February 12, 2024

Ontarian Hidden Gems Explorer

DSMM - Maple Mapping

PREPARED BY:

Auradee Castro (c0866821)

Bhumika Rajendra Babu (c0867081)

Lakshmi Kumari (c0867090)

Maricris Resma (c0872252)

TABLE OF CONTENTS

I. Project Overview

The *Ontarian Hidden Gems Explorer* project, initiated by the Ministry of Tourism of Ontario in collaboration with Maple Mapping, seeks to improve tourist experience in the region by developing a journey companion application. The objective is to guide travelers to visit lesser known but remarkable places in Ontario, fostering personalized and immersive journeys. The customization feature of the app allows itineraries to be tailored to individual interests, ensuring that travelers discover activities and destinations aligned with their preferences, ultimately enhancing the overall experience of their trip. Furthermore, this initiative aims to support smaller communities by attracting additional tourists to their locales, thereby providing support to local businesses, artisans, and cultural endeavors.

II. System Design and Feature Overview

This section provides an overview of the technical foundation of the *Ontarian Hidden Gems Explorer* application, encompassing User Profile Creation, Customized Itinerary Generation, and the Machine Learning Recommendation System. These features constitute the application's core, ensuring a personalized travel experience for people who wants to explore Ontario. The app seamlessly operates on both web and mobile platforms, leveraging AWS cloud services for scalability and optimal performance. Figure 1 illustrates the integration of these components into a dynamic, user-centric exploration platform.

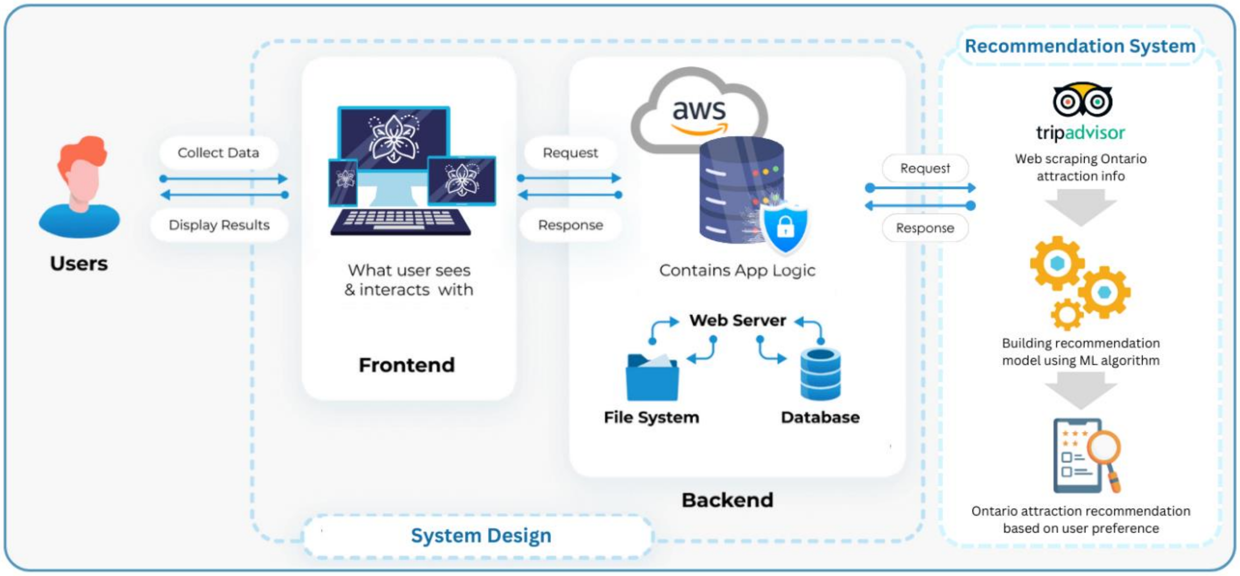


Figure 1 System Design Diagram of “Ontarian Hidden Gems Explorer” App

Application Features and Functionality

- **User Profile Creation:** Users can create detailed profiles, specifying interests, travel dates, and trip duration. Preferences for activities like nature exploration, hiking, and local cultures enhance the profile, ensuring a comprehensive user experience.
- **Customized Itinerary Generation:** The app utilizes user profile data, including interests, travel dates, and trip duration, to generate a personalized itinerary. The sophisticated algorithm suggests curated activities and destinations for a tailored experience. Real-time adjustments are possible based on user feedback or changing preferences during the trip.
- **Attraction Recommendation System:** The recommendation system employs advanced machine learning technique to analyze user preferences, suggesting unique, less-visited places in Ontario aligned with the user's interests. The continuous learning from user interactions enhances recommendations over time.

