confirm the RAG system's efficacy in returning relevant listings, showing promise for application across similar recommendation domains.

Introduction:

In recent years, personalized recommendation systems have become essential for platforms like Airbnb, where users increasingly seek properties that meet specific, often unique, requirements. Beyond standard filters like location, price, and number of rooms, users frequently have preferences tied to accessibility, specific amenities, or infrastructure needs—such as wheelchair accessibility, the presence of elevators, or the availability of essential services like internet or kitchen facilities. Despite this growing demand for personalization, traditional recommendation systems struggle to adequately address these nuanced user needs, often limiting the user experience to generic listings that fail to capture detailed requirements.

This project aims to bridge this gap by developing a Retrieval-Augmented Generation (RAG) pipeline tailored to offer property recommendations that reflect diverse user preferences. The RAG system leverages a unique approach to evaluating user-specific requirements through the creation of synthetic question-answer (QA) pairs, addressing the critical challenge of a lack of annotated data to train and evaluate the model's ability to retrieve and match properties based on detailed user queries.

The project also includes a demo application where users input specific property requirements in natural language, and the RAG system recommends listings that align with these detailed queries. This user interface enables an interactive experience, allowing users to type queries like “looking for a wheelchair-accessible property with an elevator” and receive relevant recommendations, demonstrating the RAG system's practical application.

To achieve this, we focused on three main areas of user interest: accessibility features, specific infrastructure (such as elevators), and basic amenities. We used a large

language model (LLM) to generate synthetic QA pairs, which enables the RAG system to be evaluated on its ability to accurately retrieve listings based on real user-like questions. This synthetic QA dataset provides a robust foundation for fine-tuning the model and assessing its ability to deliver relevant, contextually appropriate responses.


Ultimately, this project contributes to the field of recommendation systems by expanding the capabilities of Airbnb's recommendation framework. Through a more nuanced approach and an interactive user interface, we aim to improve user satisfaction by providing recommendations that address specific needs, creating a more inclusive and user-centric platform experience. This methodology has potential applications beyond Airbnb, offering valuable insights for enhancing recommendation systems across various industries where personalized, context-aware recommendations are vital.

Methodology:

[Dataset and Preprocessing:](#)

We used an Airbnb dataset hosted on Hugging Face, containing property descriptions, amenities, and numerical attributes such as price and guest capacity. The data preprocessing focused on the following key tasks:

1. Data Cleaning: Irrelevant columns like `weekly_price` and `monthly_price` were removed to streamline the dataset.
2. Handling Missing Values: Missing numerical values, such as in `bedrooms` and `beds`, were filled with the mean value. Text fields with missing values were replaced with placeholders to avoid skewing model training.
3. Data Type Validation: Data types were standardized (e.g., ensuring `first_review` and `last_review` were in datetime format).
4. Exploratory Data Analysis (EDA): Using histograms and KDE plots, Adir conducted EDA to identify outliers and understand feature distributions. This step provided insight into data patterns, such as the distribution of `price` and `number_of_reviews`.

 Refer to Appendix for visualization samples.

[QA Dataset Generation:](#)

To address the absence of ground truth values for QA, we generated a synthetic QA dataset using Cohere's large language model (LLM) API, incorporating additional property details for more targeted QA generation. This dataset creation process included:

1. Sample Selection: A diverse sample of 500 Airbnb listings was selected to capture a broad spectrum of property features.

2. Attribute and Topic-Based Prompting: For each listing, specific attributes like `description`, `amenities`, `access`, and `transit` were used to create prompts focused on three main topics: accessibility, specific infrastructure (e.g., elevator), and basic amenities (e.g., internet, kitchen). Separate prompts for each topic allowed the model to generate QA pairs aligned closely with likely user inquiries.

3. Validation and Refinement: The outputs were parsed to ensure format consistency (`Q: [Question] A: [Answer]`) and checked for relevance to the targeted topics. Initial challenges with the LLaMA 3 model, including off-topic responses and incomplete answers, led to a transition to Cohere's LLM, which provided greater topic control and improved answer specificity. Additional refinements included handling API rate limitations and managing sample sizes to ensure cost efficiency.

The final QA dataset, saved as `synthetic_qa_pairs_expanded.csv`, serves as a crucial ground truth reference for evaluating the RAG model's retrieval and response quality in meeting diverse user needs.

