CMPT 732: Big Data Lab 1 (Fall 2018) – Project Proposal

Date: 15th October 2018

The Battle for Bookings: A Comparative Assessment of Hotels and Airbnb's on a Global Scale

Team CodeNinja: Anurag Bejju – abejju@sfu.ca, Andrew Wesson-awesson@sfu.ca

---

Problem Statement:

Based on a recent study, Travel, and tourism is one of the world's fastest-growing sectors, with bookings hitting close to $1.6 trillion in 2017[1]. Each year, the global traveler pool is flooded with millions of new consumers from both emerging and developed markets, many with rising disposable incomes and a newfound ability to experience the world. With most travelers trying to get the best bargains on basics such as accommodations and transportation, we wanted to analyze the cost of accommodations around the globe. Since there are many options available, tourists either opt for hotel rooms or turn to Airbnb: an online community marketplace to plan their stay. As part of our Big Data project, we decided to compare the costs of hotel and Airbnb rates in 88-cities around the world by leveling the measuring parameters to see how the deals stack up.

Data Source:

Airbnb Data: InsideAirbnb[2] has made Airbnb data accessible by using publicly available information from the Airbnb site and provides a snapshot of searchable listings with various parameters like dates, prices of future available dates, reviews, and locations with listing metadata*. (Data Size – 88 Cities * 27000 listings each (approx.) * 100 parameters => 2.4 million * 100 records)*

Hotels.com Data: Web scraping techniques will be used to extract publicly available data about the hotel name, street, locality, guest reviews, rating and room pricing for all 88 cities.

Global Rental Data: Upon request, numbeo[3] has provided free access to their API (worth US$220 per month) which gives latest rental price information for 549 cities around the globe. This API will provide monthly rental prices for 1 bedroom and 3 bedroom houses.

Analysis:

In order to give an accurate estimate of the cost difference between Airbnb and hotels for each city, we need to account for other variables influencing price. These may include but are not limited to amenities, number of guests, and location. To prevent these factors from skewing the results, rather than simply taking the difference of the mean costs of each within each city, we will search for the hotel that is most similar to each Airbnb with respect to these criteria and compare their prices. We can then use this information to create a map of which option is more affordable in each location. We also intend to compare reviews and ratings in order to fully evaluate traveler preferences as well as get a snapshot of the satisfaction index of a customer.

Tech Stack:

Even though the Airbnb data is well partitioned based on the city, the hotels extracted data requires a lot of preprocessing to be done. Since we anticipate to compare them by leveling all important parameters (like the number of rooms, location etc), we have to convert the address of the hotels to geo-coordinates using the geoPy library. Later we would use geospatial and other search techniques to map all the 2.4 million Airbnb listings to its closest hotel listing. Magellan[4] will be used on top of Apache Spark in order to perform geospatial analytics on the data in an efficient and scalable manner. This library will be used to match Airbnb to the nearest hotel in order to compare the prices of each in similar locations. As this library does not have a Python API, we will need to use Scala for this portion of our data processing pipeline. After performing aggregations and analysis, we will use Geopandas[5] on the reduced dataset for further analysis and visualization.

References:

[1] Douglas Quinby, Phocuswright Conference, Florida, November 9, 2017.
[2] http://insideairbnb.com/get-the-data.html
[3] https://www.numbeo.com/cost-of-living/prices_by_city.jsp
[4] https://github.com/harsha2010/magellan
[5] https://github.com/geopandas/geopandas