III. Recommendation System Development

The *Ontarian Hidden Gems Explorer* recommendation system is crucial for enhancing user engagement. This section details the processes of data collection, preparation and analysis necessary for building recommendation model to be integrated into the main system for deployment in public cloud, as shown in Figure 2. Using a content-based approach, the development incorporates machine learning algorithm to build a model customizing suggestions based on individual preferences for an enriched travel experience in Ontario.

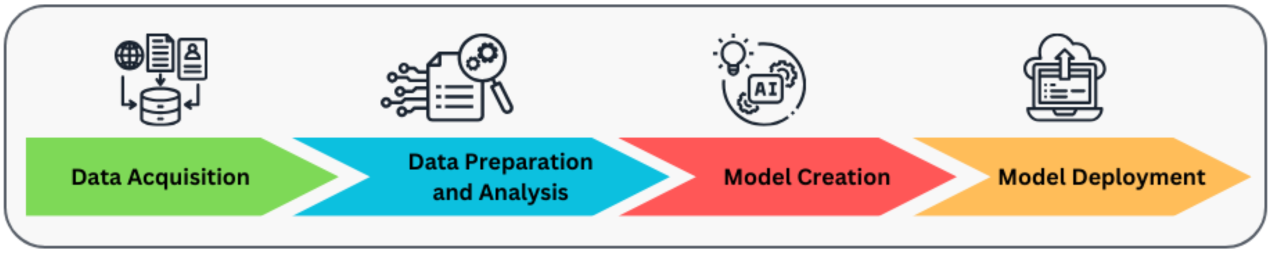


Figure 2 Data Science Process

Data Acquisition

The team leverages the information web scraped from TripAdvisor[1] platform to enhance the overall quality and authenticity of data for a more informed user experience. The dataset[2] comprises important details about various hidden attractions in the region, ensuring a personalized journey for users. Here's the key information gathered from TripAdvisor for different places in Ontario:

- **Name of the Place:** Name of each attraction, providing a quick identification but allowing for multiple entries with the same name across different locations in Ontario.
- **Description of the Place:** Summary outlining the unique features, historical significance, or distinctive aspects of each hidden attraction, enabling users to make informed choices based on their interests.
- **Location:** Specific city in Ontario where the attraction is located, offering geographical context for travelers to efficiently plan their itineraries.
- **Category:** Classification of each attraction into broader categories such as heritage places, parks, zoos, etc., assisting users in selecting attractions that align with their preferences and interests.
- **Opening Time:** Operational hours of each attraction, aiding travelers in planning their visit and ensuring they align with the venue's schedule.
- **User Rating:** A numerical representation of the overall satisfaction of visitors, providing a quick assessment of the attraction's popularity and appeal.
- **Total Review Count:** Total number of reviews submitted by visitors.

Data Preparation and Analysis

Effective data preparation and analysis are vital as they lay the foundation for accurate insights, informed decision-making, and the success of machine learning models. Each step in data preparation involves specific data preprocessing techniques to handle missing and duplicate values, transform text data, and encode categorical variables for numerical analysis. Below provides an overview of the processes performed on the data to ensure transparency and reproducibility of the analysis.

- **Feature Removal:** Dropped 'Opening Time' from the features due to 40% missing/blank data. This step is important to remove the bias and guarantee robust model performance.

[1] 'Best Ontario Hidden Gem Attractions' taken from TripAdvisor: https://www.tripadvisor.ca/Attractions-g154979-Activities-zft12156-Ontario.html

[2] Appendix A: TripAdvisor Dataset

- **Handling Missing and Duplicate Records:** Identified and handled the missing and duplicate records in the TripAdvisor dataset to maintain data integrity and ensuring the accuracy of analysis.
- **Text Preprocessing:** Processed the 'Description' data by removing numbers, symbols, and stopwords, and reducing words to their base or root form, which aims to improve text quality, simplify complexity, and enable machine learning models to extract meaningful insights from the textual information.
- **Text Vectorization:** Tokenized the text in 'Description' col into words and converted these tokens into numerical vectors using TF-IDF (Term Frequency-Inverse Document Frequency). Text vectorization is essential for machine learning models to effectively process and understand textual information.
- **Converting Categorical into Numerical Values:** Converted the 'Category' and 'Location' categories into numbers (e.g. Toronto = 0, Mississauga = 1, Ottawa = 3) using a method called one-hot encoding so that machine learning algorithm can analyze and understand them better.

