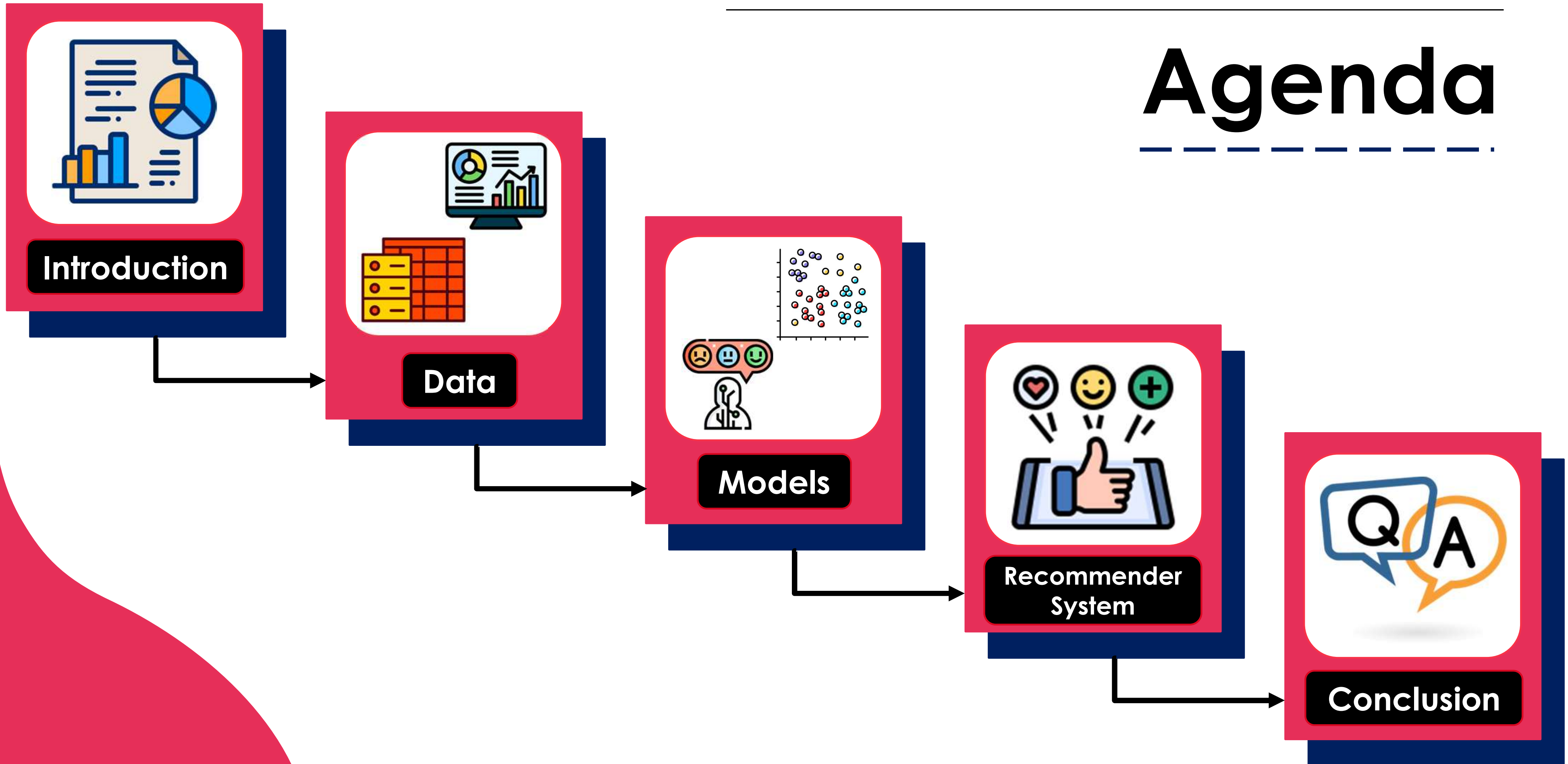
Mai Ngo

Semantic-Driven Hybrid Recommender System Using Embeddings for Chicago Airbnb Listing

11.14.2024

Agenda



Introduction

Objective

Create a **semantic-driven hybrid recommender system** specifically for Airbnb listings in Chicago using Airbnb Open Data.

By combining both user input and semantic embeddings, the system is designed to **offer personalized and contextually relevant** Airbnb recommendations.

Four main stages are employed: Exploratory Analysis (including data cleaning), Sentiment Analysis and Scoring of Reviews, Listing Clustering based on Names and Descriptions, and Recommender System Development.

