Using Instagram Posts to Compare Disney Parks –An Application of Sentiment Analysis

Sridevi Suresh

Abstract

Social media applications such as Instagram are a great resource for data in order to conduct social experiments or glean information about the opinion of the majority. In this project, we investigate the popularity of two Disney parks; namely, Disneyland and Disneyworld. We use public Instagram posts to learn about statistics in the data. As well, we use the posts to perform sentiment analysis to deduce which park is more popular for the majority of people.

1 Introduction

The goal of this project is to quantify which park is better: Disneyworld or Disneyland. Our hypothesis is that Disneyworld is better than Disneyland. We approach this task by using a type of regression to perform sentiment analysis. We look at the caption under a post made by a user on Instagram about either Disneyland or Disneyworld and give it a score from $0-4$ based off of how *positive* the post is. The higher the score, the more positive a statement is. We say that the more positive a post is, the more fun a user is having at that park,. We then say that the park where people have more fun is the better one. In addition to sentiment analysis, we will be looking at the activity within each park as well as the number of likes posts received within the respective parks. These are three pieces of evidence we will use to distinguish which park is the better one.

2 Data Collection

We retrieved a client id and secret id by requesting access to the Instagram API. Once we received this, we installed a library titled *python-instagram* in order to make the process of getting our queries from the api easier. Once we set the ids properly, we followed the correct calls from the documentation of the library to retrieve the information we wanted. We retrieved information based off of the hashtag used on Instagram.

For Disneyworld, we retrieved posts that included at least one of the following hashtags: "wdw", "disneyworld", "downtowndisney", "hollywoodstudios", "magickingdom", "animalkingdom", or "epcot." Similarly, for Disneyland we retrieved posts with the hashtags: "disneyland", "fantasyland", "frontierland", "crittercountry", "mainstreet", "neworleansquare", "tomorrowland", "toontown", or "adventureland." We chose to ignore posts that were written in a language other than English and collected the responses in a csv file.

The columns of the csv file refer to different features of the data: the first column is the user-name, the second column is the number of followers of the user, the third column is the image URL, the fourth column is an integer representing the number of likes of the picture, the fifth column is a string represent-

ing the caption for the post, the sixth column is an integer representing the number of comments of that post, and the rest of the row is filled with strings which are the individual comments. This data was collected over a period of 8 days total spanning from 4/5 to 4/12. We retrieved 1100 posts total for Disneyland and 1100 posts total for Disneyworld.

disneyne	420	https://sc	55	We're exc	0	
candy2171	116	https://sc	16	#disneysi	0	
theresaca	84	https://sc	10	#hauntedi	0	
viewofash	43	https://sc	7	Looking fc	0	
viewofash	43	https://sc	3	So great t	1	@alformentera
disneydre	901	https://sc	12	Because b	0	
cartoondi	785	https://sc	26	I can't par	1	amazing!

Figure 1: Sample Data from Disneyland Data Set

3 Preprocessing

As part of pre-processing, we decided to move all of the comments for each post into a different csv file so that we can simply read our csv file as a data frame through the python *pandas* library. While we were collecting data, we had kept track of how many entires we had by the end of the day for the two parks for every day of data collection. We further took the data frames and split them up into mini dataframes by day and by park. This made analyzing certain trends much easier.

4 Initial Data Analysis

We gathered all the data and decided to look into some initial trends and statistics before we embarked upon doing any sentiment analysis. We did this in order to get an idea of what our data looked like and to see if there might be an inkling of whether or not our hypothesis is correct.

The first observation we made was on the average number of followers that users who posted about the respect parks had. As can be seen by Figure 2, users who post about Disneyworld on average have more followers than users who post about Disneyland. At a high level, this tells us that people who are more popular post about Disneyworld.

