National Park Service Sentiment Analysis
Shelby Heise

Problem Statement

How do people feel about the National Park Service?

To explore this question, I have set up a sentiment analysis on select parks in the National Park Service using Twitter data. This information can be helpful in determining how people feel about the parks overall, including the features available at the parks. Understanding visitor feedback is essential in making informed decisions regarding resource management and park services.

Twitter is a great option for exploring visitor feedback, as visitors frequently use the social media platform to share their experiences. Further, this information is available for public collection using Twitter's API.

As the park system continues to experience increasing visitor numbers, diving into a sentiment analysis can be beneficial when gauging public opinion. This information could be especially helpful as the National Park Service adjusts and adapts to new pressures in a COVID-19 world.

Data Collection

Data for this project was gathered using Twitter's API paired with Twython—a Python wrapper that works with Twitter's API. I selected Twitter's standard search API which—at the time of this project—allows developers to query a collection of tweets over the previous seven days. Additional limitations of the search API included a cap of 500,000 tweets per month, with a rate limit of 300 tweets every 15-minutes. The standard search API features a "query" parameter that I used to search for park-related hashtags. I further filtered the search results by setting the language as "English."

Twitter data was collected across a period of three weeks in 2020; the final week of November, and two weeks of December. This time period included two holidays and three weekends, which typically see higher visitor turnout. The result of each collection was saved as a separate CSV.

To prepare for the project, I reviewed the Twitter pages for some of the most visited National Parks in 2019. I theorized that these these parks would likely have higher visitor engagement on social media platforms such as Twitter, and provide a larger pool of available tweets for analysis. For each of these pages, I checked for hashtags used by the official park account and completed a quick search to see what other tweets might be pulled with those hashtags. For example, "#Zion" was used frequently across the official Twitter page for Zion National Park. However, "#Zion" also pulled a large number of unrelated tweets. For Zion National Park, I ultimately selected another frequently used but more specific hashtag- "#ZionNationalPark". Some parks did not have a clearly favored hashtag; for consistency, I used the format "#ParkNameNationalPark". A quick search showed that this combination appeared to pull related and recent tweets for the selected parks.

My initial data collection focused on five park-specific hashtags and one hashtag utilized by the official National Park Service Twitter page: #gsmnp, #GrandCanyon, #rmnp, #ZionNPS, #Yosemite, #NPS. This initial data collection yielded a relatively small number of tweets, with many unrelated to the park system. I modified my search, and expanded to ten hashtags: #gsmnp, #GrandCanyon, #rmnp, #ZionNationalPark, #Yosemite, #AcadiaNationalPark, #GrandTeton, #OlympicNationalPark, #GlacierNationalPark, #NPS. This combination produced a higher number of relevant tweets for analysis.

Queried National Parks & 2019 Visitation Numbers

National Park	2019 Visitation Numbers	Queried Hashtag
Great Smoky Mountains National Park	12.5 million	#gsmnp
Grand Canyon National Park	5.97 million	#GrandCanyon
Rocky Mountain National Park	4.7 million	#rmnp
Zion National Park	4.5 million	#ZionNationalPark
Yosemite National Park	4.4 million	#Yosemite
Acadia National Park	3.4 million	#AcadiaNationalPark
Grand Teton National Park	3.4 million	#GrandTetonNationalPark
Olympic National Park	3.2 million	#OlympicNationalPark
Glacier National Park	3 million	#GlacierNationalPark
National Park Service	327 million	#nps

NPS/Neal Herbert. *National Park Visitation Tops 327 Million in 2019*. https://www.nps.gov/orgs/1207/2019-visitation-numbers.html

As of April 2021, it appears that Twitter has updated their standard search API, and some of these specific details may no longer be applicable.

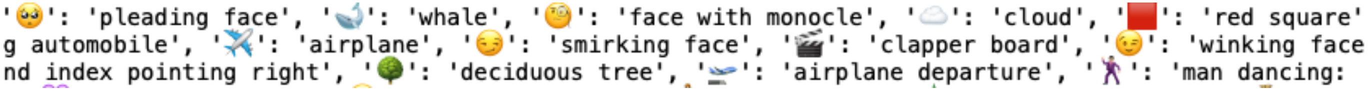
Cleaning

Multiple cleaning and processing steps were required before moving forward with analysis. For this stage of the project, I heavily relied on the Natural Language Toolkit (nltk).