To enhance user convenience, each recommended listing includes location details and the distance to the nearest CTA subway station, which is displayed on the **resulting geospatial map.**

Three dataset used are the Chicago City Airbnb listings, review data, and CTA station locations.

Data

Airbnb Listings ◻

Obs: 7,952 | Features: 18

- Numerical: 12
- Categorical: 4
- Boolean: 2

[Amenities, Location, Price, and Guest Count]

Reviews ◯

Obs: 403,291 | Features: 6

- Numerical: 3
- Categorical: 3

[Reviews and Date]

*** Multiple reviews for one listing.**

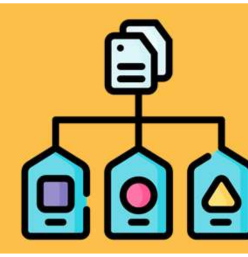
CTA Station △

Obs: 145 | Features: 3

- Numerical: 2
- Categorical: 1

[Location, Name, and Line Colors]

*** Disregard direction.**



Final Dataset

Obs: 7,952 | Features: 19

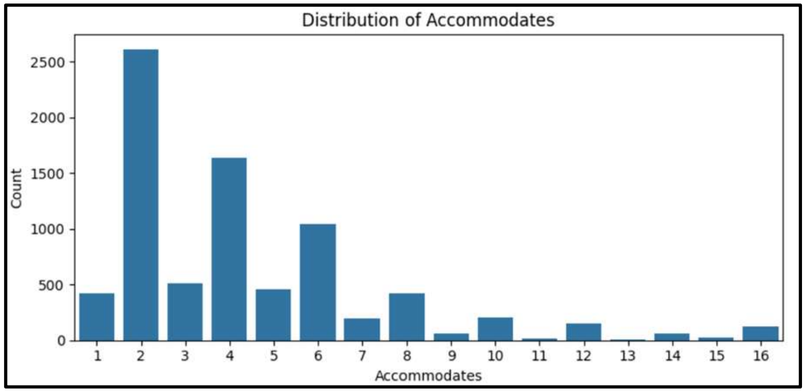
- Numerical: 10
- Categorical: 2
- Boolean: 2

[Sentiment Scores, Listing Cluster, Nearest Train Station(s)]