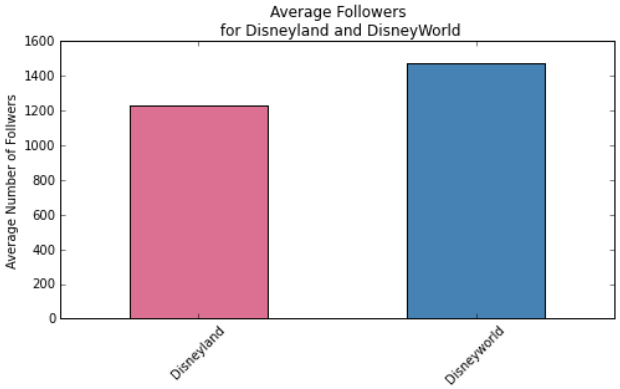

Figure 2: Average Followers

We investigated further to look at the activity on Instagram over the span of our data collection to see which park had more activity. By activity, we mean that we counted how many unique users made at least one post about either park. Looking at Figure 3, we see that in five out of the eight days Disneyworld had more activity than Instagram. This is further evidence that Disneyworld is more popular than Disneyland.

To look specifically at the indiviual posts, we plotted the average number of likes and the average number of comments received on posts over the course of our data collection. One important thing to note is that we did not do any sort of normalization, rather took the given mean value. Figures 4 and 5 show us that Disneyland had more likes and comments on posts than Disneyworld did. In both cases, six of the eight days Disneyland had a higher average than Disneyworld.

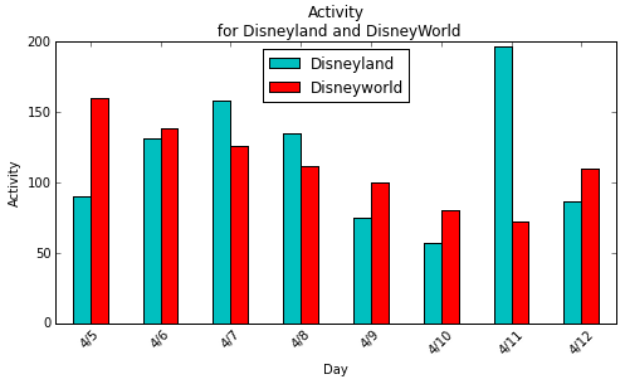

Figure 3: Activity

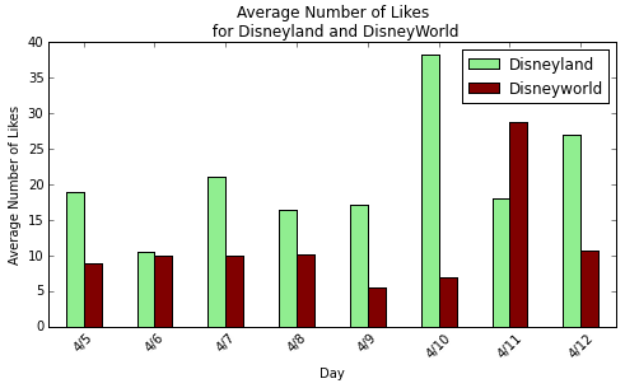

Figure 4: Average Likes

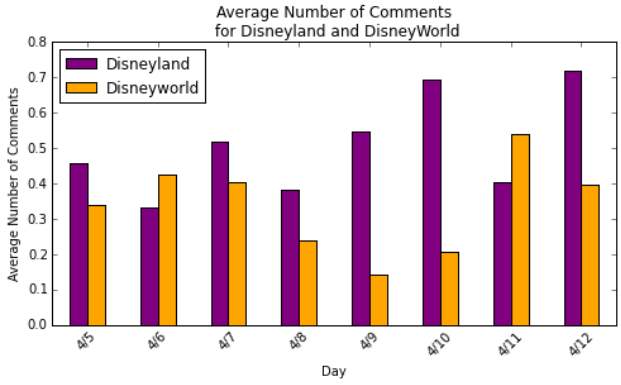

Figure 5: Average Comments

5 Further Work

The evidence that we have collected so far does not make us feel confident in our hypothesis. We have seen that while Disneyworld is more popular in terms of posts, Disneyland is more popular in terms of interaction of users for posts. However, as noted above we did not perform any normalization of the data as this was just an initial screening to see if our data could tell us anything interesting to begin with. We cannot make any strong statements about whether or not we expect the sentiment analysis to prove our hypothesis correct because of the inconclusive trends we have seen initially.

We plan on computing the averages likes and comments once more, however, this time we plan on normalizing by the number of followers of the user so this might give us a more accurate result. We also plan on performing the sentiment analysis to see what it can tell us about the happiness of people at the different parks. We continue to hope that our hypothesis will be proven true.