# logo_bw.eps *Worcester Polytechnic Institute*

# *Data Science Program*

# Case Study 1 Twitter Analysis

# Submitted By

# Oluseun Olulana Enbo Tian Dang Tran

## **Date Submitted :** 2/23/2022

## **Date Completed :** 2/23/2022

## **Course Instructor : Prof.** Ngan

# Motivation and Background

# The Beijing 2022 Winter Olympic was held from February 4th to February 20th, which was one of the most hot topics around the world during the time we started this case study. The opening ceremony was also concerned by the society all over the world. The Slogan of the Winter Olympic was: Together for a Shared Future. Together was added to the Olympic Slogan in the Summer Olympic in Tokyo, with faster, higher, and stronger. It is also a hot topic during this term of the Olympic. Many people sent their ideas for the Winter Olympics on Twitter. Thus, many tweets were collected by us from different languages and different locations. It is significant to see the attributes and speech from those countries and people from different places. By the way, there are more hot topics about the Winter Olympics and that’s what we are interested in.

# Data Sources

# The source of our data was the Twitter Search API. To get research/developer access, we applied on the developer.twitter.com portal.

# Using the search API, we ran a filter to get tweets about the just concluded Olympics Tournament in Beijing, China - Beijing 2022. As expected the tweets were centered around the activities and sports competition during the tournament. For this research and analysis, we collected a total of 1100 tweets.

# We extracted the data in json format which had several name/value pairs containing information like the timestamp of the tweet, the location from which the user was tweeting from, the text, user mentions, number of retweets, profile background color, followers count, friends count, etc.

Data Characteristics: The data was unstructured data, as it was in JSON format. The data was all text and numbers. There were urls linking to web hosted images. There was some numerical data. Some examples of the discrete numerical data were the retweet\_count and favorite\_count. There was also some nominal categorical data, for example, “follow\_request\_sent”, “favorited”, “retweeted”, which all took the value true or false.

There was some missing values in certain pairs.

# Methodology

# We chose “Beijing2022” as the keyword to gather a collection of tweets using the Twitter Search API. We collected these tweets in json format and downloaded them into a local text file.

We conducted some frequency analysis with our collected tweets using the Counter, PrettyTable, and nltk libraries in Python.

* For our word count analysis, first we used the nltk library to get rid of all the common stop words in our collection of tweets. Counter was used to easily find the most frequently used words and PrettyTable put them together in a table of the top 30 words with the highest counts.
* To get the most popular tweets, we decided to determine their popularity by the retweet counts. First we put the tweets and their retweet counts together into a dictionary of tuples. Then we simply created a function to sort the dictionary by the retweet counts and put them together into a table of the top 10 most popular tweets using PrettyTable.
* To find the most popular hashtags and user mentions we employed a simpler methodology from that of our word count analysis. We used Counter to find the most frequently used hashtags and user mentions from our tweet collection then used PrettyTable to put them into tables of the top 10 most popular hashtags and user mentions.

We conducted some friends and followers analysis on a designated user in our collection.

* To choose the user, we went through the first 100 users in our data and chose the 3rd most popular user based on the sum of their follower and friend count. We chose the 3rd one because the most popular users tend to have a much higher ratio between follower and friend count which resulted in less chances of there being mutuals between the 2.
* Using twitter api, we gathered 300 user ids of the chosen user’s friends.
  + We decided on 300 due to their being a rate limit of 900 every 15 minutes.
  + 300 allows us to troubleshoot quicker and runs the data collection for both friend and follower in quick succession.
  + We then used twitter api again to collect the screen names of the users based on their ID.
  + Finally, we used panda dataFrame to put our data together and create a table of 20 of the designated user’s IDs and screen names.
* Repeated the process above for the follower counts.
* To find the mutuals between the 2 data collections, we simply used the panda merge method to find the common user ID and plot it out in a table using panda DataFrame.
  + Due to the small collection of 300 friends and followers each, occasionally we’d get no mutuals so we’d need to collect new data or choose a new user.

For our domain question, we decided on finding the locations with the most tweets.

* The process is similar to the word count analysis where we used counter and PrettyTable but on the location attribute instead.
  + There was a lot of trash data as a lot of users left their location empty and there’s no generalized way that they follow in setting their location so several resets were needed to get meaningful data.
  + Used PrettyTable to create a table of the top 3 most popular locations.

