WANDERLIST: A travel itinerary prediction tool

Team 8 : Vineeth Aljapur, Zainab Arshad, Ramit Bharanikumar, Sachin Sarath Y Kothandaraman,

Sachin Kumar, Siddhartha Sharma

Introduction

Getting a travel plan done is one of the biggest unaddressed issues in the modern world. There are multiple websites which offer different services like flights and transportation, hotel bookings, restaurant searches and for local travel within a particular city but none of them combine all these aspects to make the user aware of how the trip might turn out to be considering all the factors. In the survey of existing itinerary prediction tools available, most tools work as a tour itinerary planners which require users to manually weigh in all his/her options by listing out the places that they want to visit to generate a timetable of their tour plan. The other complex tools require personalized user data such as social media data, prior trips or activities to generate meaningful trip suggestions. There are no itinerary prediction tool available that can help users decide on the most suitable plan according to their budget.

Wanderlist is a travel itinerary predictor capable of providing meaningful travel itinerary suggestions to users based on minimal input from users - primarily their budget and dates, and not provide every detail of their planned travel.

Definition of the Problem

Our primary goal is to provide travelers a one-stop solution containing a detailed travel itinerary, that includes data about local tourist attractions, restaurants, hotels and air travel based on a budget specified by the user. The current problems users face is having to look for information on hotels, flights on specified dates to their destination that fit their budget and having to crawl through several web sources to identify locations to visit in a given number of days and having to plan an itinerary. Though there are several sources that provide this information. They are scattered across multiple platforms and web services and they require the user to manually select the places that they would like to visit. Further, none of the tools provide an auto-generated itinerary without any user input on the places they would like to visit.

Our tool aims to address these problems by providing the users with information on the best city they could visit during a given season. Identify flights that are cheaper during the specified dates, list out options for hotels to stay and places they could visit each day of their stay. All this information is curated to fit within the user’s budget and the feasibility of visit – mainly the distance.

Finally, the output of our tools output contain the following information:

- Possible restaurants that the user can visit for a given budget, with the rating for the restaurant.
- Cheapest flights that the user can take between two destinations.
- Optimum routing from one tourist attraction/restaurant to another minimizing distance and cost.
- List of tourist attractions that people need to visit in a given place for a given budget.
- List of hotels that a tourist can stay in for a given budget.

Survey

[1] C. Choi, M. Cho, J. Choi, M. Hwang, J. Park and P. Kim, "Travel Ontology for Intelligent Recommendation System," *2009 Third Asia International Conference on Modelling & Simulation*, Bali, 2009, pp. 637-642.
Review: Description of an intelligence recommendation system based on travel ontology. Discusses the storage of content data in two parts - hierarchical information using owl and the instance information in relational databases indicating that hierarchical clusters are a better methodology for implementation.

[2] S. Bao, M. Yanagisawa and N. Togawa, "Personalized one-day travel with multi-nearby-landmark recommendation," *2017 IEEE 7th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, Berlin, 2017, pp. 239-242.
Review: Proposes a travel recommendation system suggesting the best travel route based on the time required for travel and prominent locations in a given area using weight assignment. However, the paper does not take into consideration the dynamic factors involved.

[3] K. Popat, T. Thorat, M. Vidhate, N. Rane and N. D. Ghuse, "Personalized and relational approach for travel package recommendation," *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, Noida, 2015, pp. 930-935.
Review: Proposes a tourism package recommendation model based on TAST model utilizing a probabilistic graphical model to determine weights from the user information provided to recommend specific packages and further modifies it based on a price system to identify immediate neighbours to assign travel groups.

[4] Q. T. Le and D. Pishva, "An innovative tour recommendation system for tourists in Japan," *2016 18th International Conference on Advanced Communication Technology (ICACT)*, Pyeongchang, 2016, pp. 717-729.
Review: Demonstrates a system prototype capable of suggesting optimal touring plans composed of various points of interest (POI) that takes travellers' preferences and context into account. A heuristic greedy search approach is employed for transforming the data into informative routes.

[5] H. Alghamdi, S. Zhu and A. E. Saddik, "E-Tourism: Mobile Dynamic Trip Planner," *2016 IEEE International Symposium on Multimedia (ISM)*, San Jose, CA, 2016, pp. 185-188.
Review: Proposes the Balanced Orienteering Problem combined with a recommender system for tourism suggestions for the mobile tourism guide application developed in the paper providing statistical evidence of comparison studies stating comparatively favourable performance as compared to existing algorithms with faster execution times.

