Bharath Murthy

Ryan Yoneshige

IST 652 Final Project

Tokyo 2021 Olympics Tweets

Data and Source

The primary data are composed of various tweets surrounding the Tokyo 2021 olympics. The original data were obtained from as a .csv file from Kaggle.com and consist of over 290,000 rows across 13 attributes, including the id of the user, the date/time in which the tweet was created, the language in which the tweet was written, the location of the user, number of retweets, and the actual tweet text, itself.

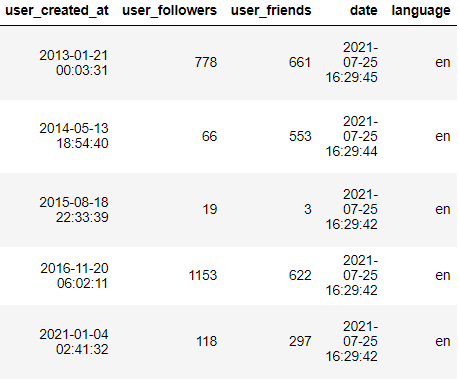
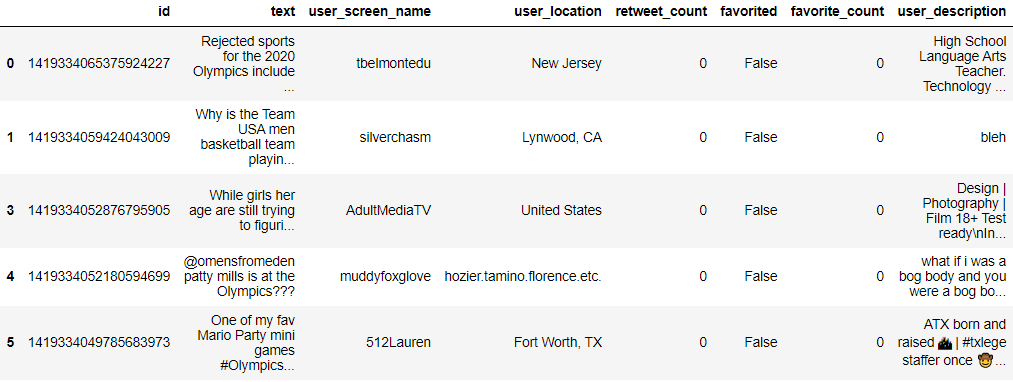
A data set from olympics.com was also included to look at the distribution of medals awarded to each country, as well as two other data sets from whereig.com and olympiandatabase.com to look at the countries who participated in this year’s olympic events, the number of athletes from each country, and the NOC (National Olympics Committee) abbreviations for each country. A Wikipedia page was also used to bring in the names of the Olympic sports involved in the Tokyo 2021 Olympic games.

Description of Data Exploration and Data Cleaning

As many of the rows in the original data set contained “NaN” values, after reading the data into a data frame, almost 90,000 rows were removed, leaving 202,768 rows and 13 columns. The names and description of the columns are as follows.

| id | numerical Twitter id of the user |
| --- | --- |
| text | text of the tweet |
| user\_screen\_name | user’s Twitter screen name |
| user\_location | location of the user |
| retweet\_count | number of times the tweet was retweeted |
| favorited | whether or not the tweet was favorited |
| favorite\_count | number of times the tweet was favorited |
| user\_description | description of user |
| user\_created\_at | date and time user joined Twitter |
| user\_followers | number of Twitter followers of the user |
| user\_friends | number of Twitter friends of the user |
| date | date and time the tweet was made |
| language | language in which the tweet was delivered |

The remaining data were further analyzed, and cleansing resumed. The columns of “retweet\_count,” “favorite\_count,” “user\_followers,” and “user\_friends” were found to contain strings, so they were all converted to integer type data. The columns “user\_created\_at” and “date” were found to also be strings, so were converted to a date/time format through the datetime library.



