# Analysis of Universal Studios Park Reviews

## Introduction

As a consumer, customer reviews are used to help make purchase decisions. The customer ratings of a product are often used to judge a product. When there is a limited amount of reviews it is possible to read the full text of the reviewer and make an informed decision. Once the reviews are of larger numbers it is harder to make informed decisions based on the review text.

For the provider, this can lead to missing on a lot of valuable information regarding what the customers like or dislike regarding their products. One issue with relying on the customers ranking is that there is no uniformity in the way customers rate their experience.

rating written\_date title review\_text branch

3087 3.0 February 13, 2019 Harry Potter area is A+, Mummy is A, all else ... Harry Potter area is A+, Mummy is A, all else ... Universal Studios Florida

10866 2.0 March 13, 2017 Harry Potter area is A+, Mummy is A, all else ... Harry Potter area is A+, Mummy is A, all else ... Universal Studios Florida

47151 4.0 May 9, 2014 A must... You must go to Sentosa Island and see Universa... Universal Studios Singapore

47197 5.0 April 29, 2014 Can't miss it! You must go to Sentosa Island and see Universa... Universal Studios Singapore

While the cases above may be extreme, they illustrate two instances of the exact same review text but with different ratings. This analysis will attempt to find a model that can predict a customer rating with decent accuracy that can then be used to score ratings. This will provide a more uniform rating system.

Another way to benefit would be to determine what aspects of the product are liked or not liked. A basic method of finding what customers are rating positive or negative is to find words and topics that are frequent in positive or negative reviews.

## The Data

The data used for this analysis, after removing duplicate rows and dropping the reviewers name deemed unnecessary for this project, consists of 50,845 rows. Each row contains the rating (1-5), written date, title review text, and the branch. This analysis does not focus on the title although it may be useful in the future. Similarly, the date was not used for any analysis. Analyzing changes in reviews over time can provide useful information. However, no such attempt was made. It would have been interesting to have the date of park attendance as well, but it is not provided.

The table below shows the distribution of the rating scores for each of the three parks.