# Results

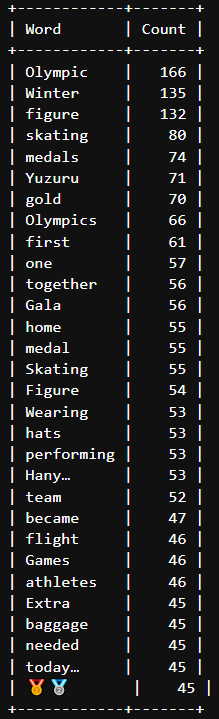
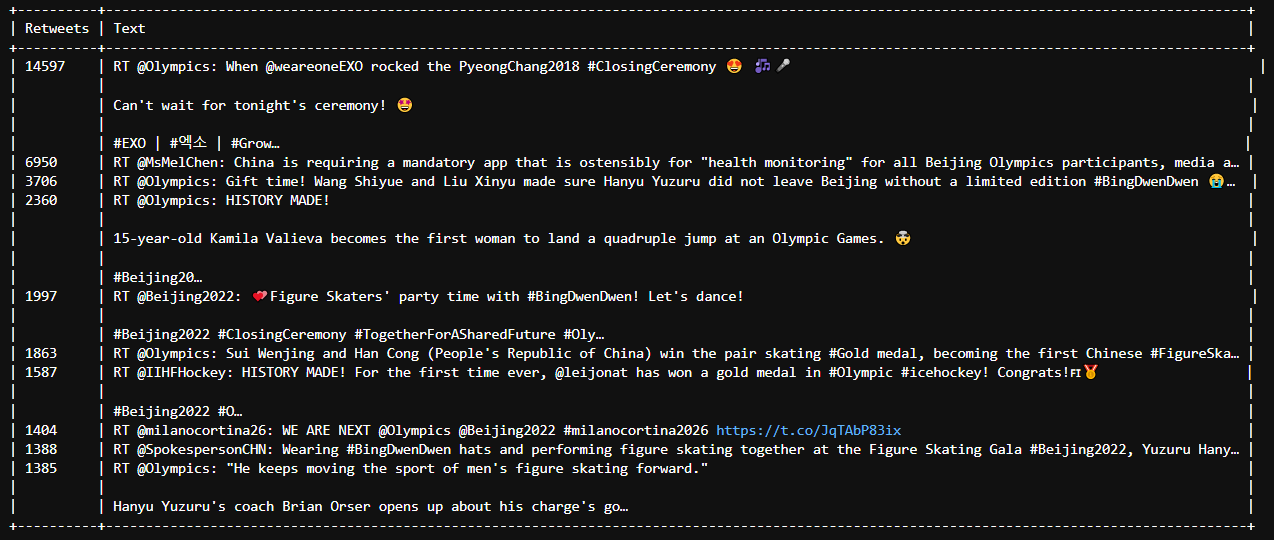


Figure 1: Top 30 Most Frequently Used Words



# Figure 2: Top 10 Most Popular Tweets by Retweets count

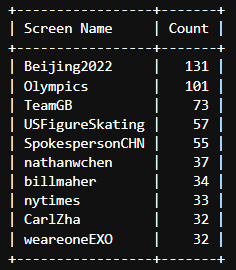


Figure 3: Top 10 User Mentions

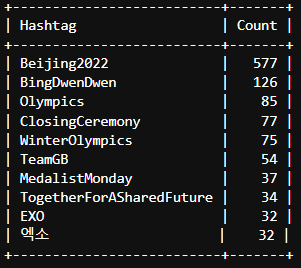


Figure 4: Top 10 Hashtags

Designated Popular User: rodmickleburgh



Figure 5: IDs and Screen names of 20 of rodmickleburgh’s followers



Figure 6: IDs and Screen names of 20 of rodmickleburgh’s friends



Figure 7: IDs and Screen names of rodmickleburgh’s mutuals between friends and followers

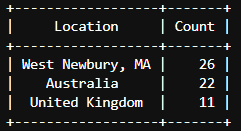


Figure 8: Top 3 most popular locations

From the results, we found that the highest tweeting location was West Newbury, Massachusetts in the United States. We also discovered the following:

Most frequently used words - Olympic, Winter, Figure and Skating

From the most popular tweet, a lot of Twitter handlers seem to be interested in the closing ceremony.

@Olympics is the user with the highest retweets

The most popular user mentions are @Beijing2022 and @Olympics

The most popular hashtags are #Beijing2022 and #BingDwenDwen

# Conclusion

From the tweets we got, and the analyzing of data, we can conclude that Skating was the most concerning sport during the Winter Olympics. In addition, people also focus on the medals, especially the gold medal. Yuzuru Hanyu is the most famous athlete in the Olympics. People were also concerned about the ceremony. Because the date of the closing ceremony was close to the date we finally searched, it is more concerning than the opening ceremony. We have the most locations of west newbury, MA, Australia, and UK, but this cannot be elucidated too much. Since the host of Winter Olympics, China banned twitter, and many countries do not speak English, The location can only give an elucidation of the English-speaking countries. And Since we are located in MA, we got more tweets from the state where we are located. **Twitter Data can give a more visual location of interest in the Winter Olympics. It will help the Olympic committee to find a good place for activities in the future.** From the User’s followers and friends we got, we can see that there are only 1 common friend and followers of the popular user. Thus, we can conclude that friends and followers are not mutual during the Beijing 2022 Winter Olympics.