The second data set, which contained the medals data, was read into a dataframe and contained 93 rows and 8 columns. As a few of the columns (namely the columns involving medal colors) were named “Unnamed: #,” the names of those columns were changed to be more descriptive. The names and descriptions of the columns are as follows.

| Rank | Rank of the team as a result of the Olympic games |
| --- | --- |
| Team/NOC | Team/Country |
| gold | Number of gold medals obtained by the team |
| silver | Number of silver medals obtained by the team |
| bronze | Number of bronze medals obtained by the team |
| Total | Total number of meals obtained by the team |
| RankbyTotal | Rank of the team by the total number of medals |
| NOCCode | National Olympics Committee country code |



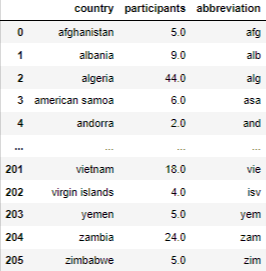
Another data set used in this analysis was a data set containing a list of the countries that participated in the Tokyo 2021 Olympic games as well as the number of athletes from each country that competed in the games. As the original countries data set split the names of the countries into three columns, the Pandas library was used to concatenate the columns containing country names as well as the columns containing athlete counts. The country names were all converted to lowercase, and a row containing all “NaN” values was dropped. The resulting data frame consisted of 207 rows and 2 columns, namely “country” and “participants.”



The final data set brought into this analysis contained a list of the countries that participate in Olympic games along with their NOC code. As the abbreviations data set had a similar structure to the countries data set, similar steps were used to concatenate the columns representing the same information. The country names and their abbreviations were converted to lowercase, and the resulting data frame consisted of 238 rows and 2 columns.



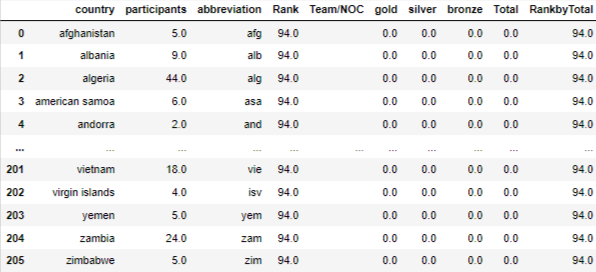
The abbreviations data frame was joined with the countries data frame on the country name and a new data frame was created consisting of country name, number of participants, and the country’s NOC code. As not all of the country names from the countries data frame perfectly matched with the country names from the abbreviations data frame, 13 rows of the new “countries\_abbr” data frame contained “NaN” values in the “abbreviation” column. After manually entering in the 13 abbreviations, the resulting data “countries\_abbr” data frame consisted of 206 rows and 3 columns, namely “country,” “participants,” and “abbreviation.”



In order to create a data frame that held the medals counts of all the participating countries, the “medals” data frame’s “NOCCode” column and “Team/NOC” column were converted to lowercase, and the “medals” data frame was joined with the “countries\_abbr” data frame on the “NOCCode” and “abbreviation” columns, respectively.



As many of the countries that participated in these Olympic games did not receive any medals, the resulting “countries\_perf” data frame had many “NaN” values that needed to be resolved. The “NaN” values in the “Rank” and “RankbyTotal” columns were replaced with the value of 94, as there were 93 countries that received medals, and the remaining countries all tied for last place. The “NaN” values in the “Team/NOC” column were replaced with an empty string, as many of the entries in that column were identical to the “country” column from the “countries\_abbr” data frame. The “NaN” values in all remaining columns (pertaining to medal counts) were replaced with the value of 0, as those countries did not receive any medals.



With all data read in and cleansed, the Analysis Questions could now be addressed.

Analysis Questions

The analysis of the Olympics tweet data was broken into four main questions.

1. Which Olympic sports were referenced the most in this tweet data?
2. To what extent did a relationship exist between user follower count and tweet content during these Olympic events?
3. How was the content in targeted tweets (those with an “@”) different from that of the non targeted tweets (those without an “@”)?
4. Which countries were referenced the most in this tweet data?

