**Project Report - Travel Advisor Recommendation System**

**Project Scope**

The scope of our project encompasses the development of a sophisticated travel advising recommendation system using state-of-the-art machine learning algorithms. Our system aims to revolutionize the vacation planning experience by intelligently gathering and analyzing data from users regarding their travel preferences. This includes detailed information such as the desired destination, preferred travel dates, allocated budget, specific hotel amenities, categories of attractions they wish to visit, and their favorite cuisine types.

The primary objective of our system is to create a personalized and tailor-made travel plan for each user, taking into account their unique preferences and requirements. By leveraging advanced machine learning techniques, we can generate accurate and insightful recommendations for accommodations, activities to do at various times of the day, and dining options for breakfast, lunch, and dinner throughout the trip.

Our project's overarching goal is to provide travelers with a comprehensive and efficient tool that simplifies the entire vacation planning process. By automating the tedious tasks of researching multiple websites and sifting through vast amounts of information, we aim to save users valuable time and effort. This, in turn, allows travelers to focus more on enjoying their vacation experiences to the fullest.

Through the detailed implementation of data science tools and methodologies, our project will offer a seamless and user-friendly interface for travelers to input their preferences and receive customized travel recommendations. These recommendations will be based on data-driven insights derived from extensive data analysis and machine learning models trained on diverse travel datasets.

In summary, our project aims to be a game-changer in the travel industry by providing a holistic and efficient solution for vacation planning, ultimately enhancing the overall travel experience for users worldwide.

**Problem Definition and Challenges**

The problem definition and challenges for implementing a travel advising recommendation system using machine learning were multifaceted and required strategic solutions. The project's goals centered on utilizing three distinct recommendation models: two collaborative filtering techniques and one hybrid technique. The motivation behind this approach was to delve into various recommender system models and devise an effective solution. The collaborative filtering techniques employed were Restricted Boltzmann Machine (RBM) and Matrix Factorization using Alternating Least Squares. The hybrid technique integrated K-Means, a content-based filtering technique, and K-Nearest Neighbours, a memory-based collaborative filtering technique.

The major modules of the project encompassed data collection through web scraping, cleaning, and integration; hotel recommendation based on user's amenity requirements; attraction recommendation for morning and evening using categories of tours; restaurant recommendation for different meals of the day based on cuisine type and food preferences; and integration of all recommendations to provide a comprehensive travel plan for each day of travel.

Several challenges were encountered during implementation:

* Dataset Availability: The required data for generating recommendations was not readily available. Web scraping techniques were essential to gather information from various sources, including attractions, hotels, and reviews from relevant websites.
* Categories of Recommendation: Implementing different models for hotels, attractions, and restaurants necessitated distinct methods for data collection, pre-processing, cleaning, and user profiling, increasing the complexity and workload.
* Model Performance Evaluation and Improvement: Ensuring the models provided relevant and accurate recommendations required continuous evaluation and refinement.
* Integration of Recommendations: Merging recommendations from multiple modules to create an efficient travel plan posed integration challenges, necessitating seamless coordination and compatibility.
* Scalability: Designing the system to handle a growing volume of recommendations while maintaining performance and user experience was a crucial consideration.

Addressing these challenges involved a combination of advanced data processing techniques, model optimization, rigorous testing, and scalable architecture design to create a robust and user-friendly travel advising recommendation system.

**Project Requirements**

**Data Collection**

Data collection played a pivotal role in our travel advising recommendation system using machine learning, given the project's requirements for datasets containing detailed information about attractions, hotels, restaurants, and their corresponding reviews. Since no readily available dataset existed for hotels and attractions, data collection became a crucial and labor-intensive task. To acquire these datasets, we conducted extensive crawling through thousands of attractions and hotels listed on Tripadvisor. This process involved utilizing Python libraries such as lxml and requests for downloading and processing HTML files using the ElementTree API. Notably, due to the unique HTML page structures of hotels and attractions, we implemented distinct crawling methods for each category. Our efforts yielded approximately 3.5k attractions and 35k hotels through scraping Tripadvisor, alongside gathering around 33K reviews for attractions and a substantial 2.4M reviews for hotels.