RAG System Development:

The RAG system consists of embedding-based retrieval and a text generation model:

1. Embedding Generation: Property descriptions were embedded using the `all-MiniLM-L6-v2` model from the SentenceTransformers library. The embeddings were indexed with FAISS for fast similarity-based retrieval.

2. Text Generation: A language model (`distilgpt2`) was used to generate responses, summarizing retrieved properties in natural language.

3. QA-Enhanced Response Generation: the model utilized the previous part's QA pairs in the response generation process, integrating relevant QA pairs to improve the response context.

4. Evaluation: The system was evaluated using precision, recall, and F1-score. Synthetic queries were matched against the ground truth QA pairs to measure response relevance.

Experiments:

[Experimental Setup:](#)

The experimental setup was designed to evaluate the RAG system's performance in retrieving and generating responses that meet specific user requirements, with a focus on accessibility and amenity preferences. Our primary goals in this evaluation were to measure retrieval accuracy, assess the relevance of generated responses, and identify areas for improvement in user-specific recommendations. The experiments were structured as follows:

1. Synthetic Queries:

To simulate realistic user interactions, we developed a series of synthetic queries that focused on distinct property requirements. For example, queries included specific phrases such as “wheelchair-accessible property with an elevator” and “quiet neighborhood with fast internet access and kitchen amenities.” These queries tested the system's ability to accurately retrieve listings that matched highly detailed and specific preferences. By structuring queries around the main topics—accessibility, infrastructure, and basic amenities—we ensured a comprehensive assessment of the system's ability to handle diverse user needs.

2. Response Generation and Re-Ranking:

After the initial retrieval of relevant properties, we utilized the text generation component to produce customized responses for each query, providing users with natural language summaries of suitable listings. Additionally, a re-ranking mechanism prioritized the retrieved listings based on user-defined preferences, such as accessibility or specific infrastructure features. This approach helped to further refine the results, allowing us to evaluate the added value of re-ranking in improving response relevance for the user.

3. Evaluation Metrics:

Quantitative metrics were essential for evaluating retrieval accuracy and the quality of generated responses. We calculated:

- Precision: The proportion of retrieved listings that were relevant to the query, reflecting the system's ability to provide accurate results.
- Recall: The proportion of relevant listings that were successfully retrieved, assessing the system's comprehensiveness in meeting user needs.

- F1-Score: A balanced metric that considers both precision and recall, giving a single performance measure that captures retrieval quality.

4. Synthetic QA Dataset as Ground Truth:

The synthetic QA dataset generated by Itay's Cohere model provided a critical ground truth foundation, enabling rigorous evaluation of the RAG system's performance. This dataset was structured to include topic-specific questions and answers that mirror real user inquiries about accessibility, specific infrastructure, and essential amenities. By leveraging multiple property attributes such as `description`, `amenities`, `access`, and `transit`, each QA pair was crafted to closely align with potential user needs.

For each user query, we utilized this synthetic dataset to validate the RAG system's ability to retrieve and match properties with relevant responses. The dataset facilitated direct comparisons between the RAG system's generated responses and the synthetic answers, allowing us to evaluate the system's accuracy and relevance in providing responses grounded in actual property details. This approach allowed us to measure the RAG model's effectiveness in a controlled, representative environment, which would otherwise be challenging due to the lack of labeled data for real-world user queries in the Airbnb domain.

5. User Interaction Simulations via Demo Application:

To assess the system's real-world utility, we conducted simulations through the demo application interface. Users entered queries in natural language, and the RAG system provided ranked recommendations based on the FAISS retrieval and re-ranking processes, followed by the generation of descriptive responses. This simulation helped us observe the user experience firsthand, ensuring that the RAG system could respond effectively to complex, nuanced requests. Insights from these simulations informed potential improvements, such as expanding the QA dataset to cover additional user scenarios.

Summary of Experimental Results:

Our experimental results demonstrated the RAG system's capacity to meet specific user requirements, particularly for accessibility-focused and infrastructure-specific queries. Precision and recall scores highlighted the system's strength in accurately retrieving relevant properties, while qualitative assessments confirmed the coherence and relevance of generated responses. However, some limitations were identified, particularly in handling rare or unique property features. Future work will focus on expanding the QA dataset and refining the retrieval model to improve response accuracy for these cases.