The first question aims to provide the Olympic Committee insight into the nature of viewership in these summer Olympic games. This may help the committee schedule popular events during times/days when many people who watch the Olympics will be able to view them.

The second question aims to provide Twitter users with insight into what popular Twitter users tweet about in order to possibly gain more followers.

The third question aims to provide the Olympic Committee and/or companies insight into what is popular enough at the Olympic games for a user to desire to share this experience with another user. This may help companies target their advertising to more efficiently spend their money.

The fourth question aims to provide the Olympic Committee insight into which countries garner the most attention across participating nations. This could aid in the Olympic Committee’s decision as to which countries could potentially host future Olympic games.

Description of Program

The program begins by importing all of the necessary libraries and modules. Next, the primary data set and medals data set are read in by use of the Pandas library, and both data frames are cleansed, following the processes described in the previous section.

In order to begin the analysis of the data, and answer the first Analysis Question, a list of the Tokyo 2021 Olympic sports was required in order to cross-reference against the tweet text. To accomplish this, the Summer Olympic Games Wikipedia page was scraped using the Beautiful Soup library and parsed for the names of all summer Olympic sports. A list called “sports” was then created. In order to cross-reference the tweet text with the sport names, the tweet text required preparation.

First, the tweet text was stored into a list and the URLs were removed from the tweet text with the help of Regular Expressions. Next, the tweet text was converted into all lowercase and tokenized, and stop words were removed using the nltk stop words. Finally, the list of lists of words were flattened and counted using the itertools and collections libraries.

The “sports” list was then tokenized and flattened, and intersected with the flattened tweet text words to create the “int\_words” list. Dictionaries for the tweet text words and the sports words were then created in which the “keys” were the words, and the “values” were the counts of each word. The “sport\_word\_counts” dictionary was then sorted by the frequency of the word, and the resulting dictionary was called the “sorted\_sports\_dict.” A plot was then created through matplotlib to show the distribution of sports words by their frequencies, and question one was ready to be answered.

To answer the second Analysis Question regarding user follower counts, the tweets were binned by the number of followers a user had. A “followers\_bin” column was added to the primary data frame in which “bin 1” referred to users that had less than one million followers, “bin 2” referred to users that had between one million and ten million followers, and “bin 3” referred to users that had more than ten million followers.

A “tweet\_length” column was then added to the primary data frame to keep track of the length of each tweet, as that may aid in the analysis. The primary data frame was then summarized and grouped by the “followers\_bin,” and the means of all numeric columns were shown. The primary data frame was then split into three data frames based on the “followers\_bin” in order to more easily analyze the tweet data within each followers bin group. The same process of tokenizing, removing stop words, flattening, and counting that was used on all of the tweets was repeated within each followers bin data frame, and three similar plots were created. The second Analysis Question was now ready to be answered.

In order to address the third Analysis Question on targeted tweets, the primary data frame was split into two data frames based on whether or not the tweet text included an “@” symbol. The same process of tokenizing, removing stop words, flattening, and counting was performed on both data frames, but plots were not created, as the third Analysis Question could easily be answered simply by looking at the most common words in each resulting dictionary of word counts.

In order to address the fourth Analysis Question on the countries’ references in the tweet data, the two other data sets involving the country names, participants, and abbreviations were read in through the Pandas library, and both were cleansed and joined with the “medals” data frame as described in the previous section to create the “countries\_perf” data frame. Next, a list of the words in the tweet text that referenced either the name of the country or its abbreviation was created and word frequencies were counted. A dictionary of the most common 30 “country\_words” was created along with a plot to show the same information, visually.

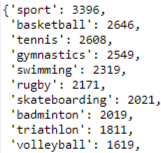
For further analysis into the words referencing specific countries, a column that calculates the proportion of the total medals won by a specific country was added to the “countries\_perf” data frame as “medals\_prop,” and the “countries\_perf” data frame was sorted by “Rank.”