Additionally, to enhance the restaurant recommendation aspect of our system, we utilized data from the original YELP dataset. This comprehensive dataset encompassed diverse information related to users, businesses (including but not limited to restaurants), reviews, check-in times, user tip reviews, and business photos. While the dataset contained details of various types of businesses on Yelp without explicit category features, our challenge lay in extracting only the restaurants. The techniques employed to accomplish this task will be elaborated on in the report. Ultimately, we successfully obtained data for approximately 12K restaurants and close to 5M reviews from the YELP dataset specifically related to restaurants, significantly enriching our recommendation system's data repository.

**Data Cleaning and Integration**

After gathering the extensive datasets for attractions, hotels, and restaurants, the next major undertaking for our travel advising recommendation system was data cleaning and integration. One of the significant challenges encountered during this phase was the inconsistencies present in the Tripadvisor website, leading to missing values for essential attributes such as prices, ratings, and location details in both the scraped datasets.

* **Hotel dataset**: we employed GeoPy, a Python library, to geolocate hotels based on their addresses and obtain latitude and longitude positions. In cases where GeoPy couldn't output coordinates from the address, we used locality and province information to estimate the coordinates. Missing prices and ratings were filled hierarchically based on city, province, and dataset-wide averages, assuming similar pricing and ratings patterns within localities. The amenities, crucial for interpreting hotel preferences, underwent extensive cleaning, including splitting, exploding, and converting them into a list format. Continuous unique IDs were assigned to users and hotels for streamlined data handling.
* **Attractions dataset:** only priced tours and sightseeing were included. Due to the absence of location information on the website, we created addresses from scraped data and utilized the Google API to obtain coordinates. Missing coordinate values were imputed based on assumptions about attraction categories' spatial distributions within a city. Similarly, missing prices and ratings were averaged based on city and attraction category.
* **Restaurants dataset**: initial data cleaning involved filtering businesses relevant to restaurants only from the vast Yelp dataset. We also refined the reviews dataset to align with restaurant businesses of interest, removing closed restaurants from consideration. The categories and attributes columns were crucial and underwent thorough cleaning to convert them into usable formats.

Additionally, sentiment analysis was conducted on reviews to gauge positivity levels and enhance personalized recommendations for users. Throughout these processes, meticulous attention was paid to maintaining data integrity, converting textual features into structured formats, and ensuring compatibility across datasets to facilitate seamless integration and analysis within our travel advising recommendation system.

**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) played a crucial role in our travel advising recommendation system using machine learning by providing valuable insights into the integrated datasets. This thorough analysis aimed to understand the distribution of key features such as prices, ratings, and other attributes across various categories within the dataset. EDA was instrumental in guiding our decisions regarding the selection of specific attributes for user profiling, as it highlighted the variance in data across different attributes. Some of the pivotal analyses conducted during EDA included examining the average ratings for different attraction categories, analyzing the distribution of ratings across attraction categories, exploring the average price and rating distributions across different provinces in Canada, comparing price distributions with average prices for attraction categories, and assessing the number of attractions at the province and city levels. Furthermore, EDA delved into the distribution of poorly and highly rated restaurants across provinces, rating distributions for restaurants province-wise, and the number of restaurants at the province and city levels. One significant aspect of our EDA was the sentiment analysis performed on user reviews and tips within the restaurants dataset.

This analysis aimed to understand the sentiment polarity of reviews and tips, particularly focusing on identifying positive and negative sentiments. Since some reviews and tips were in French, we initially utilized the Google Translator API to translate them into English before feeding them into the VADER sentiment analyzer, which is specifically designed for the English language. We computed positive-to-negative ratios and normalized scores across the dataset to gain deeper insights into the true positivity of user reviews. Overcoming the challenge of multilingual data through translation and employing specialized sentiment analysis tools enabled us to derive meaningful sentiment insights from the dataset during our EDA process.