[6] K. Sylejmani, A. Muhaxhiri, A. Dika and L. Ahmedi, "Solving tourist trip planning problem via a simulated annealing algorithm," *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, 2014, pp. 1124-1129.
Review: Presents a simulated annealing algorithm for solving the tourist trip planning problem. The hard constraints include limited trip duration, working hours of points of interest (POIs) and tourist budget, while tourist preferences and traveling time comprise the soft constraints, which take part in a two-component function for evaluation of quality of solutions.

[7] Y. Xu, T. Hu and Y. Li, "A travel route recommendation algorithm with personal preference," *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Changsha, 2016, pp. 390-396.
Review: Proposes a personalized travel recommendation system based on preferences by the general public and real-time congestion between different spots. However, historical data for traffic congestion is used in the algorithm due to lack of real time data.

[8] S. S. Arote and R. L. Paikrao, "A Modified Approach Towards Personalized Travel Recommendation System Using Sentiment Analysis," *2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*, Sangamner, 2018, pp. 203-207.
Review: Aims to develop a personalized recommendation system for users based on sentiment analysis from social media. Proposes to utilize a TSP algorithm understanding the common tourism related times and destinations with social media information used as a source of data.

[9] W. G. R. M. P. S. Rathnayake, "Google Maps Based Travel Planning & Analyzing System (TPAS)," *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, 2018, pp. 1-5.
Review: Details the utilization of Google Maps API to sort multiple intermediate destinations based on distance from a start point provided by the user. Our approach can take similar into consideration while designing city-based itineraries depending on the time available to the user in a part of the city.

[10] Daniel Guttentag (January 2013) “Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector”
Review: Details Airbnb’s rise in popularity in a relatively short period of time by market disruption. It explains the potential shortcomings of Airbnb like service quality, friendliness, and customer security and provides a viewpoint to justify our selection of Airbnb as a data source.

[11] W. Groves and M. Gini. “Optimal airline ticket purchasing using automated user-guided feature selection.” In IJCAI ’13: Proc. 23rd Int’l Joint Conf. on Artificial Intelligence, 2013.
Review: Addresses the correct time to buy flight tickets using a prediction model which uses feature extraction, fine-tuning, constructing partial least square regression and selecting the optimal model. It suggests that expertise in hierarchy construction improves efficiency by having better feature selection.

[12] Liu, Xudong, et. al. “Combination of Diverse Ranking Models for Personalized Expedia Hotel“ URL: http://arxiv.org/pdf/1311.7679.pdf
Review:
Describes usage of various models like logistic regression, random forest, lamdaMART, deep learning to rank hotels to maximize purchases. Details the shortcoming of deep learning as lacking training data which we try to address by combining different data sources to generate a large enough dataset to overcome our problem.

[13] A. Hicks, S. Comp, J. Horovitz, M. Hovarter, M. Miki and J. L. Bevan, "Why people use Yelp. com: An exploration of uses and gratifications.," *Computers in Human Behavior, 28(6), 2274-2279.,* 2012.
Review: Aims to determine why customers use Yelp based on a 144-user survey, concluding the reasons of usage for information-seeking, followed by entertainment, convenience, interpersonal utility, and pass time by a large disparity.

[14] A. Noulas, S. Scellato, C. Mascolo and Pontil, "An empirical study of geographic user activity patterns in foursquare.," *M. In Fifth international AAAI conference on weblogs and social media.,* 2011.
Review: Describes the results of a user behaviour study in Foursquare of over 700 thousand users over a 100-day period, concluding that user activity is variable across the day and week. Our application aims to take this analysis into account to provide precise recommendations from a temporal and spatial point of view.

[15] Taylor, D. Christopher and J. B. Aday, "Consumer Generated Restaurant Ratings: A Preliminary Look at OpenTable. com," *Journal of New Business Ideas & Trends 14.1,* 2016.
Review: Describes the variables influencing restaurant rating on OpenTable.com and the effect of pricing and number of meals served daily on the ratings using statistical analyses such as correlation and ANOVA.

[16] Vermeulen, Ivar E., and Daphne Seegers. "Tried and tested: The impact of online hotel reviews on consumer consideration." Tourism management 30.1 (2009): 123-127.
Review: Describes the various factors considered when an end-user is making hotel bookings online and the impact of reviews on this decision-making process. This provides a good idea while building a hotel database for our application. However, the paper fails to take into consideration the genuineness of the reviews.