Next, a “country\_words” dictionary was created that held all of the words that referenced countries in the tweet text. This dictionary was then used to create another dictionary that calculated the proportion of times each specific country word was used out of the total times any country word was used. The resulting “country\_prop” dictionary was then sorted, and the fourth Analysis Question was ready to be answered.

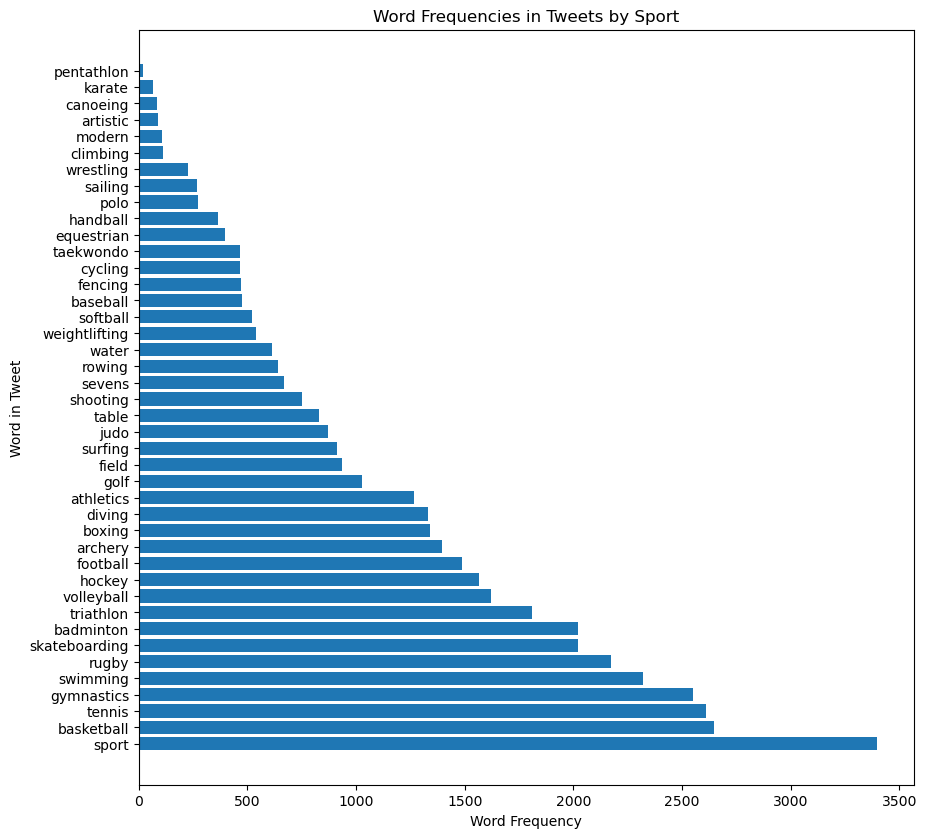
Description of Output Files

As all cleansing of data read into the program was described and shown in the previous sections, those images will not be included here.

In answering the first Analysis question, the sports words were intersected with all of the tweet words, and the top ten words by frequency in the “sorted\_sports” dictionary were as follows.



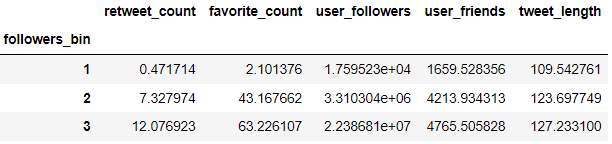
The same information is displayed in graphical form below.



In the dictionary as well as the plot, it is clear to see that the sports that were referenced the most in tweets this year were basketball, tennis, gymnastics, swimming, and rugby.

Although the dictionary and counts are accurate, the limitations of performing the analysis in this manner is that among Twitter users, many people may not tweet the exact name of the sport/event they are referencing. For example, although the “track and field” type events are extremely popular, they do not show up on this list, as all of the “track and field” events fall under the sport “athletics.” However, this method of analysis works very well for sports such as basketball and tennis, which most users would reference by their official sport name.