First, any duplicated tweets and retweets were removed from the dataframe. Tweets were then separated by dropping all columns in the dataframe except for the “text” column. While the other columns contained interesting information, the “text” column housed the tweet bodies. For consistency while processing, the text column was also converted to lowercase letters. This was saved as “tweet.”

Emojis presented an interesting challenge, as removing them in the cleaning process would potentially eliminate the sentiment behind the emoji. To preserve the meaning behind the emojis, demoji was utilized to find and replace emojis with their text description.

Example of demoji text description



```
'🥺': 'pleading face', '🐋': 'whale', '🧐': 'face with monocle', '☁️': 'cloud', '🟥': 'red square'
g automobile', '✈️': 'airplane', '😏': 'smirking face', '🎬': 'clapper board', '😉': 'winking face
nd index pointing right', '🌳': 'deciduous tree', '🛫': 'airplane departure', '🕺': 'man dancing:
```

All remaining non-alphanumeric characters and hyperlinks were removed from the tweets using regular expression. This process produced a cleaned version of our original tweets that was ready for processing.

NLTK Processing

Following the initial cleaning, I completed a series of data processing steps using nltk. The first step was tokenization, which is the process of splitting each tweet into individual words. TweetTokenizer was applied on the cleaned tweets to generate tokens.

To allow for better processing, stop words needed to be removed from the tokenized words. Stop words are commonly appearing words such as “in” and “is” that do not provide additional feedback during analysis. For this project, stop words were removed from the tokenized tweets using nltk’s stopwords.

After stop words had been removed from the tokenized tweet, the final step was to apply nltk’s lemmatizer. Lemmatization acts as part of the word normalization process and involves reducing “inflected words” to their root word. For example, converting the word “hours” to “hour.”

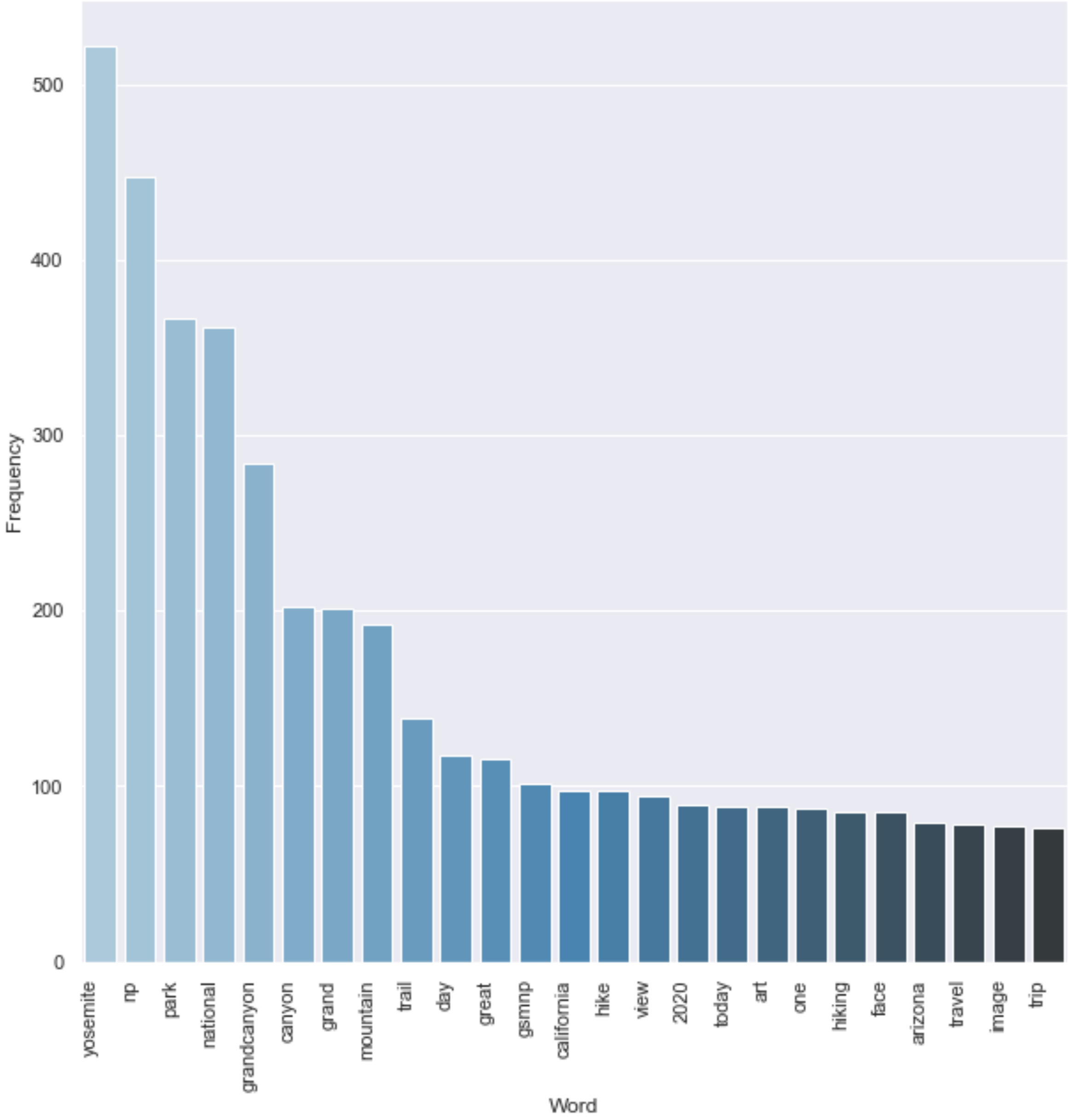
The tokenized and lemmatized tweets were saved as “tweet_final”.