[17] Kang, Hanhoon, Seong Joon Yoo, and Dongil Han. "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews." Expert Systems with Applications 39.5 (2012): 6000-6010.
Review: Suggests an algorithm to improve the accuracy of the sentiment analysis of restaurant reviews by the use of a senti-lexicon along with a Naive Bayes algorithm to remove the bias towards positive echoed sentiment in reviews and provide a bias free review.

[18] Jeacle, Ingrid, and Chris Carter. "In TripAdvisor we trust: Rankings, calculative regimes and abstract systems." Accounting, Organizations and Society 36.4-5 (2011): 293-309.
Review: Develops upon the concept of commercialization of trust and how it has helped TripAdvisor become a trusted source of travel information. It provides us with great insight as to what should be taken into consideration while building our application for it to be successful.

Current practices and its limitations

The current applications in existence provide an end-user with a lot of information regarding travelling to a new place. Application like Expedia, TripAdvisor, Yelp and Airbnb provide details about travel, accommodation, restaurants and tourist spots. Travelling within the city can be done through the usage of Google Maps.

However, the above-mentioned options provide more of a general database of the knowledge necessary to build an itinerary but do not cater to the specifics of the user necessities, forcing the user to go through every application multiple time to build the correct itinerary for themselves. There exist no options which integrates all the above-mentioned information and provides a basis of comparison and sorting through based on user requirements and outputs this curated information into a singular itinerary planner catering to all the specifications of the user.

Data Collection

We have accessed data through multiple APIs and scraper for setting up the databases as well as retrieving data in case of specific inputs from the user. The following sources address each of the datasets:

- RapidAPI – For collecting the flight price info through the Skyscanner API. To accommodate for the low cost, the flight info has been collected for 6 months in advance for 3 days per week. Specific flight prices and dates in case of user input are queried real-time.

- Yelp Fusion API – For collecting restaurant information and user reviews including restaurant rating, price levels, location and cuisines for five cities. Results for the query are sorted by an adjusted rating value that considers the number of ratings, similar to a Bayesian average.

- Hotels - Used a scraper available online for collecting hotel data from TripAdvisor for the given constraints implemented. The results can be collected based on Ratings, Popularity and Price. The scraper provides details about the hotel facilities, price per night for an individual traveler and number of reviews. Another scraper is available for use to capture the address and overall rating of the hotel.

- Google Places API – For collecting places of interest in the five city options. The places info is retrieved as a JSON and stored on to the DB.

- Google Directions API – For mapping routes between places of interest for a default of 3 days of trip and adjusted according to the user input for dates.

Limitations for data collection

Collection of hotel data: There exists no freely available API currently which allows collecting data for hotels and Airbnb stays. The available scrapers as such are highly limited in functionality and thus, there is no available technology which allows us to collect hotel data on a large-scale basis. Further, we are currently trying to modify the available scrapers to collect data on a large-scale basis without running into any issues. Hence, for accommodation purposes we are only using hotel data that was obtained from tripadvisor.

The skyscanner API only allows a total of 50 queries in a day, and hence all team members are using an API to generate more queries.

The Yelp Fusion API can only return a total of 1,000 businesses at this time for a specific location. The API doesn't allow more than 50 per request. You need to use the offset parameter to get the next page of results.

Proposed method

The proposed method takes a little revision from the initial proposal with databases being setup for a select few cities for a few months in the future and with a default budget and number of days of trip as a default. This is to tackle large number of combinations in case of no inputs provided from the user. The current implementation would cover 5 cities, 6 months ahead, 300$ budget and 3 days of trip as default constraints.