Continuing on to the results for Analysis Question two, once the tweets were binned by follower count and summarized by calculating the mean of the numeric columns, the following output was generated.

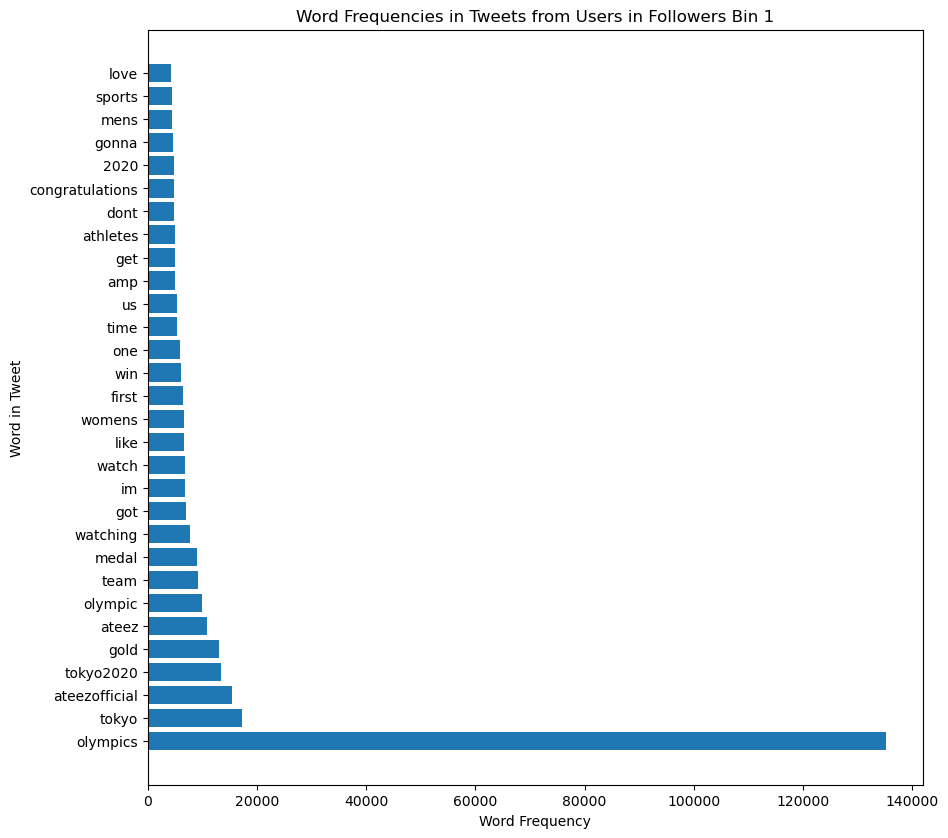


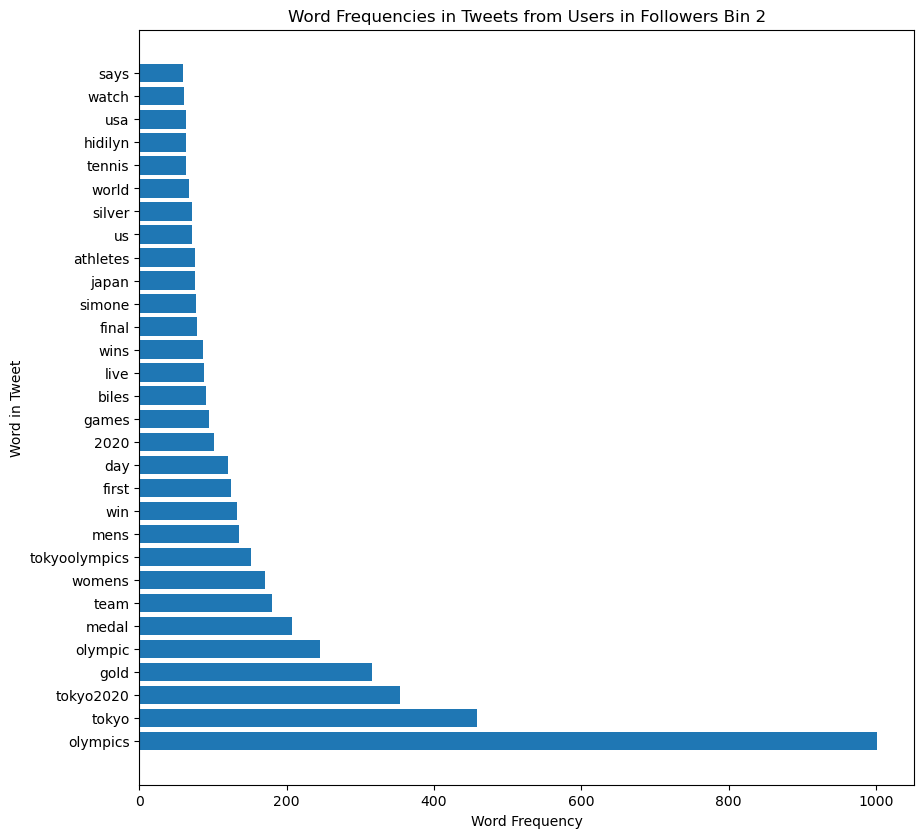
From this summary table, it is clear to see that although most of the differences across the followers bins are a result of having more or less followers, the difference