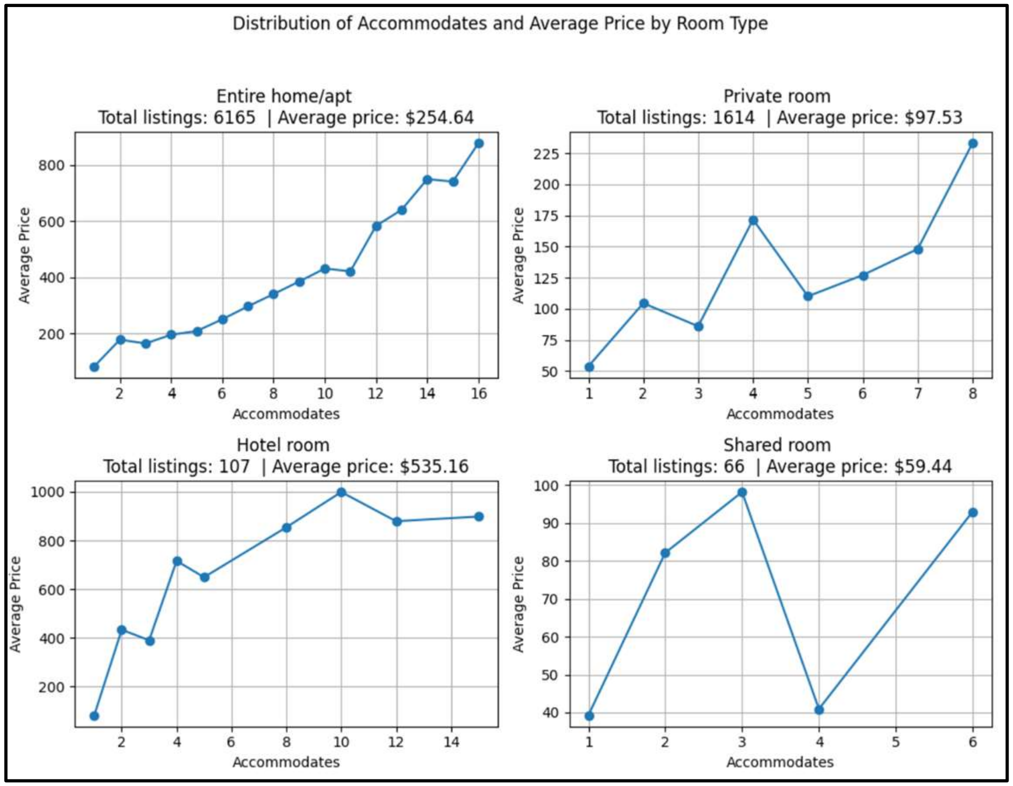
SOURCE: INSIDE AIRBNB | CTA | CHICAGO SHAPE FILE

Visualization

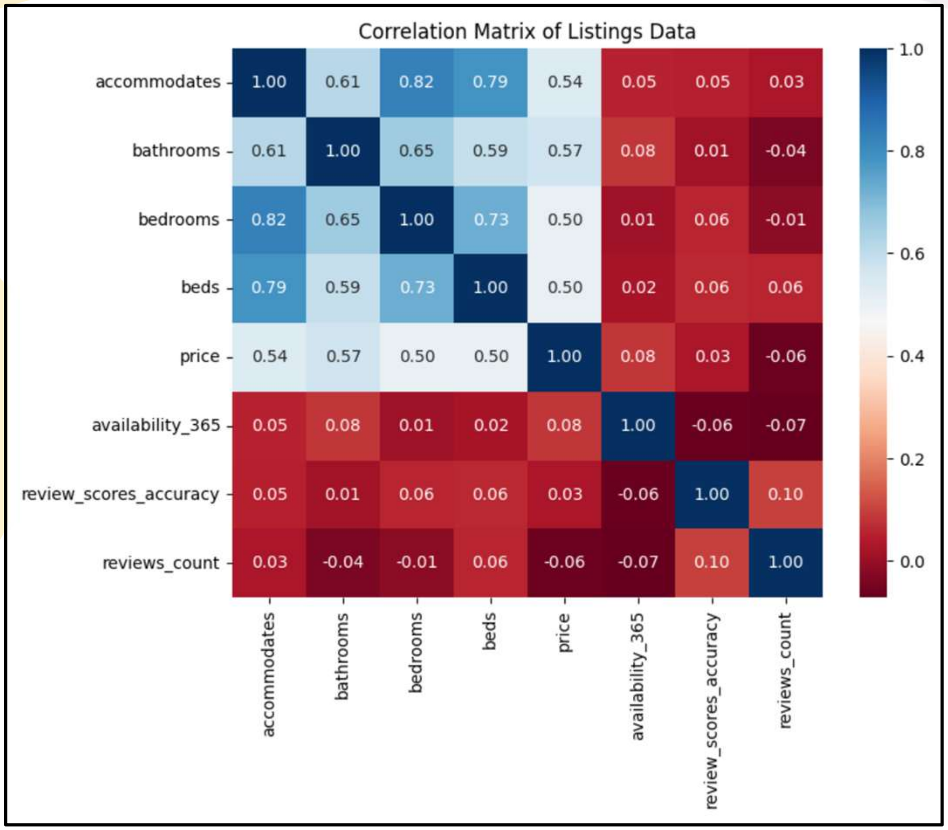
Listings Data [*Accommodates – Head count]



- 🏠 Entire Home/Apartment listings make up **77.5%** (6,165) of the data. Most listings accommodate 1 to 8 people.
- 🏠 Entire Home/Apartment, prices rise with accommodation capacity, suggesting a high demand for larger accommodations that likely appeal to groups or families.
- 🏠 Hotel room prices vary widely by capacity, likely due to luxury features.
- 🏠 Private and shared rooms show irregular price trends, influenced by location, property quality, or host ratings.



Visualization



- 🏠 Strong positive correlations between the number of guests and variables like bedrooms and beds.
- 🏠 Moderate positive relationship between price and the number of bedrooms and beds, as higher capacity generally leads to higher prices.
- 🏠 Interestingly, “review scores” show minimal correlations, indicating that customer booking decisions may not be heavily influenced by reviews alone.

Geospatial Maps

[Listings]

[CTA Stations]

Road Map

Step 1

Clean Data

(Listings, Reviews, CTA)

Step 2

Sentiment Analysis [Reviews]

(VADER, Logistic Regression, Linear SVC, and Naïve Bayes)

Step 3

Clustering Listing

using Embedding

(Hierarchical Clustering}

Step 4

Nearest Train Station(s)

Less than 10 minutes walk or < 0.8 km in distance.

