Final Project Report

Christian Kevin Kusuma – U1302568

Introduction

This project will create a professional and valuable analysis from customers' reviews on specific products/activities of Polynesian Cultural Center, a cultural theme park in Hawaii. While primarily designed to fulfill a required course work from MKTG 6640, this project's secondary role is to be used by the management of the Polynesian Cultural Center as a credible reference for their pursuit of service excellence. Through its CMO/vice president, the company has agreed to grant access to a subset of its customer satisfaction survey data for this project.

Polynesian Cultural Center is a non-profit organization that dedicates its efforts in preserving Polynesian cultures, including Hawaii, Tahiti, Samoa, Tonga, Aotearoa, and Fiji, while giving opportunities to underrepresented students from Asia-Pacific regions to pursue higher education at nearby Brigham Young University – Hawaii. The analysis resulted from this project will go beyond assisting the company to better its services. It will also support the preservation of Polynesian heritage, educate future generations, and provide opportunities to thousands of less-fortunate students.

This attraction company offers authentic Polynesian cultural experiences (which include Hawaiian culture) to visitors from every corner of the world, with one million visitors annually. They offer a wide range of activities and services to customers such as bus pick up, shows, cultural activities, retail stores, restaurants, concessions, and tours. To enhance customer experience and maintain product quality, the company collects satisfaction survey data that include activity reviews from their customers. The survey dataset that contains customer reviews

will be the main ingredient for this text analytics project to create a comprehensive analysis that will answer these business questions: 1) What are the most frequent meaningful words used by customers that associated with positive and negative sentiments? 2) What deeper insights can be drawn from those frequently used words to enhance customer experience? 3) Which customers will most likely recommend the theme park to others?

Targets/Plans

This project will use several text analytics techniques and methods that have been discussed in class, such as tokenization and stop-word removal, to pre-process the data before the actual analytics process.

The analysis will contain single and bi-gram word-cloud visualizations to illustrate the most frequent meaningful words used by the customers in their reviews. The single word-cloud graphs will be grouped by their sentiments, whether positive or negative, for each of the different activities.

The most common words found on the word cloud visualizations will be analyzed carefully using word embeddings techniques. A word-embedding model will be trained to gain deeper insights from those frequently used words to provide recommendations for the management to enhance customer experience.

Lastly, the project will use a random forest algorithm to create a classification model from the text data to predict which customers will most likely recommend the theme park to others.

Method & Discussion

Accessing the Customer Survey Data

Polynesian Cultural Center has a quite comprehensive satisfaction survey that includes comment/suggestion from their recent customers. The CMO/vice president of the company agreed to give access to a subset of their satisfaction survey for the purpose of this exercise. The survey data is stored in their DOMO instance where SQL was used to subset the data as shown below:

```
SELECT `responseid`
, `d.service`
, year(`d.service`) `Year`
, `islands.com`
, `guide.com`
, `pageant.com`
, `dinner.com`
, `ha.com`
, concat(`islands.com`, `ha.com`, `guide.com`, `pageant.com`, `dinner.com`) `All Comment`
, `recommend`
, `r.islands`
, `r.guide`
, `r.ha`
, `o.sat`
from `guest_exp_comments_kr_com`
where year(`d.service`) in (2018, 2019)
```

Data Dictionary:

- responseid: unique id assigned to each customer
- d.service: the date when they visit the park
- Year: the year extracted from the d.service
- islands.com: customer reviews of the islands
- guide.com: customer reviews of the tour guide
- pageant.com: customer reviews of the pageant
- dinner.com: customer reviews of the dinner
- ha.com: customer reviews of a show called HA
- All Comment: the combination of all comments
- recommend: the likelihood of a customer to recommend
- r.islands: the ratings given to the islands experience by customers
- r.guide: the ratings given to the tour guide experience by customers
- r.ha: the ratings given to the HA show experience by customers
- o.sat: the ratings given to the overall satisfaction experience

Creating Subsets

The dataset is then broken down into six different subsets, where each contains the data of five different activities/services and one combination of all activities/services.