in average tweet length is not. It appears that users with more followers tend to tweet longer tweets than those with less followers, on average.

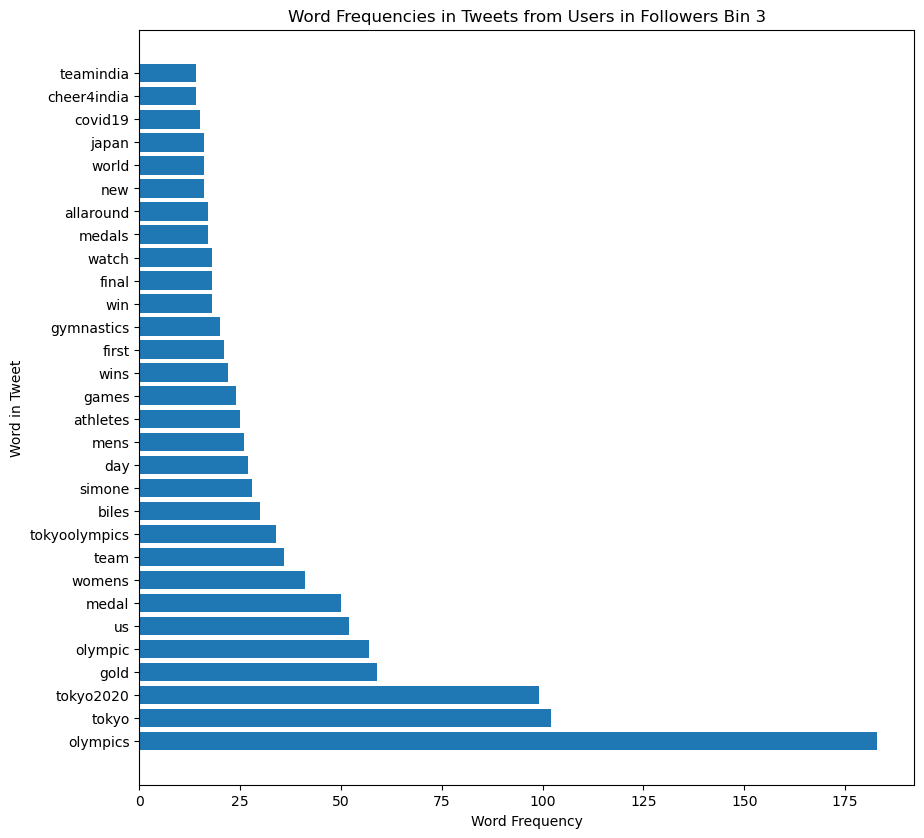
The top ten words by word frequency in each follower bin were as follows.

| Followers Bin 1 | Followers Bin 2 | Followers Bin 3 |
| --- | --- | --- |
|  |  |  |

Accompanying plots of the top 30 words in each follower bin show slightly more information.