6.1. Scenario 1 - Input dates, budget and source

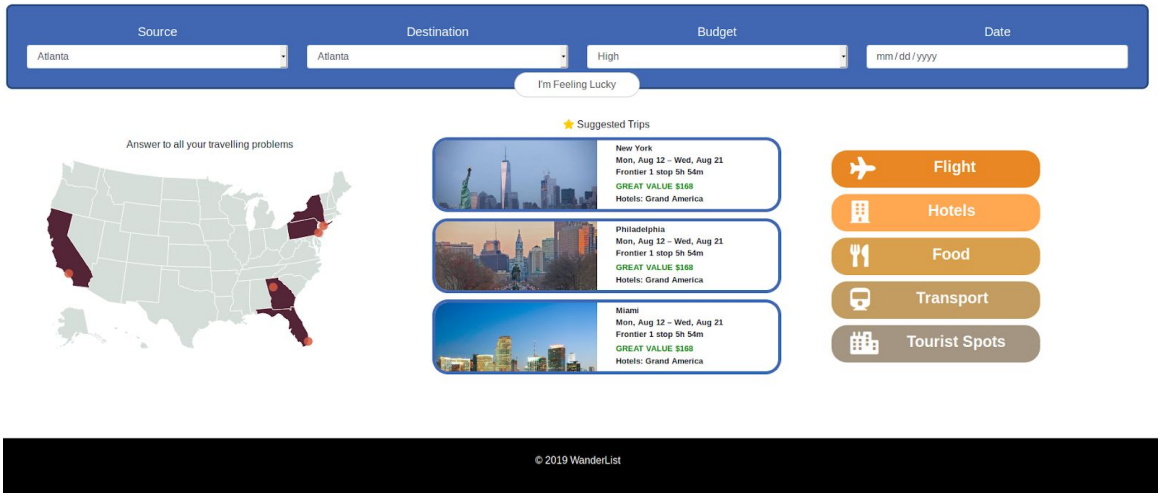
This would consider the input from the user for the dates, their source city and a given budget and the specific hotel reservations and flight prices would be retrieved for those dates and the city with the cheaper prices for this combination for the given date will be displayed to the user. An itinerary for this result will be generated.

6.2. Scenario 2 - Input dates, budget, source and destination

This would consider the flight and hotel reservations for the city specified by the user on the given dates from the source city. City specific itineraries are displayed to the user with the budget.

6.3. Scenario 3 - I am feeling lucky!

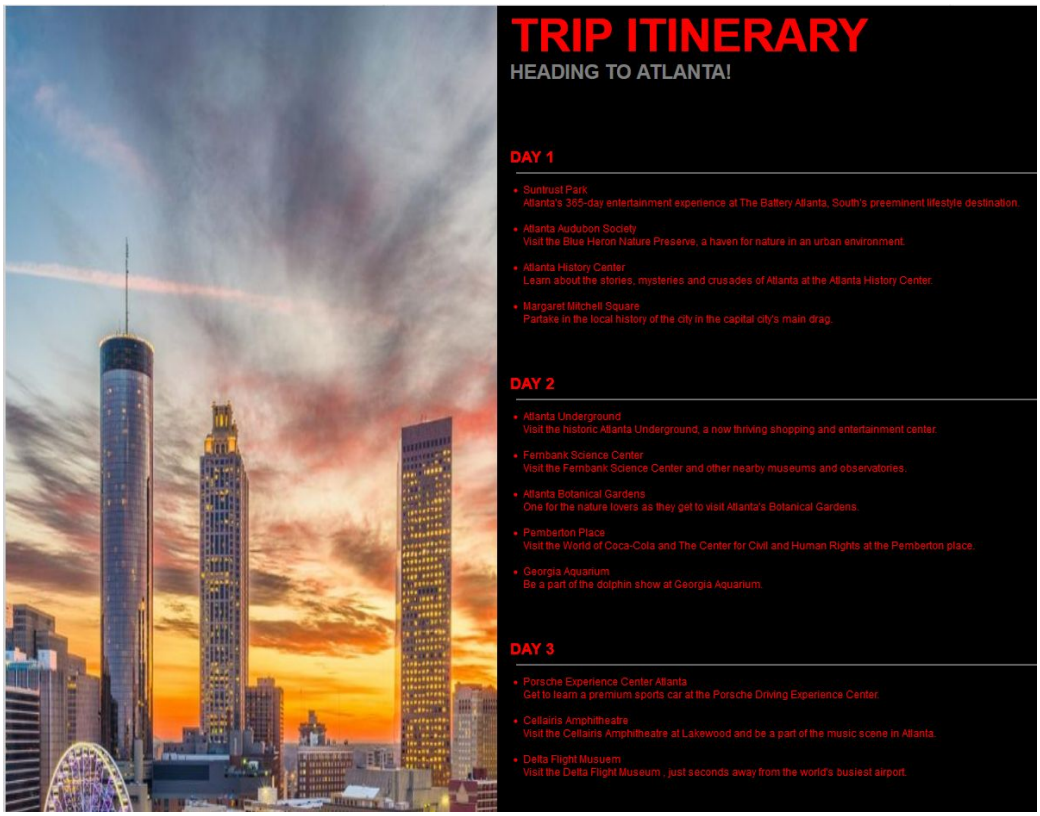
This would keep a default minimum budget range and identify cities offering combinations of flights and hotel rates within the given budget for a specific season (recent dates) for the user to display constrained itineraries fitting the budget. Here, the weight assignment becomes critical to balance out the flight, hotel and suggested restaurant prices.