1. data_islands
2. data_guide
3. data_ha
4. data_pageant
5. data_dinner
6. data_all

Sentiment Analysis

Creating a corpus for each of the six subsets to feed to the liwcalike() function with the NRC dictionary. Positive and Negative sentiments are the only ones that will be used from the outputs of the liwcalike() operation. A new binary variable will be created to mark whether an observation has an overall positive or negative sentiment. The new binary variable will be called "Sentiment" and will be calculated by looking at both positive and negative variables. If the positive variable is higher than negative, then the observation is "positive," otherwise "negative." These Sentiment variables will be added to the six datasets respective to their activity type.

Word Cloud

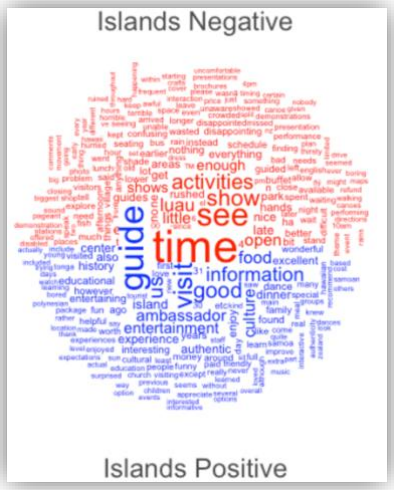
Corpuses are once again created from the 6 datasets to create bases for word-cloud visualizations. DFM's then will be the results of conversions of those six corpuses.
Using textplot_wordcloud() function, word-cloud visualizations are created from each of the six datasets joined by their respective sentiments.

All Comment Word Cloud

This word cloud shows that “food”, “good”, “guide”, “visit”, and “tour” are the top frequent words used associated with positive feeling/sentiment. On the other hand, “buffet”, “seating”, “shade”, “pig”, and “shade” dominate the negative frequent words.

Islands Word Cloud

This word cloud shows that “guide”, “visit”, “information”, “food”, and “culture” are the top frequent words used associated with positive feeling/sentiment. On the other hand, “time”, “luau”, “show”, “activities”, and “see” dominate the negative frequent words.

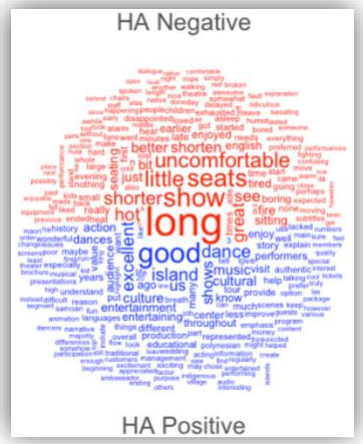


Guide Word Cloud

This word cloud shows that “guide”, “tour”, “good”, “friendly”, and “information” are the top frequent words used associated with positive feeling/sentiment. This finding suggests that guides are most likely to be perceived positive when they’re friendly and informative. On the other hand, “time”, “group”, “smaller”, “spent”, and “people” dominate the negative frequent words.

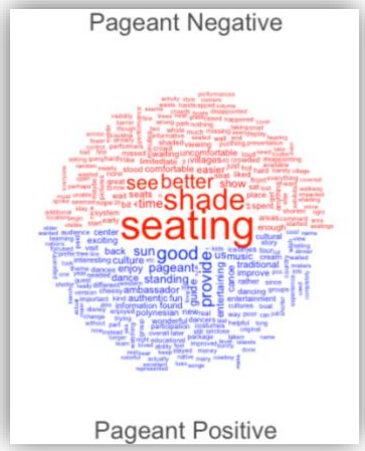
HA Word Cloud

This word cloud shows that “good”, “dance”, “island”, “music”, and “excellent” are the top frequent words used associated with positive feeling/sentiment. On the other hand, “long”, “show”, “seats”, “uncomfortable”, and “tired” dominate the negative frequent words.