It is clear from the top ten words in each bin as well as the plots of the top 30 words in each bin that the users, regardless of their number of followers, used similar words in their tweets. However, there are small differences in the tweets from Bin 1 when compared to the tweets from the other two bins. Specifically, Simone Biles, the USA gymnast who is arguably the “greatest of all time”, but withdrew from most of the events in which she was meant to compete, does not seem to appear in the top 30 words mentioned in tweets from the users in Bin 1. However, Biles is definitely mentioned quite a bit from users in Bins 2 and 3. Instead, users in Bin 1 tended to talk more about the Korean pop group Ateez, whose songs happened to play at the Tokyo 2021 Olympic games.

In response to Analysis Question three, the primary data frame was split on the inclusion or exclusion of an “@” symbol in the tweet text. After counting the word frequencies in each group and finding the top ten words by word frequency, the following outputs were generated.

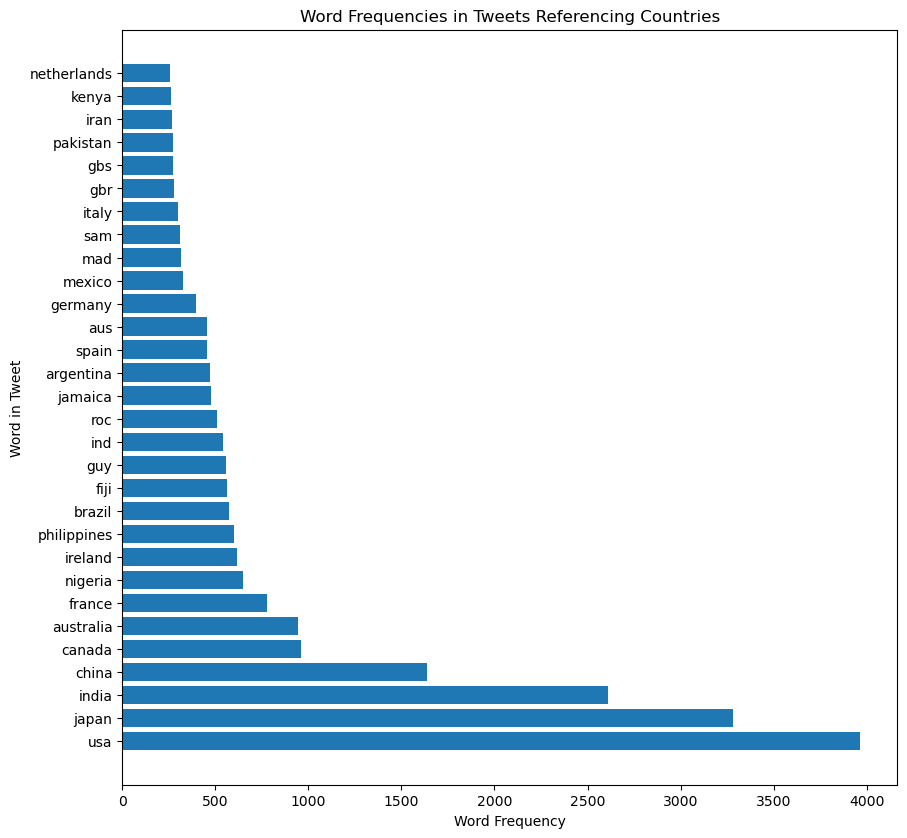
| With “@” | Without “@” |
| --- | --- |
|  |  |

From this output, it is clear that the tweets including an “@” were mainly targeted at the Ateez group, and those without the “@” followed a similar breakdown to the word frequencies based on followers bin.