count mean std min 25% 50% 75% max

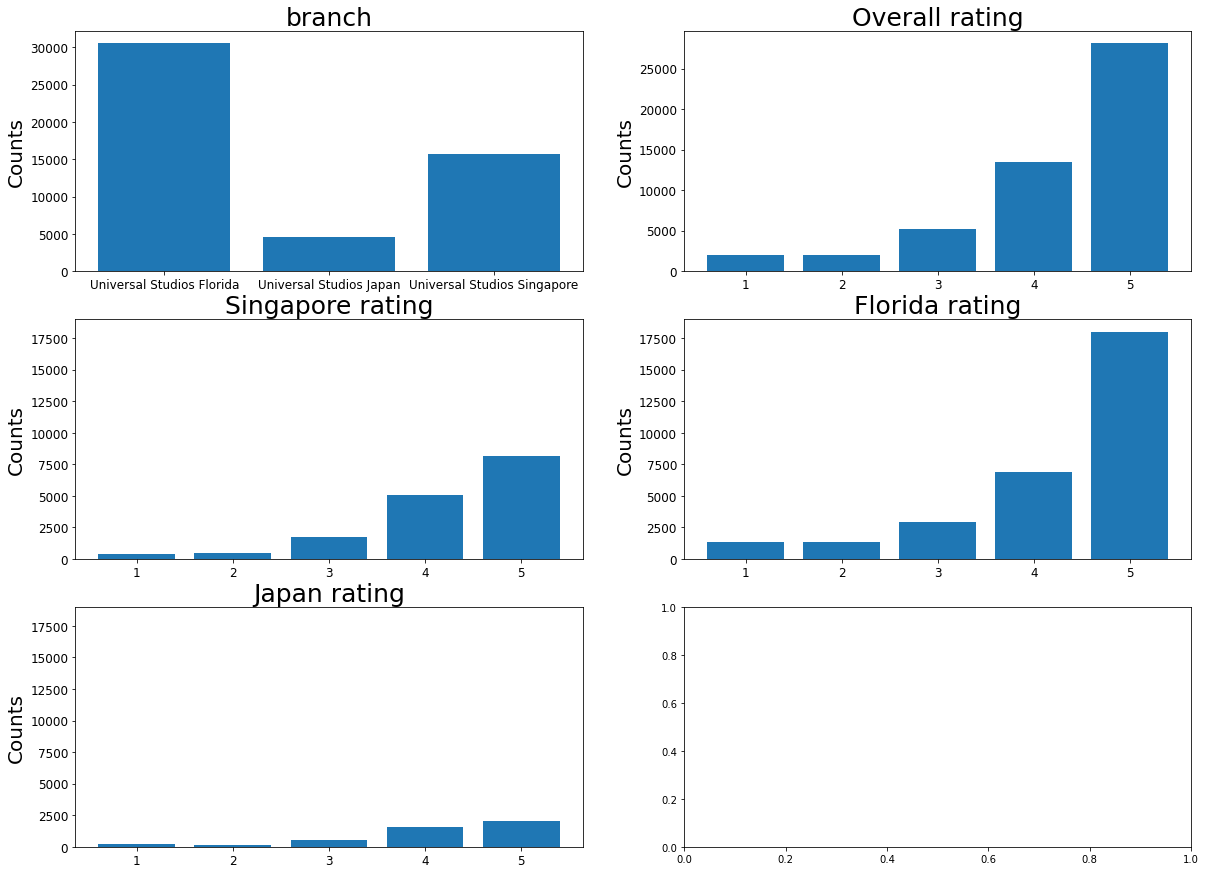
branch

Universal Studios Florida 30581.0 4.266963 1.092853 1.0 4.0 5.0 5.0 5.0

Universal Studios Japan 4527.0 4.130992 1.049306 1.0 4.0 4.0 5.0 5.0

Universal Stu Singapore 15737.0 4.274639 0.946289 1.0 4.0 5.0 5.0 5.0

As can be seen most of the ratings are positive. Both Florida and Singapore branches have more than 50% of their ratings at 5. All three have 75% of ratings above 4. This distribution can be seen in the graphs below.



## Sentiment Analysis

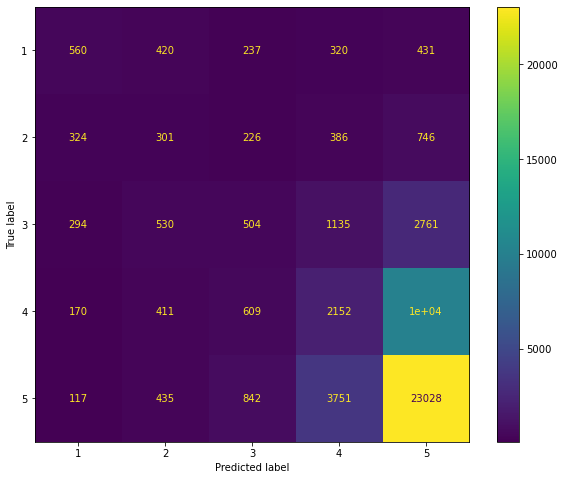
Using a general sentiment analysis from the nltk library predictions can be made for all reviews. The results are,

* Accuracy: 0.5220769003835185
* Precision: [0.38225256 0.14353839 0.20843672 0.27789256 0.62034967]
* Recall: [0.28455285 0.15179022 0.09647779 0.15944284 0.81737834]
* F1 score: [0.32624527 0.14754902 0.13190264 0.20262699 0.70536343]

The poor performance of this model may be contributed to two factors.

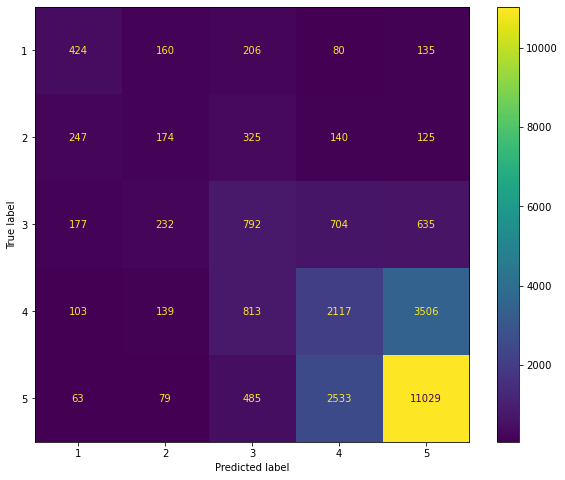
1. As detailed above there is little uniformity in the rating of consumers. It would be hard to predict the exact rating given from the sentiment analysis score. One person may be very negative regarding the point of their review but only deduct one or two points.
2. There are comments in the reviews that are industry specific. A sentence such as “The wait was over an hour” may receive a neutral review from a sentiment analyzer but would lead to an actual negative rating.

An exact view of the sentiment analysis performance can be seen in the confusion matrix below.

