Nikhil Jagatia

Data Science Task

Introduction

This dataset consists of a subset of different businesses, their reviews, user information, photo information and tip information. Each subset contains fields that have descriptive attributes such as 'business_id' , 'caption' ect. Each dataset is in the form of JSON files therefore to be able to be read through a textfile function the JSON module needed to be imported. The datasets were located on a directory based in Poincare's Hadoop File System which was available to be accessed and then fed through into Apache Spark by using Pyspark and the . Each subset of datasets had the given attributes with their details, in a Key-Value style structure.

Goal and importance of Data Science Questions & plan for each question:

- *Which reviews have been rated as useful by more than 30 users and funny by more than 20 users?*

Having a count of 30 or higher and 20 respectively it shows that the reviews would have a higher validity than a lower count. In this situation with the more reviews being counted as useful it showed how accurate the initial review is, and funny is the accuracy of the review in the context of the business which invokes humour as the emotional response from a relatable occurrence. Yet the reason a person would find this useful or funny whether it is in a positive way or negative is not given.
This is achieved on Spark by using 2 filter functions where each condition 'useful' and funny' will be defined using a lambda function and then set using a boolean operator which in this case is their number (30 and 20). Take 2 shows the incidences.

- *Which businesses based in Las Vegas that are identified as Nightlife have been rated 4.5 stars or higher?*

Due to the nature of Las Vegas being a tourist hot spot it is not unreasonable why people would want to know the best places to visit. This is a good reason of why a rating of 4.5 stars is selected as a threshold.
The file used to perform analysis is the business file which is imported and used as a RDD. Following this the city will be filtered as Las Vegas, followed by another boolean to filter 4.5 stars this will be a new RDD. After, the term 'Nightlife' will be passed through a filter given the key attribute 'categories' to select from. Finally, it can be assigned as a new RDD which then can be inspected by using *first* and *count* to identify the satisfying features.

- *What are the top-10 reviewers, in terms of the absolute number of reviews marked as useful by other users, of Nightlife businesses in Urbana-Champaign?*

This is important as it can show who are the best in terms of well documented users and their reviews. Investigation into the overall reviews against businesses can be beneficial however the joining of the useful and business RDDs would be joined via the 'review_count' being the mutual key formed by the map function. This is done after the usefulness is ranked from highest to lowest, and the two cities are merged.

Nikhil Jagatia

Data Science personal question

- *Which State out of Arizona (AZ) and Wisconsin (WI) has the greatest percentage of users that have liked the tips given for restaurants that are also rated greater than 4 stars?*

This question aims to compare the difference in the two states (AZ and WI), not only would it assess the difference in the number of restaurants which an individual such as a tourist might enquire about it would show the level of involvement of those that attend those restaurants that advise potential customers
This would be achieved by having and comparing two separate RDDs whereby they would be filtered to obtain the satisfying conditions, then reassigning the Keys. After the likes would then be ranked by setting a threshold of greater than 0 to find the liked instances. Finally each RDD for AZ and WI would be joined separately to the ranked tips RDD. This count can then be used when finding the percentages in the given states out of all the restaurants, to which have been liked.

Conclusion

Those reviews that had gained over 30 ratings for useful and 20 for funny had a count of 4011 which shows a large number that met both criteria. Ratings are a good classifier of how good something is, this is reflected by the star ratings, as found in Las Vegas where Nightlife was selected as the category, only 382 satisfied these variables, if the number of stars was purely set to 5 this would be a lower number which could infer as being the very best nightlife spots in Las Vegas, which could be very useful to individuals such as tourists.
Finding the best reviewers in the specified cities raised a problem as finding a those that show '*Urbana-Campaign'* did not exist, therefore they required to be searched individually before merging the two, this was checked as both are located in Illinois, USA. After joining this to the ranked 'useful' RDD the list showed the names of the top 10 reviewers were (in descending order with useful count): Chuck (36907), Maria (4710), Jose (2379), Paige (1818), Anna (1458), Benny (1314), Russ (1169), Elizabeth (1142), Ashley (1060) and Aaron (954).
Finally, the findings of the personal question showed for those restaurants with over a 4 star rating 1412 were from AZ and 200 from WI. Therefore the percentage of those that had been liked were 10.5% of WI restaurants and almost double for AZ (25.4%). This shows an individual could identify restaurants to be more popular in Arizona which could be a reason their tips gained more likes. A cause could be due to popularity where individuals who read the reviews of the desired restaurants require the tips in order to have a better customer experience. However just because there are greater numbers for AZ it does not directly infer that those restaurants are better, as there could also be people that have visited the restaurants that did not document or respond by creating reviews. Further work that could be done with this dataset could be by identifying the correlation between the star ratings given to restaurants and those that have commented on text such as 'very tasty' . Although even if correlation is seen, again it should be noteworthy that correlation does not infer causation, with the addition of potential erroneous data being involved.