In addition, various visualizations were performed to understand the data, uncover trends and identify patterns, underscoring the significance of visual exploration for comprehensive understanding of the TripAdvisor dataset. The following bar charts offer insights into the distribution of attractions across cities and the prevalent categories of these attractions within the Ontario region.

Figure 3 Top 10 Cities in Ontario with Highest Number of Attractions

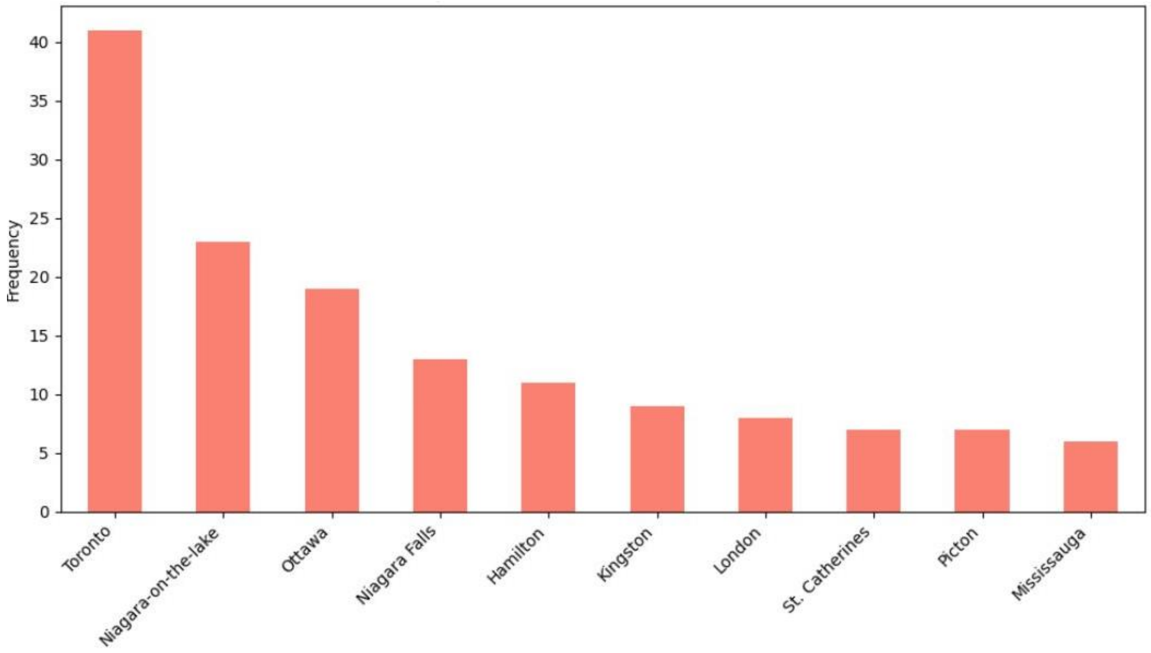
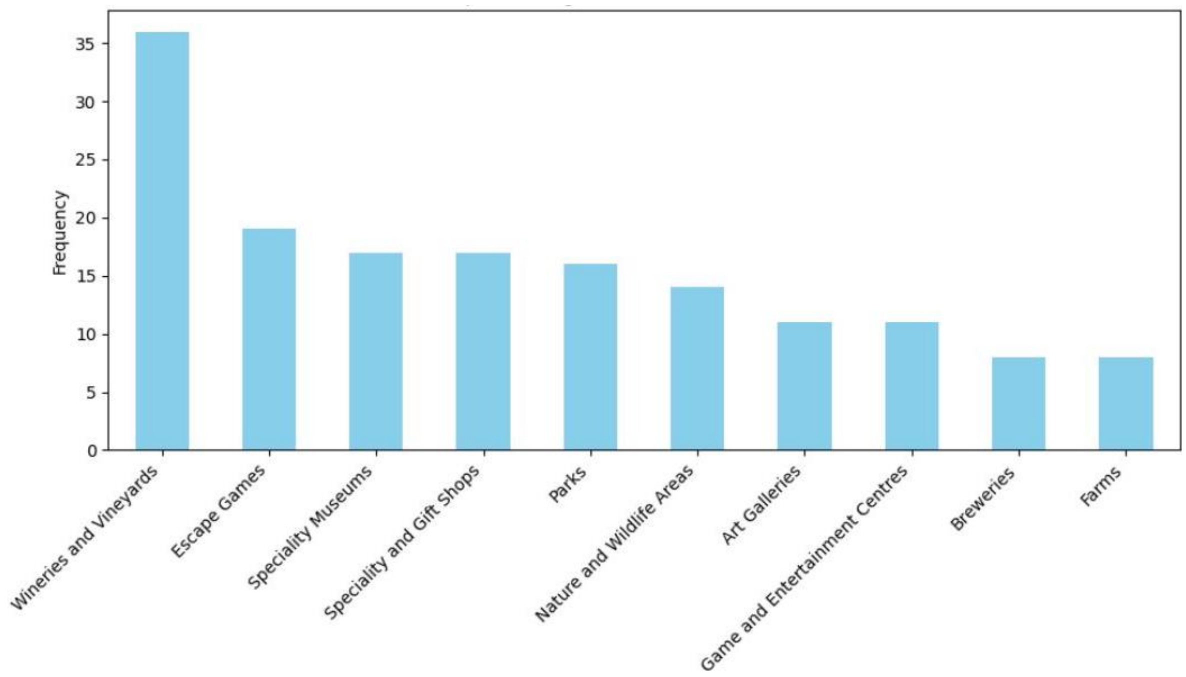