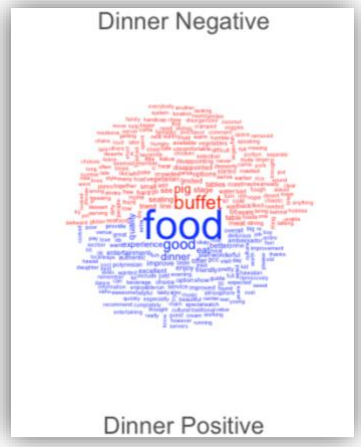
Pageant Word Cloud

This word cloud shows that “good”, “cultural”, “sun”, “entertainment”, and “music” are the top frequent words used associated with positive feeling/sentiment. On the other hand, “seating”, “shade”, “see”, “better”, and “time” dominate the negative frequent words.



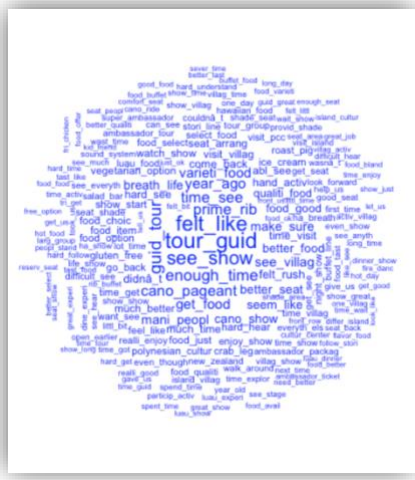
Dinner Word Cloud

This word cloud shows that “food”, “good”, “dinner”, “eat”, and “quality” are the top frequent words used associated with positive feeling/sentiment. On the other hand, “buffet”, “pig”, “stage”, “tables”, and “limited” dominate the negative frequent words.



Bi-gram Word Cloud

This word cloud shows that “tour_guide”, “felt_like”, “see_show”, “enough_time”, and “prime_rib” are the most frequent 2-word combination.



Word Embedding

The text file to feed a word-embedding model is created from the combination of all comments from the main dataset. The text file then gets some cleaning including lowercase and 2-grams before fed into the model. The word-embedding model then is trained with these parameters:

- vectors = 200
- threads = 4
- window = 12
- iter = 5
- negative_sample = 0

```
# Training the Word-embedding Model
model = train_word2vec("pcc_review_file_ready.txt",
                       "pcc_review_file_ready_bin.bin",
                       vectors=200,
                       threads=4,
                       window=12,
                       iter=5,
                       negative_samples=0,
                       force = T)
```

The model provides a way to analyze relations between or among words. That means that that there is a nice tool to dig deeper into the most frequent words as shown in the word-cloud visualizations.

Digging Deeper Insights from the Most Frequent Words in Islands

In the Islands word-cloud graph shown previously, we can uncover hidden relationships that will enable us to understand better why a customer has a negative feeling toward this particular activity. By running closest_to() function from word2vector package on the words "time" and "luau", we can see that the closest words that are associated with those two words are "7", "rush", "felt_rushed", "4_00", and "6pm".

word <chr>	similarity to "time" + "luau" <dbl>
time	0.7889103
luau	0.7161741
7	0.5066757
rush	0.4987104
felt_rushed	0.4858454
4_00	0.4758342
6pm	0.4747911
i've	0.4741338
allocated	0.4565416
30pm	0.4543393

This finding is very interesting because right after the islands' activities, each guest is hoarded to one of the luaus – one at 4pm and the other at 6pm. Those who are hoarded to the 4pm luau may feel very rushed because it's quite an early dinner and they may still want to experience the islands' activities. Most guests prefer to have the 6pm's dinner instead of the 4pm's.