The data obtained on restaurants, flights, hotels and places of interest were cleaned and a database was created on mySQL local server. The possible combinations of prices for hotels, flights and restaurants were extracted and filtered based on the budget criteria. A cross join was implemented to combine the data points from the results and a projected expense is identified for the user's entire period of stay. This value is then compared with the user's budget and the combinations that satisfy the budget requirements are filtered out, then sorted based on ratings.

These values are then clustered based on their coordinates to obtain clusters of places and restaurants nearby. We clustered the restaurants and places of interest data using the latitude and longitudinal coordinates using k-means clustering. This clustering aids in minimizing the need for individuals to travel far and gives a fair sense of idea as to which restaurants are close to each other.

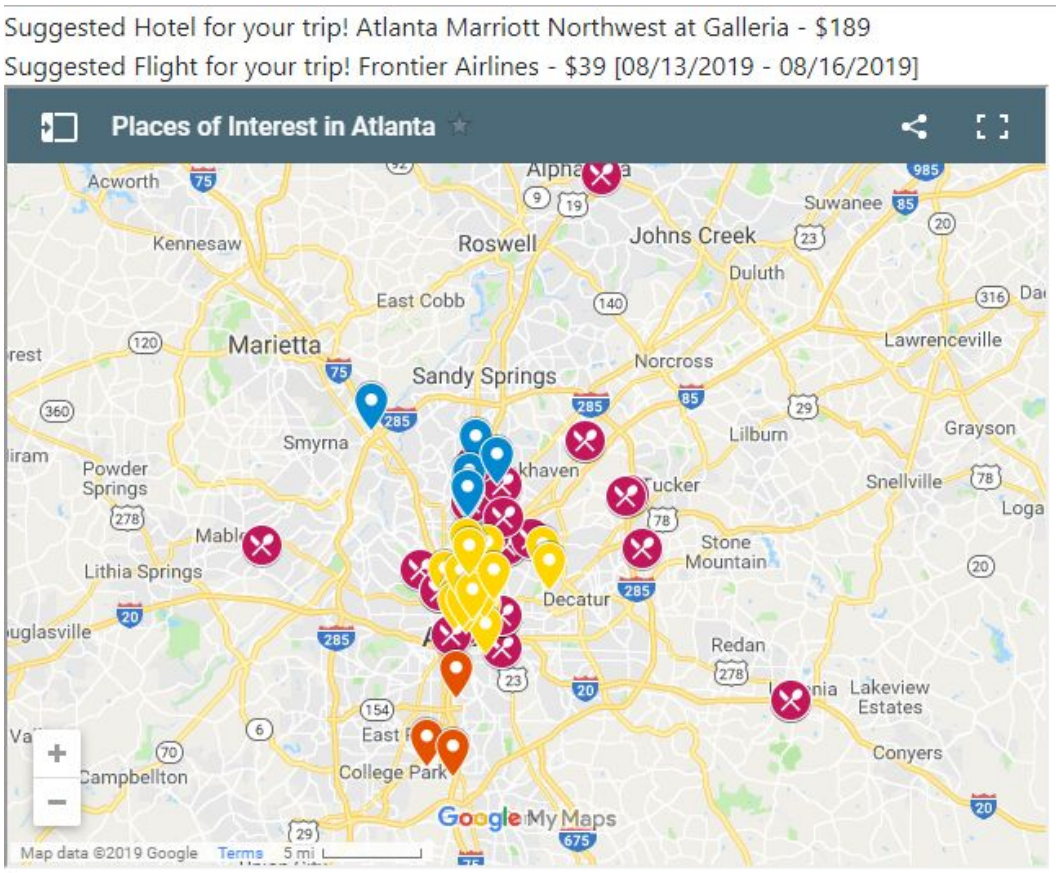
The K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster based on the nearest distance. For every cluster the algorithm calculates a centroid using mean of the data points and considers this to be the new centroid. This process is repeated until the data points do not change the cluster or to a specified depth. The clusters formed between the places of interest and the clusters for restaurants are compared to get pairwise distance between centroids. This distance matrix is used to suggest the restaurants based on places of interest, thereby providing the best possible combinations of places to visit and the restaurants nearby for each day of your itinerary.