Figure 4 Top 10 Categories with Most Attractions in Ontario



Model Creation

This project involves building a recommendation system model for a variety of undiscovered attractions in Ontario, utilizing machine learning algorithms and making use of the TripAdvisor dataset that has been cleaned and processed. The focus lies in creating a personalized experience for users through a content-based recommendation approach. Content-based recommendation is a method used to suggest items or content to users based on the characteristics or features of the items they have shown interest in before. In the context of *Ontarian Hidden Gems Explorer* system, it involves recommending attractions in Ontario based on specific details like the description, category, and location of each place[3].

Utilizing cosine similarity, the recommendation system calculates attractions' similarity. Each attraction is represented as a vector based on its features (description, category, and location). When a user shows interest in a certain attraction in Ontario, the system computes cosine similarity between that attraction's vector and others. Higher similarity suggests comparable features, leading to recommendations with similar characteristics to the user's preferences.

[3] Appendix A: Codes for Ontarian Hidden Gems Explorer's Recommendation System

The *Ontarian Hidden Gems Explorer* is designed to have three recommendation models to cater different user preferences, as shown in Figure 5:

- **Attraction Preference:** Recommends places based on attractions the user has previously enjoyed.
- **Location Preference:** Recommends top places to visit based on the user's preferred location, ensuring recommendations are convenient.
- **Category Preference:** Recommends top places to visit based on the user's preferred category of attractions.

Depending on the user's input, the system can combine outputs from these models to offer a holistic set of recommendations. This comprehensive approach ensures users receive suggestions aligned with their diverse preferences.