Digging Deeper Insights from the Most Frequent Words in Guide

By running closest_to() function from word2vector package on the words "group" and "smaller", we can see that the closest words that are associated with those two words are "size", "large", "microphone", "scooters", "speaker", "training", "voice", and "awkward".

word <chr>	similarity to "group" + "smaller" <dbl>
group	0.8178513
smaller	0.7750491
size	0.7211902
large	0.7013886
microphone	0.5807166
scooters	0.5729576
speaker	0.5609889
training	0.5603624
awkward	0.5591356
voice	0.5540000

Each tour guide normally would take less than 20 people for each tour. However, guides would take up 50 people in one tour during high seasons such as summer and Christmas. The people in

big groups may have a hard time to hear the guide's voice, thus making them have negative sentiment toward the guide or the experience.

Digging Deeper Insights from the Most Frequent Words in HA Show

By running closest_to() function from word2vector package on the words "long", "shorten, and "shorter", we can see that the closest words that are associated with those three words are "boring", "repetitive", "relaxing", "asleep", "dragged", "chaotic", and "cheesy".

word <chr>	similarity to "long" + "shorten" + "shorter" <dbl>
shorter	0.8370071
long	0.8140662
boring	0.7376388
shorten	0.6721612
repetitive	0.6605124
relaxing	0.6455185
asleep	0.6433961
dragged	0.6427236
chaotic	0.6294857
cheesy	0.5828835

After spending the whole day participating in the islands' activities and eating their dinner one of the big buffets, each guest will take a seat at a huge theater for the main show from 7pm – 9pm. This 2-hour show may be too long for some people who have spent their entire day in the park.

Digging Deeper Insights from the Most Frequent Words in Pageant

By running closest_to() function from word2vector package on the words "seating", "shade", and "better", we can see that the closest words that are associated with those three words are "importantly", "viewing", "comfortable", "tiered", "shaded_areas", and "provide".

word <chr>	similarity to "seating" + "shade" + "better" <dbl>
shade	0.8267619
seating	0.8025479
shaded	0.7288840
importantly	0.7217531
viewing	0.6620526
comfortable	0.6409060
better	0.6384172
tiered	0.6314958
shaded_areas	0.6085985
provide	0.6061790

During the day, the park has a canoe pageant that features dances of the different nations of Polynesia. Since the performances are carried out on canoes, guests have to sit or stand along the lagoon (artificial river). Many guests have to be under the sun for the duration of the performance to watch it and may feel the need to add shaded areas around the lagoon.

Digging Deeper Insights from the Most Frequent Words in Dinner

By running closest_to() function from word2vector package on the word “buffet”, we can see that the closest words that are associated with that word are “mediocre”, “prime”, “fried_rice”, “item”, “entree”, “kalua_pork”, and “serving”.

word <chr>	similarity to "buffet" <dbl>
buffet	1.0000000
mediocre	0.6904486
prime	0.5771133
fried_rice	0.5594368
buffett	0.5569260
items	0.5457231
entree	0.5393057
kalua_pork	0.5359800
cooked	0.5333794
serving	0.5311106

From those words that are associated with “buffet”, only “mediocre” that seems to have an interesting insight. Aside from the Luaus, the buffets are designed to be a service for those who opt out from Luau which is usually significantly more expensive than the buffets.

Creating Prediction on Which Customers Will Most Likely Recommend

Polynesian Cultural Center is interested to know which guests/customers are most likely to recommend visiting the park to their friends and families so that the company can reach out to them for marketing purposes.

We will create a prediction using a machine learning algorithm, such as a random forest. The predictors are taken from customer comments/ reviews combined with several other variables taken from the survey to enrich the model.

The random forest model is trained using train() function from the caret package and "ranger" as the method instead of the classic "rf". Random forest from "ranger" is more efficient and faster than "rf" and typically generates similar results.

To further cut the run time of the model training, mere "cv" will be used instead of "repeatedcv". The MacBook that is used to run the model for this project automatically utilizes the full power of its multiple cores without the need to set up parallel computing.

After tokenizing, creating dummy variables, and combining necessary predictors into a single Data Frame, training and testing datasets will be created with the proportion of 70% - 30%. The training dataset will be used to train the model, and the testing dataset will be used to measure the model's performance.