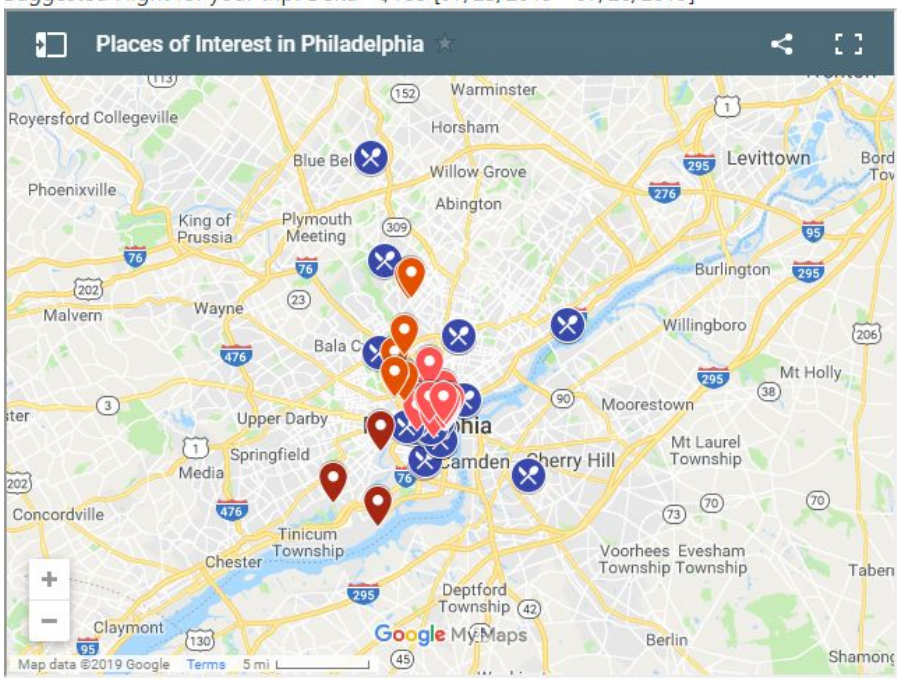
Visualizations

Finally the obtained results from the clusters' places of interest and restaurant pairings are shown as a list to the user. The user will be provided details on the lowest priced flights for the specified dates from the source to destination. The list contains the list of places to visit each day, and choices of restaurants with their distances specified.

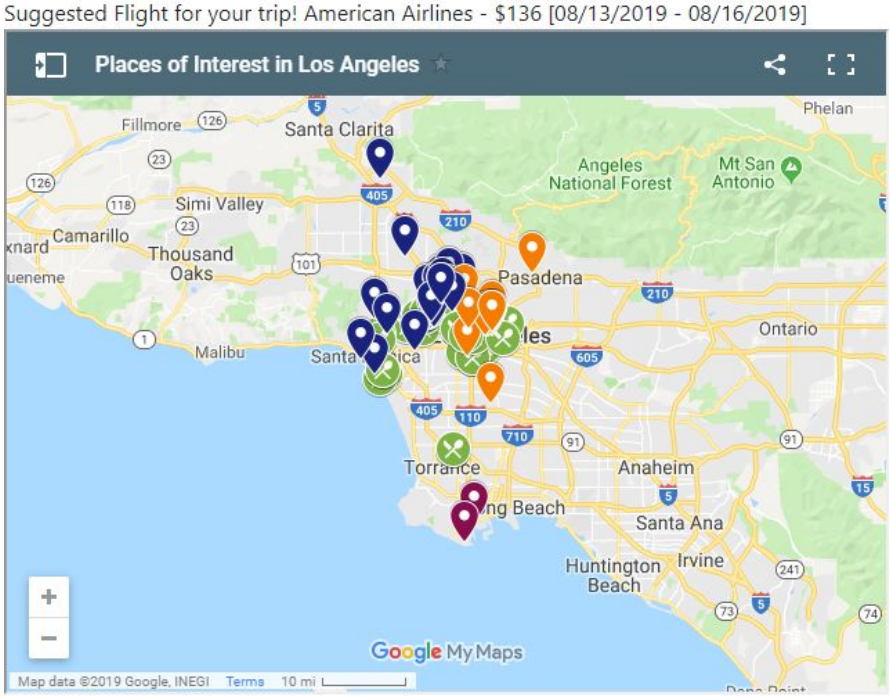
A google map functionality visualizes the optimal routes for the clusters to visit the listed places for each day utilizing the shortest path algorithm, for the user to verify the places and their relative distances.