(Geopy)

Step 5

Recommender System

(Embedding)

Sentiment analysis on reviews is conducted to label sentiments as positive, neutral, or negative, improving recommendation relevance by understanding guest feedback.

Sentiment Analysis

- Due to lacking ground-truth reviews data, using output from VADER as a benchmark and verify through word clouds and pre-trained models on both original and lemmatized text.
- Naive Bayes relies heavily on the original form of words and loses predictive power when generalized forms are used.
- Linear SVC classifies the word “good” as positive sentiment, while VADER classifies it as neutral → **Using VADER output for next step.**

VADER
- **Positive:** 91.8%, 6.55%
- **Neutral:** 6.55%
- **Negative:** 1.65%

Logistic Regression
- **Original text:** 85.9% Accuracy | Run time – 415.4 secs
- **Lemmatized text:** 89.6% Accuracy | Run time: 263.8 secs

Linear SVC
- **Original text:** 92.4% Accuracy | Run time – 134.3 secs
- **Lemmatized text:** 91.8% Accuracy | Run time: 212.7 secs
- **Positive:** 93.9%, **Neutral:** 4.87%, and **Negative** 1.23%

Naïve Bayes
- **Original text:** 89.1% Accuracy | Run time – 20.7 secs
- **Lemmatized text:** 73.1% Accuracy | Run time: 15.8 secs

Sentiment Analysis

Overall Sentiment Score for A Listing

- 🏠 Since each listing is associated with multiple reviews from broad time range, it is essential to account for the effect of time on sentiment scores and emphasize recent feedback.
- 🏠 For each listing, every individual review's sentiment score is multiplied by a weight determined by date of the listing's latest review. The weights are structured to prioritize more recent reviews.

- 🏠 **Reviews within the last 90 days:** Weight = 1.0
- 🏠 **Reviews from 91 to 365 days:** Weight = 0.9
- 🏠 **Reviews older than 365 days:** Weight = 0.8

The formula for calculating the weighted average sentiment score is:

$$\boxed{Weighted\ Average\ Sentiment = \frac{\sum(Sentiment\ Score \times Weight)}{\sum(Weight)}}$$

where:

Weighted Average Sentiment: The overall sentiment score for a listing based on its reviews.

Sentiment Score: The sentiment score from each individual review.

Weight: The assigned weight based on the review's recency.

Listings are clustered based on semantic similarity in names and descriptions, allowing the recommender to suggest listings based on thematic content.

Listing Clustering

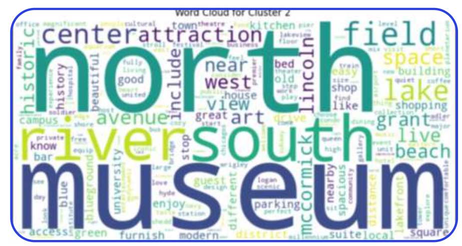
- 🏠 For each listing description, a 50-dimensional mean embedding was generated to capture the semantic information of the text.
- 🏠 Clustering based on the similarity of these embeddings, using Euclidean distance as a measure.

Cluster 1

"Vibrant Social Spaces“

(4,769 listings)

Listings that emphasize social spaces and modern amenities, appealing to those interested in lively environments.



Cluster 2

"Cultural and Scenic Attractions"

(2,483 listings)

Listings located near cultural sites and scenic landmarks, including museums and rivers. This cluster is ideal for tourists seeking proximity to attractions.

Cluster 3

"Accessible with Public Transportation“

(700 listings)

Listings highlight convenient access to public transportation options, ideal for travelers who prioritize ease of transit.

Recommender System

Step 1: Get user preferences.

- [Price, Guest Count, Neighborhoods] -> Cross Tabulate

Step 2: Get user comments.

- Ex: 'Somewhere nice and sunny.' -> 50-dimensional Embedding.

Step 3: Filter data given user requirements.

- Narrow down the data.