Random Forest Model

```
Random Forest

2290 samples
  77 predictor
   2 classes: 'No', 'Yes'

Pre-processing: centered (77), scaled (77)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2060, 2062, 2061, 2062, 2061, 2060, ...
Resampling results:

  Accuracy   Kappa
  0.8144332  0.6228358

Tuning parameter 'mtry' was held constant at a value of 12
Tuning parameter 'splitrule' was held constant at a value
 of extratrees
Tuning parameter 'min.node.size' was held constant at a value of 10
```

After trying various parameter values for mtry, splitrule, and min.node.size, it was determined that mtry=12, splitrule=extratrees, and min.node.size=10 are the best parameters to train a model with the highest accuracy. In this case, the highest accuracy achieved was 81%.

Model's Performance on Testing Data

The 30% testing data is used here to generate this confusion matrix on the right. The model's prediction accuracy on the testing data is very close to the training data and this means that the model is not overfitting and generalize well to unseen data.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  329 118
       Yes  83 450

               Accuracy : 0.7949
                 95% CI : (0.7682, 0.8198)
    No Information Rate : 0.5796
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.584

 Mcnemar's Test P-Value : 0.01648

            Sensitivity : 0.7985
            Specificity : 0.7923
         Pos Pred Value : 0.7360
         Neg Pred Value : 0.8443
             Prevalence : 0.4204
         Detection Rate : 0.3357
   Detection Prevalence : 0.4561
      Balanced Accuracy : 0.7954

       'Positive' Class : No
```

Conclusion and Recommendation

Based on the findings from sentiment analysis, word-cloud visualizations and word-embedding model above, we can identify specific problems related to specific activity/service that make customers have negative sentiment toward that particular activity/service. We can then recommend several operational adjustments to the management of Polynesian Cultural Center to enhance its customer experience, ergo improving customer sentiment from negative to positive.

Islands

Problem: Guests feel like they're rushed through the islands' activities because they have to catch the 4pm luau dinner. This may create a feeling of missing out of some activities since the islands' activities still open during their 4pm luau dinner.

Recommendation: Move the 4pm dinner to later time that the guests will have the chance to experience all the islands' activities before going to the dinner location.

Guide

Problem: Guests feel like it is hard to hear what the guide is saying because the group is too large.

Recommendation: Hire more tour guides during the high seasons so that the tour-group sizes can be kept low.

HA Show

Problem: Guests feel like the HA Show is too long especially after spending a whole day at the theme park. Many of them also need to drive 1-2 hours back to their hotels, so shortened show would help reduce their fatigue.

Recommendation: Compress the overall story of the HA Show to shorten the watch time. This may help improve guests' engagement as well.

Pageant

Problem: Guests feel uncomfortable with the amount of shades during the hot-day canoe pageant.

Recommendation: Invest more on better seating arrangements and shades.

Dinner

Problem: Guests feel that the buffet experience is mediocre.

Recommendation: Better food selection and possibly add fried rice into the menu.

Polynesian Cultural Center can now effectively run their targeted marketing to reach out to those customers who are most likely to recommend visiting the park to their friends. The random forest model trained in this project has around 80% of accuracy against training and testing datasets. This means that the company can lower their marketing cost and potentially reach better audience.

Additional Information

This final project report will not be presented to the management of the Polynesian Cultural Center because this report is specifically designed to satisfy MKTG-6640 final project requirements. A separate report will be created and presented to the management in a more persuasive and less technical approach.

Works Cited

"Oahu, Hawaii's Top Attraction." *Polynesian Cultural Center*, www.polynesia.com/.

"Home." *Hawaii Tourism Authority*, www.hawaiitourismauthority.org/.

"pcc_reviews.Csv." *Google Drive*, Google, drive.google.com/file/d/15--fYWkngxh1ibnsg1BzTRWcVa0wFL8Q/view.

"Machine Learning with Caret." *DataCamp*, learn.datacamp.com/courses/machine-learning-with-caret-in-r.