Capstone project

Hotel Recommender system

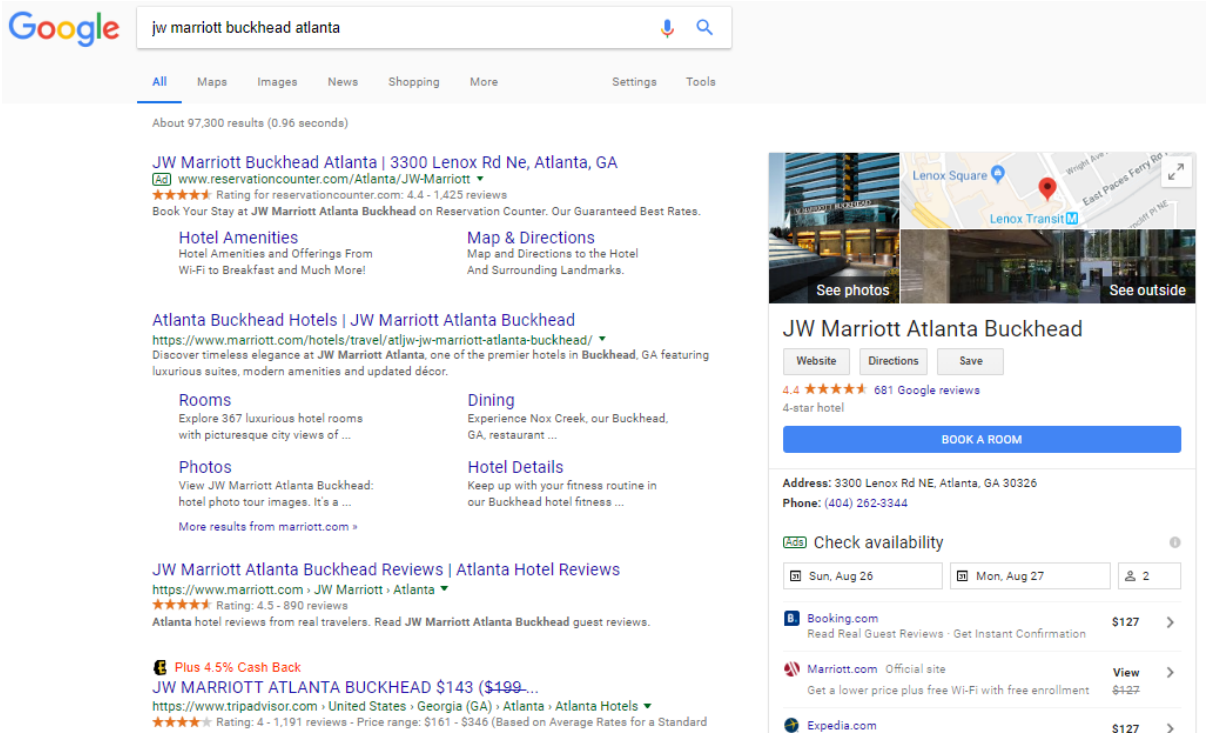
Data wrangling

A set of hotel ratings were obtained from https://github.com/kavgan/, dataset on which she published a 2011 scientific article [1]. The data consists of four .csv files with a hotel name, id, link to reviews from Tripadvisor, address, number of ratings and average rating for a set of parameters. The data is essentially clean, but obsolete. Some street names or numbers are missing.

In less than 10% of the data there was a missing zip code. Those rows were dropped.

To complete the dataset I used python selenium to search on google the hotel name and city, and find the star classification (1-5) and the mean price of the ones listed on the hotel description on the right. In the absence of details, it is filled with -1 to maintain the number structure.

For the recommendation system, if there is no price or rating available, the algorithm should use a ratings-only approach as it has been implemented elsewhere.



To create a set of virtual users, references 2-6 were used as model. Initially, a virtual user would be given an age and a zip code. A mean household income would be selected according to the demographics, and a trip purpose. The economic status of a traveler is, according to many references, the most important factor in deciding on the hotel, and it will be in the model. Then, the virtual user picks a hotel randomly, and rates it according to a simple set of conditions. The data generated will agree with the averages obtained from the dataset.

This platform is a growing system, so data management is important. A simple database in SQL will be the starting point, with a set of tables:

- Hotels: contains a hotel ID, its information, price, and mean review scores
- Users: has the user ID, username and password, city and zip, and its age. Gender or name does not seem to play an important factor in the decision-making.
- User ratings: each rating will have a record with the hotel ID, user ID, and the values. No open text required.

And initially, virtual users and virtual ratings. The recommendations will be calculated *ad hoc* for every search based on the most recent information.

---

[1] Ganesan, Zhai, (2011) “Opinion-based entity rating.” J. Information Retireval.

[2] Min, Min, Emam, (2002) "A data mining approach to developing the profiles of hotel customers", Intl J Contemporary Hospitality Management, Vol. 14 Issue: 6, pp.274 - 285, https://doi.org/10.1108/09596110210436814

[3] Palakurthi, Parks, (2000) "The effect of selected socio-demographic factors on lodging demand in the USA", Intl J Contemporary Hospitality Management, Vol. 12 Issue: 2, pp.135-142, https://doi.org/10.1108/09596110010281791

[4] Minnaert, (2017) NYU US Family Travel Survey. https://www.scps.nyu.edu/content/dam/scps/pdf/200/200-4/200-4-16/P1718-0036-2017_Family_Travel_Survey.pdf

[5] Bureau of Labor Statistics, (2010) Spotlight on statistics travel. https://www.bls.gov/spotlight/2010/travel/

[6] Dolnicar, Otter, (2003) “Which Hotel attributes Matter? A review of previous and a framework for future research”