Suggested Hotel for your trip! Akwaaba Philadelphia - $212
Suggested Flight for your trip! Delta - $185 [07/23/2019 - 07/26/2019]

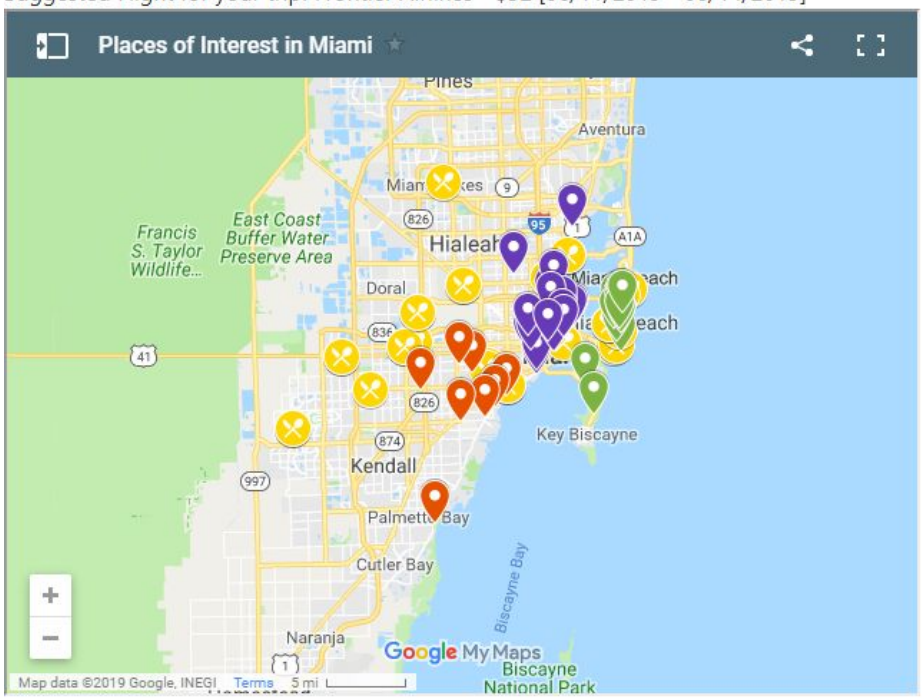


Suggested Hotel for your trip! Hotel West Hollywood - $349
Suggested Flight for your trip! American Airlines - $136 [08/13/2019 - 08/16/2019]



Suggested Hotel for your trip! Arlo Nomad - $281
Suggested Flight for your trip! American Airlines - $79 [09/18/2019 - 09/21/2019]

Suggested Hotel for your trip! Beacon South Beach Hotel - $193
Suggested Flight for your trip! Frontier Airlines - $52 [06/11/2019 - 06/14/2019]