From here, I reviewed the frequency of words appearing in the Twitter data using nltk’s FreqDist. Both the 100 most common words and 100 least common words were generated to explore what words are most commonly used by visitors to the parks. This information was used to create two visualizations of the most common words: a bar plot and word cloud.

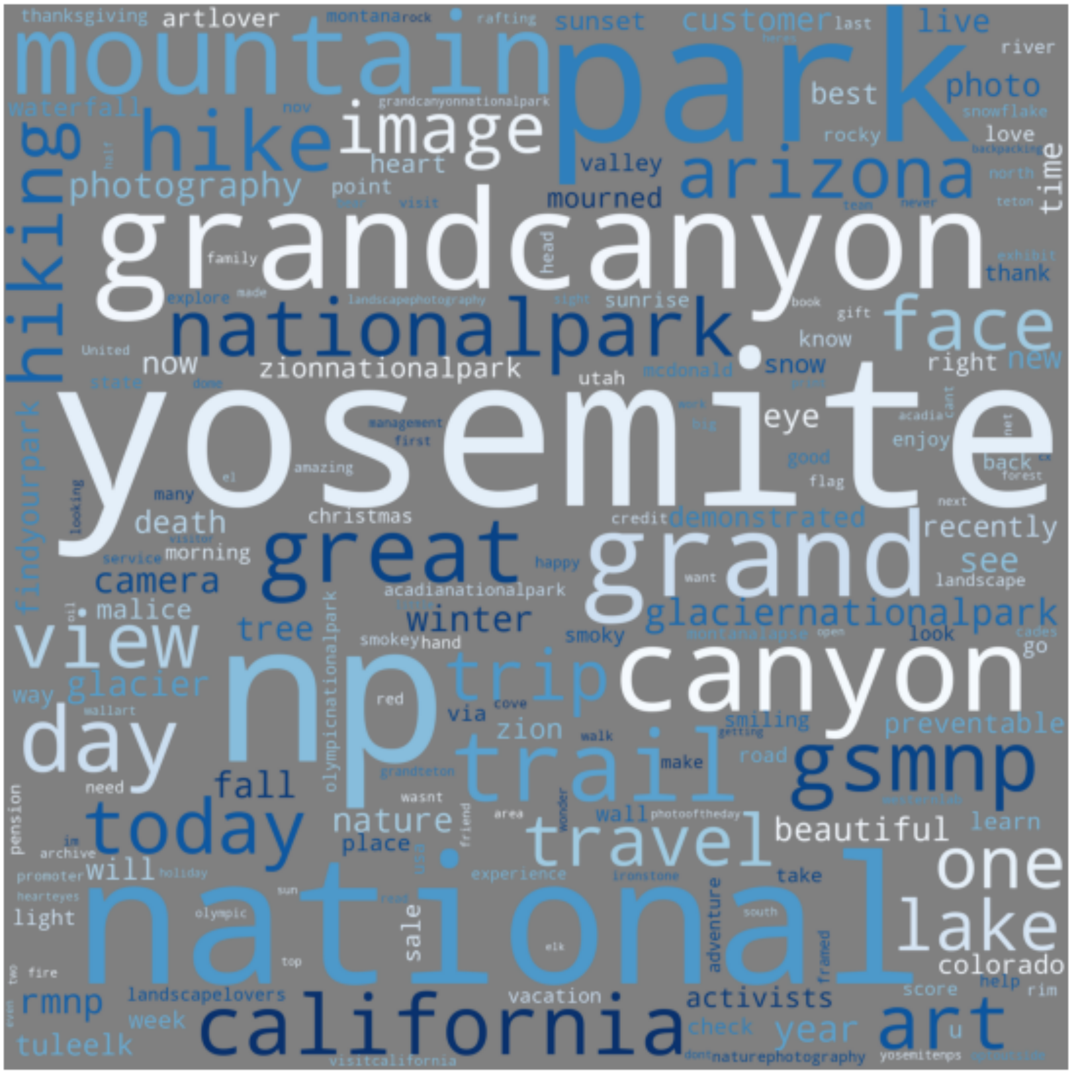
The final step in this project was to utilize nltk’s SentimentIntensityAnalyzer on the data. In order to assign a sentiment value, the SentimentIntensityAnalyzer requires context. For this reason, it would not produce accurate results on the tweets that had undergone additional nltk processing such as tokenization and removal of stop words.