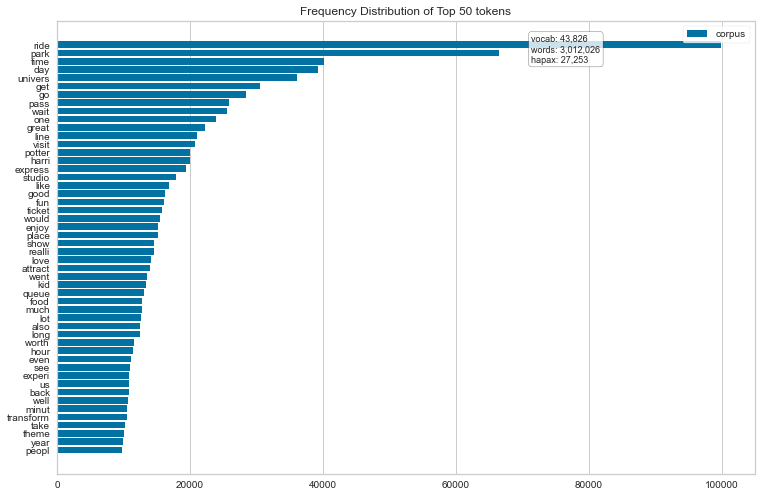


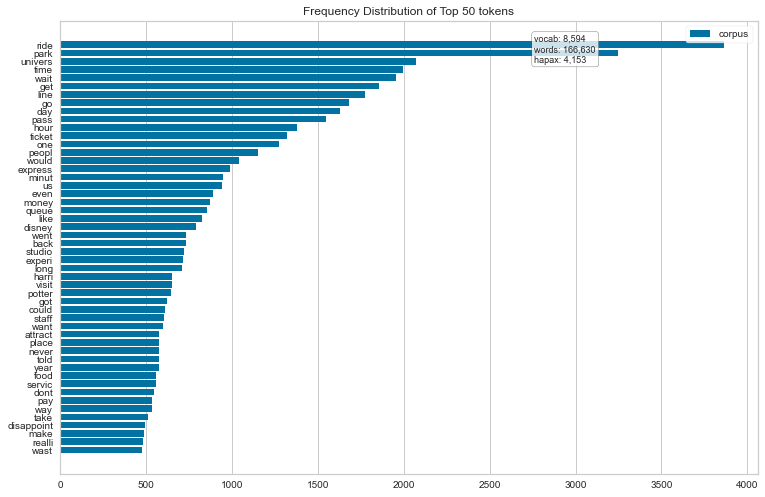
## Machine Learning

A better approach would be to use the data to build a model to determine the rating for each review. An attempt was made using the Keras Classifier model. A three-layer neural network was used to predict one of five rating classes. The model was trained on half the data set equal to more than 25,000 cases and tested on the other half. It would be interesting to retest the model using ‘leave one out cross validation’ to determine the power of the model using the full data set in predicting all cases. This was not attempted due to lack of computer memory and computation power. The results for the test performed are,   
Accuracy: 0.5717657239507532  
Precision: [0.41814596 0.22193878 0.30217474 0.37979907 0.71477641]  
Recall: [0.42189055 0.17210682 0.31181102 0.31701108 0.77729227]  
F1 score: [0.42000991 0.19387187 0.30691726 0.34557623 0.74472467]

This model performs noticeably better on all levels than the sentiment analyzer above. One reason may be as explained above that the model is trained on an industry specific data set. This accounts for instances that can affect ratings that may not be picked up by a generic sentiment analyzer. It would still be hard to determine the exact rating given by each individual. It would be hard to create a measure of accuracy for the model that is not measured against the provided ratings. Manually rating a sufficiently large set of reviews would be time consuming and may not be cost effective. A more practical approach would be to improve the model to an acceptable range and then manually review outliers to see the effectiveness of the model. A confusion matrix of the results is included below. 

# Words of Interest

There are many ways to divide words to determine interest. For this project the most basic extractions were performed. Lists of words for different subsets of the data were created based on frequency and tfidf scores. The distribution of the full data set’s most frequent words looks like this.

This can be compared to the same for the most negative rated reviews. 

One noticeable difference is where the word hour appears in the two graphs. Negative reviews contain the word hour at a much higher percentage. This indicates along with other words that time spent waiting is something that reviewer’s rate negatively.

Such differences can be seen in the lists below. Each one is of a different rating.

['regist', 'cashier', 'food', 'cash', 'minimum', 'ruin', 'horribl', 'review', 'whole', 'work', 'close', 'servic', 'place', 'experi', 'minut', 'one', 'get', 'wait', 'time', 'vacat', 'uncar', 'disrespect', 'rude', 'manag', 'camp', 'bubba', 'especi', 'restor', 'disrispectful', 'super', 'staff', 'long', 'there', 'go', 'vomit', 'like', 'smell', 'outdat', 'ride', 'ever', 'worst', 'less', 'great', 'capac', 'take', 'id', 'wont', 'guest', 'care', 'clearli']

The above list is stemmed words for a ratting of 1. Comparing this to the list for a ratting of 5 as below

'bourn', 'stuntaculari', 'xs', 'jason', 'enough', 'stress', 'knew', 'wrong', 'worth', 'anyon', 'ok', 'noth', 'pack', 'wasnt', 'entir', 'cost', 'go', 'watch', 'absolut', 'cant', 'think', 'never', 'time', 'everyon', 'trip', 'movi', 'miss', 'orlando', 'sure', 'awesom', 'buy', 'everi', 'make', 'dont', 'take', 'see', 'would', 'went', 'amaz', 'wait', 'studio', 'pass', 'univers', 'expect', 'long', 'bit', 'enjoy', 'crowd', 'clean', 'impress']

The negative list contains words such as, “ cashier, food, service, wait, time, disrespectful,’ This seems to indicate negative feelings in regard to customer service (cashiers, disrespectful) and wait time. In the positive list the first two words refer to a ride which seems to be well liked.

## Conclusion

Although more work is required in perfecting the model used to rate the reviews, there is a benefit in creating such a model. It may not be required to divide the ratings into five levels. This may improve accuracy a little. A working model should be obtainable. Having a uniform rating system can help provide a basis for analysis of feedback.

Using the ratings to extract features of interest provides an overview of what customers find positive or negative in the product. A further step may be to use the part of speech tags to extract the words for each group. Knowing what topics are present in the rating classes can help focus resources in the correct places.