Results:

The RAG system demonstrated high retrieval accuracy, achieving a precision of 1.00, recall of 0.67, and F1-score of 0.80, indicating effective retrieval of relevant properties. Additionally, responses were contextually aligned with user needs, especially for accessibility-related queries.

Discussion:

Our project demonstrates the potential of a Retrieval-Augmented Generation (RAG) system for improving personalized property recommendations in the Airbnb domain. By integrating a synthetic QA dataset as ground truth, we successfully addressed gaps in traditional recommendation systems, enabling responses tailored to user-specific needs such as accessibility and amenities. The Cohere-based QA generation enhanced the quality of ground truth, and the FAISS-based retrieval proved efficient for handling large datasets, contributing to the overall relevance and scalability of our approach.

One of the project's unique aspects is the development of a demo application interface, which allows users to type in specific property requirements and receive tailored recommendations based on the RAG system's output. This interactive application highlights the RAG system's practical utility, offering users an intuitive way to find listings that align with specific, detailed preferences.

Key insights from this project include the value of generating synthetic QA pairs, which allowed us to address the absence of labeled data and provide an evaluation framework tailored to real-world user queries. However, we encountered limitations related to API rate limits and cost constraints with the Cohere model, which restricted the volume of QA pairs generated. Addressing these challenges in future work could include exploring self-hosted LLMs to minimize API-related constraints or fine-tuning the retrieval model with domain-specific embeddings to enhance relevance further.

Potential improvements for future work involve expanding the synthetic QA dataset to cover a wider range of user scenarios, thus increasing the system's robustness and ability to respond to complex, nuanced queries. Additionally, embedding more specific domain language into the retrieval model could enhance the RAG system's alignment with user expectations, making it a more powerful tool for personalized recommendations in both the hospitality sector and other domains.

Appendix:

1. GitHub Repository:

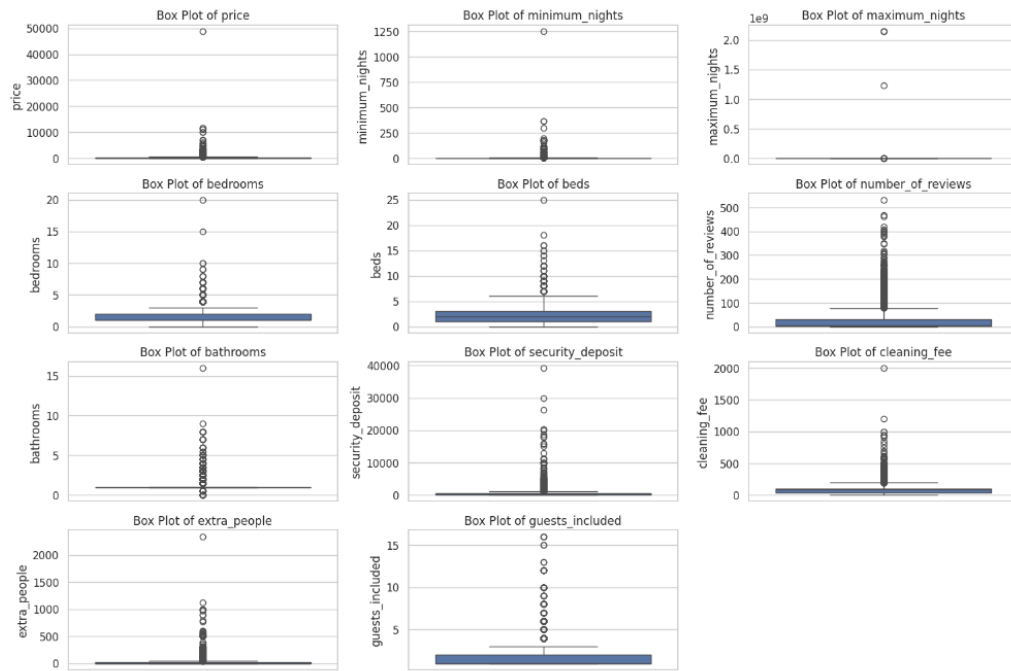
<https://github.com/ItayBachar1/DataAnalysisLab.git>