Analysis

Twenty-Five Most Common Words



Word Cloud



The results of the sentiment analyzer revealed that overall, the collected Twitter data showed a neutral-to-positive response to the National Park Service.

Sentiment Analysis Scores of Collected Twitter Data

Negative: 3.2%	Neutral: 76.4%	Positive: 20.4%

This is somewhat expected, as a number of the most frequently appearing words in the dataset were interpreted as “neutral.” For example, some of the most commonly appearing words like “mountain” and “park” are processed as neutral and have no negative or positive connotation. Further, the cleaned Twitter data also contained stop words that would be assigned a neutral value. If you examine *just* the negative score and the positive score of the sentiment analysis, the overall general opinion of the parks is a positive one.

Tools

There are a number of Python libraries and packages that can perform sentiment analysis and also work well with Twitter's API. For this project, I opted to utilize both Twython and nltk over other packages for a few basic reasons. Most importantly, was ease of use. Nltk is a fully-fledged toolkit with easy-to-understand documentation, as well as a plethora of resources for new users. The existing nltk userbase was a great resource for natural language processing questions. Of the available Twitter API wrappers, Twython was also very easy to use and beginner friendly.

Takeaways & Future Research

There were many limitations that I faced throughout this project that may impact the overall effectiveness of its findings. First and foremost, was the amount of data collected. After modifying my initial search to include a total of ten hashtags, I was able to collect a larger number of tweets. However, for a more well-rounded analysis, it would be helpful to have a greater number of available tweets. This could be accomplished by adding additional search criteria, and expanding the number of parks included in the query.

Ideally, it would also be beneficial to collect this information year-round. Several of the parks experience dips in visitor turnout as the seasons change, and continual data collection may help offset these fluctuations. For this project, data was collected in the winter months when some of the most popular parks exhibit some of the lowest turnout numbers. This can potentially lead to a smaller number of available tweets.

For future research, it may also be beneficial for the park system to deploy a large-scale survey of park visitors and make that information publicly available. In conjunction with a sentiment analysis of Twitter data, this information could be helpful in completing a more comprehensive gauge of public opinion.

Despite the limitations of this project, overall, the National Park Service is viewed positively by the public. This information can be used to help shape future park resource management decisions, especially as we continue to experience the effects of COVID-19.