Travel recommender

First note:

This seems to be the first data science challenge hosted by deloitte and tdwi. For future challenges it would be a good practice to follow the guidelines outlined by "Kaggle.com" the main place for data science challenges. My main problems where:

- There was not clear description what exactly has to be done other that "build a recommender engine", "use data" and "train a model".
- There was no "real" provided dataset

I'm looking forward how you will evaluate all given codes/project because without a given dataset and clear description this seems to me the evaluation would be difficult.

data used:

because there was no real dataset provide I created my own using web scrapping from

https://www.worldtravelguide.net/country-guides/

I did not uses tripadvisor data because as stated on their homepage the API is not intended for data scrapping.

data storage:

is used a Sqlite3 database. Because of the small size of the dataset (size under 100mb) I did not see the benefit of using a noSQL database like graph based neo4j. But it would be simple to store my information in a graph database if necessary.

nlp:

For natural language processing (NLP) is used a combination of term frequency - inverse document frequency (tf_idf) and latent semantic analysis (LSA). The result of the tf_idf is used in the LSA model. The importance of words (output of tf_idf) is used in the concept building of my LSA model. Singular value decomposition is used to reduce the number of rows. I used numpy to calculate the best number of topics which would be n=95 in my case. This seems low but only because I used every column in my database independently. So the LSA model for the "about" page is independent from the LSA model in "culture". This guarantees an individual concept for each column. Improvements could be made by using user input data (for example what information is requested most often) to combine for example "history" with "culture" if they are requested together.

Secondly, I used a simple weighted measure for certain types the users request. Implemented is "culture", "nightlife" and "activity". The matrix of my weights can be seen in the table "matrix" in the database. Here is much improvement needed. Because I don't know much about tourism / travel an expert would be needed to evaluate which types of travels are requested most often and what kind of service the datasubject values in each type. I just gave it my best guess.

future:

The designed model is based on a cold start concept. There are not ratings or any user input involved. A future design should use user input data (for example just thump up / down) do reevaluate the LSA Model.

web interface:

I'm not familiar with flask so I did not use the provided app for the web interface. But it should be easily possible to GET/POST the provided data.

usage:

you just need to create a virtual environment and install the req.txt, go into the main folder (travel_recommender) and run the app.py.

You can reach me at:

Benjamin Pohl
Benjamin.pohl95@gmail.com

Thank you for the challenge J