CSE6240: NLP Travel Advisor

Somdut Roy*
Georgia Institute of Technology
Atlanta, Georgia, USA
somdut.roy@gatech.edu

Vitaly V. Marin
Georgia Institute of Technology
Atlanta, Georgia, USA
vmarin3@gatech.edu

Devanshee Shah
Georgia Institute of Technology
Atlanta, Georgia, USA
dshah330@gatech.edu

Figure 1: "Often, users have to make multiple trip-related choices and combining them is tedious and time-intensive"

ABSTRACT

We propose novel travel recommendation system that is able to combine three recommendations: 1) food 2) travel and 3) hotel.

KEYWORDS

natural language processing, recommender systems, trip advisor

1 INTRODUCTION/MOTIVATION

In recent years, travelling to new destinations has gained more popularity, especially among the millennials. The quality of travel is not only limited to the kind of destination but is more about the overall experience. From staring at beautiful desktop wallpapers, we have come a long way to constantly browsing through the internet and striving to secure optimum deals on flights, stay, food, activities, car-rentals and others, to get bang for the buck we spend on traveling. Are we always successful in doing so?

This project combines three main aspects of travel - flight, stay and food in the travel package (novel idea). Popular platforms like Expedia, Yahoo Flight, and Yelp etc. provide us with so many options that we get bounded with time to evaluate all the permutations for the best possible offer. This limitation seeds our project - to make a fast, easy, and personalized travel package recommendation system via a query. Many customers would gain from such a "one-stop shop" platform, where they would answer a few travel queries and our system would not only recommend a flights itinerary, but also various options of stay in specific city and places to eat as per the individual's food preference.

2 THE EXISTING TECHNOLOGIES AND STUDIES

Several studies have been performed over the past few decades to build search engines [1-3]. Wissner et al. patented a search system that generates user-customized search results with the use of user-defined semantic types even as late as 2015 [3]. There have been successful attempts to create a search engine for hotel recommendations for Expedia users by training open-source data from Kaggle with SVM and decision tree to direct a given textual query

to "hotel cluster" corresponding to it [4]. Among the most recent studies that tie hotel recommendations for special purposes like vacation or business into flight information was done by Thomas et al. [5] where they devised a system as a cascaded machine learning pipeline, which would be equipped to identify an optimal list of relevant hotels based on all information about the trip. There have also been numerous studies that have used Yelp datasets in order to design recommender systems for predicting user preferences [6], or ratings in general [7].

For a user starting a search for a trip with all three key aspects in mind, the overall experience can be taxing and time-intensive. And even such a tedious manual process can not ensure satisfactory result both in financial and mental aspect. In this project, we would like to leverage the knowledge from current studies made in the individual flight, hotel or food aspects to build a hybrid three-aspect user-customized search engine.

3 OUR APPROACH

In our approach we combine three recommendation systems: 1) food 2) travel and 3) hotel where users can get the personalized recommendations via queries made to three separate databases/datasets. The idea of our approach is similar to the study made on "hotel cluster" recommender search engine [4]. We are going to create feature vectors using word2vec using unigrams with SVM as a classifier (baseline algorithm). And compare its performance to decision tree with ensemble learning on all three datasets. We are going to rank the results based on user query/specific preferences. We will use different different datasets than the other study used then we will then extend that study on to food queries on yelp dataset. As a metric of performance we are going to use predicted accuracy of each one of the results to the ground truth.

(1) For flight details, we have to web-scrap data from Expedia.
(2) Based on the returned flight query results, We will follow the steps suggested in "hotel cluster" study [4] to perform the hotel query based on the nature and location of the trip.
(3) To tie the hotel-flight combination to the food aspect of the trip, we have to use yelp reviews and apply word2vec on them to make the food query.

4 DATASETS

For Flights and hotels we will use the Expedia datasets [8]. For food, we are going to use Yelp dataset based on specific city [10]. Potentially, we may need to do some web scraping for some of the attributes.

5 RISK AND CHALLENGES OF APPROACH

Our work will have several challenges in different fronts.

(1) **Data Extraction:** We are relying on multiple sources of data. To get most recent data sets, we would have to web-scrape from Expedia (for flights and hotels) and Yelp (for food). The challenge with having multiple data sources will then be integrating those data-sets.
(2) **Creating Evaluation Parameters:** To test the effectiveness and validate the proposed methodology, it would be essential to devise a rating system to manually score the

different permutations of packages based on the appropriateness for a given text query. That would give us a parameter to test and quantify the performance of the model that we expect to design.

(3) **Creating Baseline:** As we did some literature review, we came across different data sets. As we are integrating different query systems, like the one done with flight and hotels, or existing Yelp recommender systems, we intend to match our query performances to the work done on the individual avenues.

6 SOURCE CODE REPOSITORIES

We don't have any code repository for this readily available. We have codes for different aspects of the proposed work and as a part of this project, we will make an effort to bring them under one umbrella to build the proposed search engine.

7 RESOURCES

[1] Kruger, Andries, et al. "DEADLINER: Building a new niche search engine." Proceedings of the ninth international conference on Information and knowledge management. 2000.
[2] Gulli, Antonio, and Alessio Signorini. "Building an open source meta-search engine." Special interest tracks and posters of the 14th international conference on World Wide Web. 2005.
[3] Wissner, James M., and Nova T. Spivack. "Generating user-customized search results and building a semantics-enhanced search engine." U.S. Patent No. 9,037,567. 19 May 2015.
[4] Li, Susan. "A Machine Learning Approach - Building a Hotel Recommendation Engine." Medium, Towards Data Science, 6 Dec. 2018, towardsdatascience.com/a-machine-learning-approach-building-a-hotel-recommendation-engine-6812bfd53f50.
[5] Thomas, Eoin, et al. "Cascaded Machine Learning Model for Efficient Hotel Recommendations from Air Travel Bookings." RecTour 2019 (2019): 9.
[6] Nikulin, Vladimir. "Hybrid recommender system for prediction of the Yelp users preferences." Industrial Conference on Data Mining. Springer, Cham, 2014.
[7] Farhan, Wael. "Predicting Yelp Restaurant Reviews." UC San Diego, La Jolla (2014).
[8] Kaggle. "Expedia Hotel Recommendations.", 2016. https://www.kaggle.com/c/expedia-hotel-recommendations/data.
[9] Flight Dataset. https://www.altexsoft.com/blog/datascience/data-science-and-ai-in-the-travel-industry-9-real-life-use-cases.
[10] Yelp Dataset. https://www.yelp.com/dataset.
[11] Kaggle. source code. https://www.kaggle.com/zfturbo/leakage-solution.