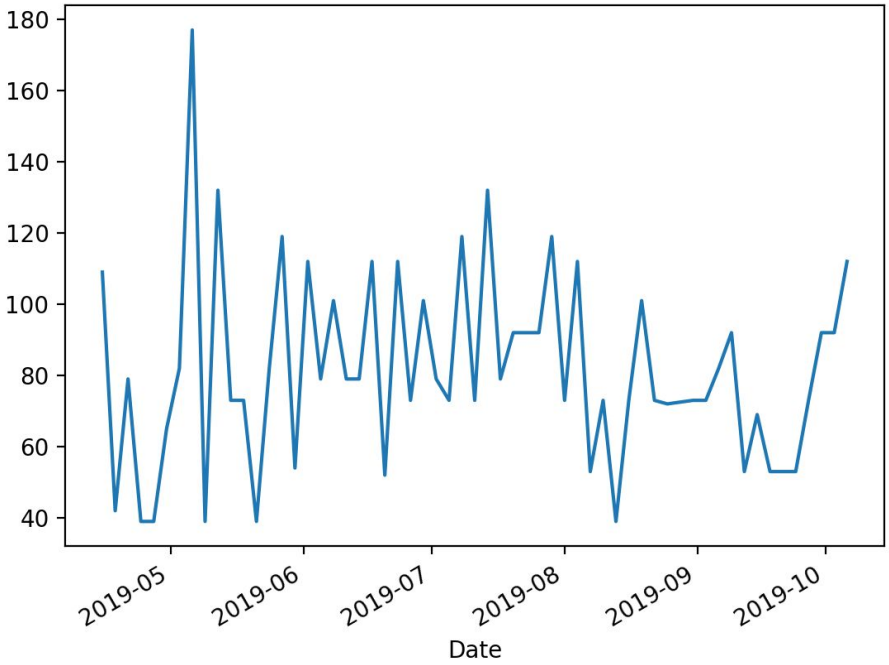
Innovations

- The model utilizes key components in a given itinerary, i.e., flights, hotel bookings, restaurants and local tourist attractions. The model is capable of predicting the destination based on the seasonality and provided budget by the user.
- Further, the model can automatically generate an itinerary of places to visit in a day by clusters identified from their relative distance providing the user a clear plan and ensuring ease of locomotion between these locations by keeping the distance within 5 miles.
- Provides the user with necessary information about the itinerary, such as a map of the shortest route to take between the listed places for each day, information on restaurants near the specified places of interest, price, and ratings so that the onus of a trip's sketch doesn't fall on the user.

Experiments and Evaluation

We explored using hierarchical clustering and CLARA for the clustering, along with the K-means algorithm. We generated results for 60 sets of dates. In 180 combinations each for 5 cities we identified that the k-means clustering was more effective and decided to use k-means as it was faster.

- We use the R-squared analysis to evaluate our predicted results. The model comprises of three variables - Flights, Hotels, and Restaurants. The prices from these variables will form the basis of our model. The parameters are assigned weights and clustered to identify clusters that fall within the required user budgets.
- We also perform a time-series graph to identify the prices to determine the seasonality of the rates between the cities. This helps to suggest a particular city to the user in case the user does not enter the destination city and only the budget.
- We are utilizing the shortest-path algorithm from google directions API to obtain the shortest path in between locations to determine the places to visit between two given locations. This helps to identify if the clustered results are consistent with the real-life scenario based on the coordinates.
- Once the application is automated and made available online our model follows the existing informational websites like tripadvisor, to gain user reviews on the usefulness of the itinerary generated based on their given inputs.



The above image is a time series plot for a trip to be taken between Atlanta and Miami, users are provided with this data for each origin and destination for a given date, so that they can take a more informed choice as to when would it be an optimal time for them to travel from a budget perspective.

Results and Future Work

Currently, the database has been built and model functions well. It has been tested to produce consistent results based on the specified input. Further the web interface development is complete. However, further steps are required to link the web interface to the backend to create a docker capable of functioning automatically. We have provided the sample data and web interface along with results obtained for specified runs along with our report. Our scope for future work involved classifying the places to visit according to the user's interests and providing a better-curated list of places to visit specifically to the activities the user would be interested in.

At present we have run the model for 60 sets of dates to generate output itineraries. Since we were not able to integrate the frontend and backend, we are generating results for a single run fo your analysis. The HTML scripts and maps are provided to verify the results along with the algorithms.

The frontend files uploaded include the Home Page, the itinerary values for 5 cities created for 3 different budget ranges and webpages showcasing the clustered places of interest and all the restaurants mapped into a dynamic Google Maps visualization.

Since we could not complete the web interface in time, we were not able to completely link the website to the backend and thus could not release the website as a completely functional application, but the back-end analysis and visualizations function for the 5 cities and 6 month range included in the dataset.

Conclusion and Discussion

The report introduces Wanderlist, a travel itinerary prediction tool that is capable of auto-generating a travel itinerary specific to a user's budget. The tool takes in user provided budget and dates to suggest a city to visit by seasonal analysis, uses k-means clustering based on relative distances to provide an itinerary of the places to visit on each day. The tool also provides a google map visualization of the routes to

We believe that our tool addresses an essential problem in tour planning, especially within a user's budget constraints, by reducing the user's effort in collecting information such as places to visit, flight prices, hotel and restaurants to choose in a given locality and drawing out an itinerary based on this information. By providing detailed information to each of these stages and generating an itinerary and planning their routes we remove the hassle of planning a vacation and worrying about the budget, the tool enables ease of travel and reduces user effort in searching for each of this information. This serves a very important function of assisting the user in providing a well-curated itinerary which can be used easily by the users for planning their vacation.

Effort Distribution

Everyone has an equal contribution to the tool development, document, and to further activities.