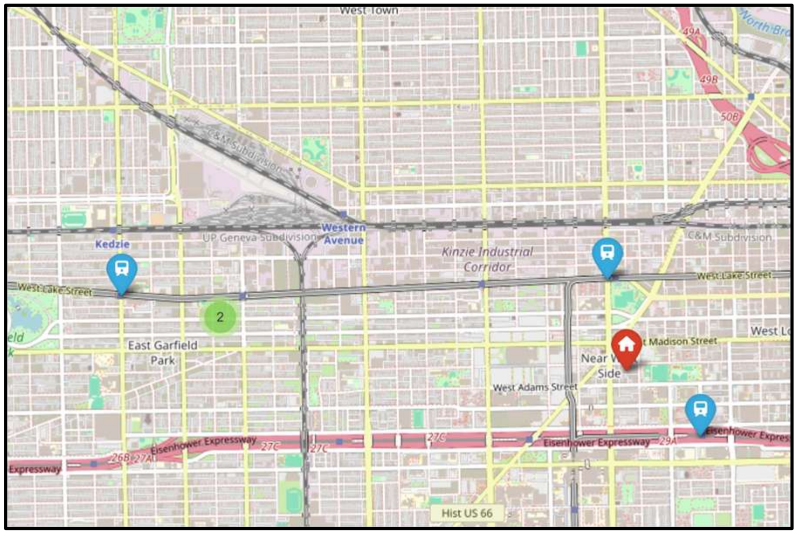
Step 4: Further rank with sentiment score.

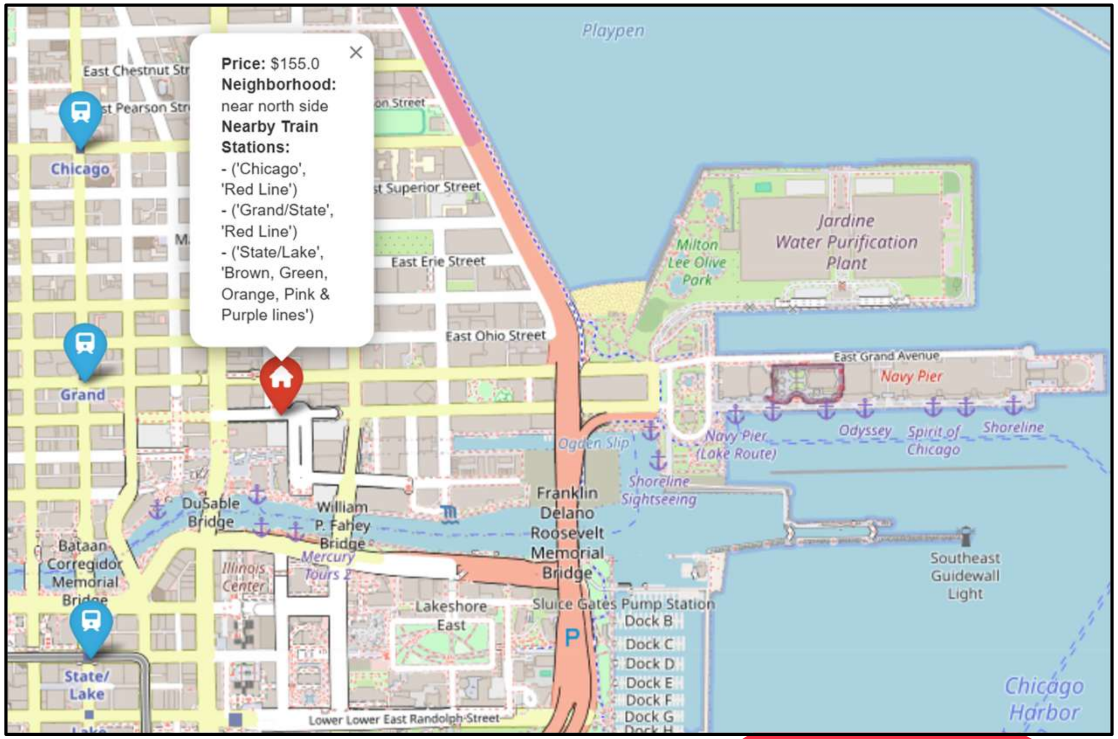
- Update ranking include sentiment score. Weights:
- Embedding similarity - 0.6 | Sentiment Score - 0.4

Step 5: Get top k Listings and Geospatial Map.

- User enter desired k. Output map of listing location and nearby train station(s).

Recommender System





Conclusion

Impact

Personalized Airbnb recommendations tailored to user preferences in a popular travel destination, Chicago.

The system adds value for travelers seeking convenient public transportation access.

Innovation

Hybrid approach combines semantic embeddings with user preferences, enhancing listing relevance.

Sentiment analysis using multiple models (VADER, Logistic Regression, Linear SVC) to validate guest feedback.

Clustering listings allowing the system to recommend properties with similar themes and characteristics.

Usability

Include neighborhood, price, amenities, and nearby subway stations, makes the recommendations actionable for users.

A **geospatial map** displaying listings and nearby CTA stations provides an intuitive, visual way to assess location and convenience.

Thank You!