Continual Assessment 2 - Individual Report

What were the main challenges faced during the data preparation and how were these solved?

Sourcing of data to fit the project's requirements

The dataset that the team eventually settled upon is Yelp's data on businesses, reviews and user data. The data is comprehensive, and sufficiently meets the team's objective of creating a restaurant recommender. The dataset comes in 7 files and segments the data into business location (details and review count and ratings), business attributes (meta information like Bike parking, restaurant attire, etc), business hours, check-in details, review, tip (similar to review, but condensed version), and user information.

The available data was more than sufficient for the team's objective of providing a restaurant recommender. In fact, the amount of data was big to the extent that the initial exploration caused the system to crash.

The team thus decided to firm up the business outcome and then select the required dataset. Data on business location, review and user information was selected. Tip was dropped since review data contain more textual content for demonstration of text analytics. The team decided that although data like business attributes are a good-to-have, they are too granular to be listed in full in a recommender's output.

Merging and subsetting data

There were few missing values in the data and not much cleaning was required, barring conversion of some review text to ASCII. However, some challenges came up during data merging from the 3 files. The initial approach taken was to randomly select 100k reviews and join it with the business dataset. However, we found that the dataset was to diverse, and businesses actually tagged under multiple business categories, not merely restaurants. The resulting merged set is sparse with almost 25k unique users. A large number of restaurants had few reviews.

An alternative approach was taken instead and we revisited the data preparation stage. Filtering was performed to scope the data, by selecting only business from Nevada and businesses categorized as 'Restaurants'. 2000 restaurants were then randomly selected before merging with the review data. This ensured that there is a greater number of reviews for each restaurant.

What were the main challenges faced during solution building and how were these solved?

Cold start issue

At the onset, the team decided to address the cold start issue. One of the business objectives is to allow a new user to come on board and still enjoy recommendation based on his or her preferences.

To get around the cold start issue, the team needed to seek explicit feedback from a new user. Given the data on hand, where each business is tagged with one or more categories, and availability of fairly lengthy reviews, it was decided that two inputs will be solicited from the user. A cuisine term to match to the one of the categories and a list of keywords to be matches to the review text.

Even though there is still an absence of information about a new user's visits to restaurants, the inputs enabled further filtering of the 2000 restaurants and helped to map similar users whom the content-based (CB) and collaborative filtering (CF) recommender systems based upon.

Evaluation of models

A challenge the team faced was in the evaluation of the two models which will affect how the results can be combined since the recommendations from each of the system have to be compared. Each result set would most likely have restaurants that are common and uncommon to both sets, so a means of comparison is required.

The results from both models had no basis for comparison since the CB model essentially outputs the ranking of the normalised score of the ratings by the expert user on each feature. This is score is normalised to 1 and it is indeed the ranking that is of more importance. In the CF model however, is a prediction based on ratings from 1-5 as given in the Yelp dataset.

To standardise both sets of result and to have a basis for comparison, the score of the CB model was normalised from 0 - 1 to 1 - 5. The benefits are two-fold. First, it allows the CB results to be evaluated against the rating provided on the restaurants. Secondly, it allows comparison to the CF predicted ratings. Eventually, the final recommendation is derived by applying the error-weighted penalty (based on RMSE) on the CB and CF results.

What conclusions and individual lessons have you learned

Remember the business context and aligning to business outcome

There were round trips made at each stage of the project as the team the business outcome was not firmed up to a certain level of details. For example, while sourcing for data and performing the initial data exploration, there was no consensus on the deliverables and features of the recommender system so time was wasted looking at multiple datasets in detail.

Data manipulation and large data sets

This is especially important when it comes to working on a dataset end-to-end, as opposed to prepared data for workshops. Time can be saved when there is familiarity with manipulating the data into the desired forms. In dealing with larger datasets, when factoring the computing power we have on hand and the time taken to execute certain commands, considerations like performance of the script and optimization comes into play. What I also found helpful is filtering and subsetting of unnecessary data to reduce the memory required for processing.

Project management

Some parts have dependency while others don’t. Project management is an important aspect that is often ignored in small-group assignment like this. An assignment of a project manager role to one of the team members would help in the delivery and time-management.