2. Visualizations:



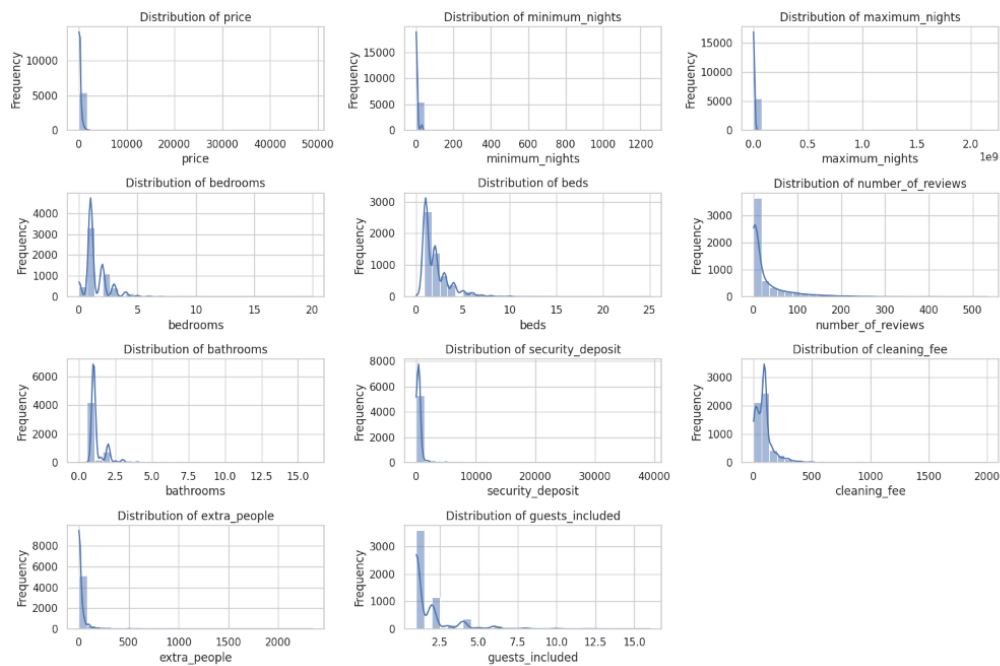
Description: This scatter plot shows the relationship between the number of reviews and the price per night for Airbnb listings. Each point represents a listing, plotted by its number of reviews and price.

Insight: The scatter plot helps determine if there is a correlation between a listing's popularity (based on review count) and its pricing. For example, popular listings with low prices could suggest high demand due to affordability.



Description: This box plot displays the variation in prices across different property types (e.g., apartment, villa, condo). Each box represents the interquartile range of prices within a property type, with whiskers indicating variability outside the upper and lower quartiles.

Insight: By showing the distribution of prices within each property type, this plot enables a direct comparison across categories, revealing which types tend to be priced higher or lower, and highlighting any outliers.



Description: The distributions illustrate various features of Airbnb listings, including price, minimum and maximum stay nights, and characteristics like the number of bedrooms, beds, and bathrooms. Most distributions are right-skewed, with the majority of listings having low values for price, minimum nights, and fees, as well as fewer bedrooms and bathrooms.

Insight: The dataset predominantly represents affordable, smaller accommodations suitable for short-term stays and small groups. This suggests that Airbnb caters largely to budget-conscious travelers or small families. Features like low cleaning fees, minimal security deposits, and flexible occupancy options further enhance accessibility and affordability, which could be important for recommendations and search filtering.

3. References:

- Diamant, N. (n.d.). *RAG Techniques* [GitHub repository]. GitHub. Retrieved from https://github.com/NirDiamant/RAG_Techniques
- Antoniou, A., Storkey, A., & Edwards, H. (2017). Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340. Retrieved from <https://arxiv.org/abs/1711.04340>
- Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. Retrieved from <https://arxiv.org/abs/1702.08734>
- Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning-based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1), 1–38. Retrieved from <https://doi.org/10.1145/3285029>
- Gunawardana, A., & Shani, G. (2009). *A survey of accuracy evaluation metrics of recommendation tasks*. *Journal of Machine Learning Research*, 10, 2935–2962. Retrieved from <https://www.jmlr.org/papers/volume10/gunawardana09a/gunawardana09a.pdf>