Continuing on to Analysis Question four, after the “countries\_perf” data frame was created and the country\_words list was created, the top ten country words by word frequency were as follows.



Once again, displaying the same information visually, yields the following plot.



Clearly, from the top ten country words, as well as the accompanying plot, the countries referenced the most often were USA, Japan, India, China, and Canada. It is understandable that the United States of America is at the top of this list, as the language of all the tweets in the sample was English. It is also understandable that the next most referenced country is Japan, as the Olympic games were held in Tokyo, Japan this year. However, India being the third most referenced country was a very interesting outcome, as they were nowhere near the top in medal count.

One limitation of doing the analysis in this manner is, as mentioned before, only looking at tweets that were delivered in English may skew the results to be very heavy on the side of the USA. Also, the fact that Japan was the host of the Olympic games this year may confound the usual tweet popularity of the country in Olympic games. Another limitation of performing the analysis in this manner is the inability to account for every way a country may be referenced in tweets, as many people may tweet partial names of countries and/or different abbreviations than the NOC abbreviations. Lastly, the counts from the NOC abbreviations were not combined with the counts from the country names, so the results may be slightly different if there is a somewhat even split between users tweeting the actual name of the country and its corresponding NOC abbreviation. With all of those limitations taken into consideration, the results are still very interesting, and somewhat match the success of the countries in the Olympics games this year.

To further analyze the relationship between tweets referencing countries, a proportion of total medals won by each country was added to the “countries\_perf” data frame, and once the data frame was sorted by “Rank,” the top ten countries were as follows.



Running the same proportion calculation on the country word frequencies, gives the same top ten countries as mentioned earlier, but gives the proportion of times the country was referenced when looking only at tweets where a country was referenced.



What is interesting about these two collections of top ten countries, is that five out of the ten occur in both collections. This shows that when countries win medals, it is likely that they will be referenced in a tweet. However, India stands out significantly in this data, as they are the third most frequently referenced country despite not winning very many medals. As shown below, India won 0.0065, or 0.65%, of the total medals, but were referenced in 0.079, or 7.9% of tweets that referenced a country.



Although no sentiment analysis was performed on the data, these results most likely speak to India’s pride as a country, and support of their athletes regardless of their medal performance in the Olympic games.

Summary of Results

As shown in the analysis and answers to the four Analysis Questions, many conclusions can be drawn from different aspects of tweet data. What was found in this study was that the top five most referenced sports were Basketball, Tennis, Gymnastics, Rugby, and Skateboarding. With that in mind, the organizers of the Olympic games may wish to schedule those sports during times in which the largest number of people can watch them.

What was also found in this study is that the most common tweeted words, other than “olympic(s),” “tokyo,” and “tokyo2020” across all follower bins were “gold,” “medal,” “team,” and “womens.” This shows the competitive nature of the games, but also the aspect of a team. The fact that “womens” is one of the most highly tweeted words in this data set across all follower groups could mean that women’s sports events in this year were much more popular than the men’s events.

Another finding was that the targeted tweets this year, generally centered around a Korean pop group, may show the acceptance of the Asian culture and music in the general population. In a year where #stopasianhate was a popular hashtag, the many tweets around Ateez was a nice thing to see.

Lastly, as mentioned earlier, the most referenced countries were the USA, Japan, India, China, and Canada. Possibly the most interesting finding of this study is India’s great country pride and support of their athletes. As a country that did not perform exceptionally well in the olympic games, India was still able to make the top three in regards to referenced countries, possibly due to the fact that India won their first gold medal in three Olympic games (13 years).

Group Member Tasks and Roles

Group members collaborated on gathering the data from multiple sources, as well as in the data cleansing and analysis. Both members of the group were responsible for creating the final project presentation slide deck as well as the final project report.