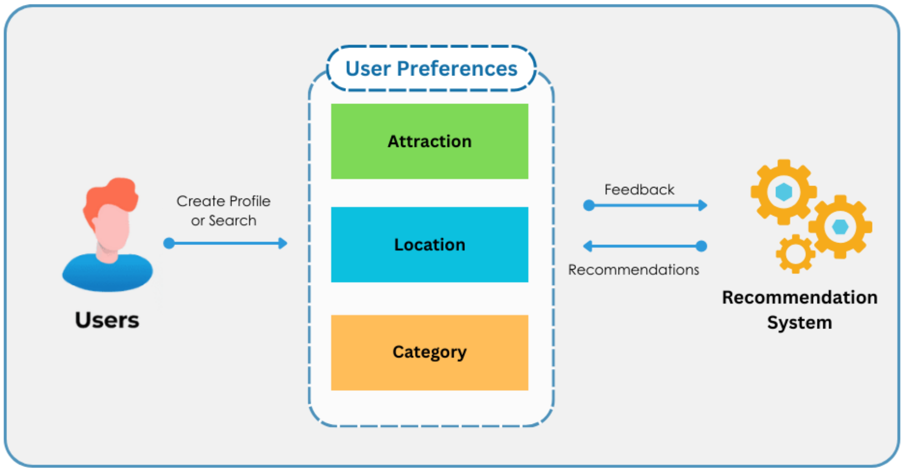


Figure 5 Recommendation System Design

Model Deployment

The recommendation models the team has created will be integrated with the *Ontarian Hidden Gems Explorer* system, enhancing its capabilities to deliver personalized attraction suggestions. The deployment in AWS public cloud is chosen for scalability and reliability, ensuring optimal performance and accommodating potential increases in user activity. This integration aims to refine the user experience by offering tailored recommendations based on diverse preferences, establishing a resilient infrastructure capable of handling varying loads. The deployment on AWS aligns with industry best practices, assuring both efficiency and dependability in delivering personalized and memorable travel experiences for Ontarian Hidden Gems Explorer users.

IV. Recommendations for System Enhancement

This section outlines proposed improvements and refinements aimed at optimizing the functionality and performance of the existing system. These enhancements are designed to elevate user experience, address identified limitations, and ensure the sustained evolution of the system in alignment with evolving requirements.

- **Improve recommendation system with hybrid approach:** Enhance the recommendation system by combining user-collaborative and content-based methods, providing a more comprehensive and personalized approach to suggestions.
- **Integrate live location for real-time recommendations:** Utilize GPS or location services to dynamically suggest nearby attractions based on the user's current whereabouts.
- **Enhance user preference data collection:** Implement interactive forms, surveys, or feedback mechanisms within the app to gather and analyze user preferences more comprehensively.
- **Integrate images for attractions:** Enrich the "Hidden Gems of Ontario" platform by incorporating visual content for all listed attractions, providing users with a more immersive exploration experience.

V. Glossary

Term	Definition / Explanation
Cosine Similarity	Cosine similarity is a mathematical metric that measures the cosine of the angle between two vectors. In the context of content-based recommendation, these vectors represent the features or characteristics of items. The closer the cosine similarity is to 1, the more similar the items are.
One-Hot Encoding	One-hot encoding is a technique used to convert categorical data into a numerical format suitable for machine learning algorithms. In this method, each category or label is represented by a binary vector, where only one element is '1' (hot) and the rest are '0' (cold). The position of the '1' corresponds to the index of the category, creating a unique binary pattern for each category. One-hot encoding is particularly useful for algorithms that require numerical input, allowing them to effectively process categorical information.
Stopwords	Stopwords are commonly used words in a language that are typically excluded from text processing and analysis because they are considered to be of little value in determining the meaning of a document. These words are frequently used but do not carry significant semantic meaning, and their presence in a text can add noise and complexity to natural language processing tasks. Examples: "the," "a," "an", "and," "but," "or", "in," "on," "at" Removing stopwords is a common preprocessing step in text analysis to focus on the more meaningful words and improve the efficiency of algorithms.
Text Vectorization	Text vectorization is the process of converting textual data into numerical vectors, allowing machine learning algorithms to process and analyze text

	effectively. In this transformation, each word or term in the text is represented by a numerical value, usually within a high-dimensional space. Text vectorization is essential in natural language processing (NLP) tasks, as it enables the application of mathematical and statistical models to textual data. This numeric representation allows machine learning algorithms to analyze relationships between words, identify patterns, and make predictions based on the content of the text.
TF-IDF (Term Frequency-Inverse Document Frequency)	TF-IDF is a numerical statistic used in natural language processing and information retrieval to evaluate the importance of a word in a document relative to a collection of documents (corpus). The TF-IDF score increases with the frequency of a term in a specific document and is offset by the number of documents in the corpus that contain the term, helping to identify the significance of words in a document within a broader context.

VI. Appendices

Appendix	Item	Location
A	Codes for Ontarian Hidden Gems Explorer's Recommendation System	Codes available on link
B	TripAdvisor Dataset	Dataset available on link