Locality of Interest among Twitter Users during the 2018 U.S. Midterm Elections

Liam Scholl and Morgan Vigil-Hayes
School of Informatics, Computing, and Cyber Systems
Northern Arizona University
[las589, morgan.vigil-hayes]@nau.edu

Abstract

In recent years, Americans have utilized social media to stay informed on current events (Shearer & Matsa, 2018). Over the past few election cycles, Twitter has emerged as a critical platform that citizens used to stay abreast of political elections. In this study, we examined a set of 24,272,614 tweets that were collected during the 2018 U.S. Midterm Elections. Critically, we seek to understand the locality of interest that users maintain during an election and how this locality relates to the geographic patterns. We used a K-means clustering approach to cluster the locations mentioned in the content tweeted by every user in the data set. After examining 5,335,564 users, we found that 2,142,370 tweeted content referenced a location. On average, users tweeted about 1.3 different places and had an average locality of interest of about 116 miles. Future work will investigate any relationships between a user's profile location and the places that they tend to tweet about, as well as explore how the geographic interests of users change over time.

Research Questions

RQ1: How many different places do users typically Tweet about?

RQ2: How wide-ranging are Twitter users' locality of interest?

Methodology

First, data was labelled with locations using a natural language processing approach. Next, we identified the latitude and longitude coordinates of each mentioned location. Then, we analyzed the number of places mentioned by each user. Finally, we found a locality of interest for each user based on their location data.

Clustering

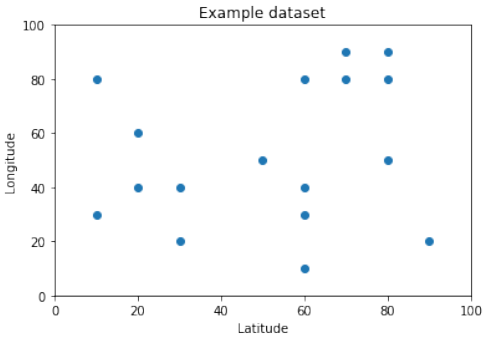
Various locations that were mentioned by individual users were collected in large data sets to be evaluated. These locations were then converted to their corresponding latitude and longitude using Mordecai, a Python package that extracts location from text using natural

language processing and gazetteers. One goal was to calculate how many clusters of locations a user mentioned, as well as the center of each cluster, or the centroid. To achieve the optimal number of clusters for each unique user, a python script was created that used K-means clustering machine learning analysis (Garbade, 2018). Given a two-dimensional set of latitude and longitude points, K-means sets out to split the data into a fixed number of clusters. One method used to find the optimal number of clusters is the elbow method, which finds the within cluster sum of squares, or WCSS score. Plotting a two-dimensional line graph with an x-axis of whole numbers and a y-axis of the WCSS scores creates an L-shaped line that has a particular point usually referred to as the elbow. The x-value at the elbow is the optimal number of clusters found Figures 1A and 1B are examples of what the elbow method looks like from a graphical standpoint, with Figure 1A being a sample data set of latitude and longitude points, and Figure 1B being a visual representation of the elbow method. According to Figure 1B, the optimal number of clusters is four. A problem that with the elbow method was the complex retrieval of the optimal number of clusters that was found because of the number of users that had to be analyzed, so it was decided to use another method known as the silhouette method. The silhouette method finds a different set of scores called silhouette scores. The silhouette scores are calculated as followed:

$$Silhouette\ Score = \frac{a - b}{\max(a, b)}$$

Where a is the mean distance to the instances of the next closest cluster and b is the mean intra cluster distance, or the mean distance to the other locations within a cluster. These silhouette scores are graphed on the y-axis, while the x-axis contains whole numbers. The silhouette method creates a graph that displays a line that has a visibly much larger area under a specific part of the line. The x-value with the largest silhouette score is the optimal number of clusters that the silhouette method found. This number is much easier to obtain as it is simply the largest number in the created set of silhouette scores. From the largest silhouette score, we can find the number of optimal clusters and the WCSS score for each user. With K-means, we were also able to find the centroid of each individual cluster as well as the amount of locations within each cluster. To find the centroid of each cluster, the fit_predict function was used to compute the cluster center and create a list of numbers relating to the size of each cluster. The number of locations, or size of each cluster was found by counting the number of referenced clusters within the list created by fit_predict (). In special cases where a user had two locations mentioned, a silhouette score was unable to be calculated, but clusters could still be determined using the distance formula. If the user's two locations are an arbitrary distance apart, this would conclude with an optimal number of clusters of two with the centroid of each cluster being each mentioned location. If the two locations are close in distance, the optimal number of clusters was one and the centroid of the cluster was calculated using the midpoint

formula between the two mentioned locations. This data was placed in a data set in JSON format and could be used in future work for cluster analysis.



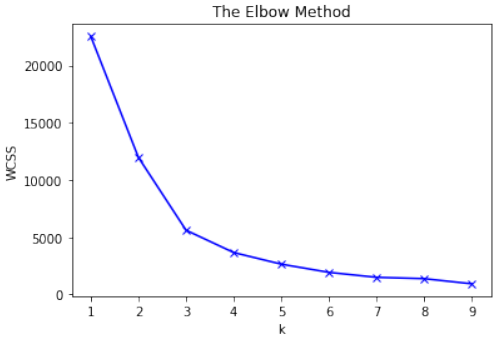


Figure 1A

Figure 1B

Radius of Gyratio

Locations mentioned by 2,142,370 Twitter users were analyzed to discover the center of each user's mentioned locations as well as the user's radius of gyration. The center of the user's mentioned locations was found by taking the average latitude and longitude points for that user. The centers of mass of each user were entered into an FCC geo census that would return the county FIPS and state FIPS of the entered data. Using the center of their locations, the radius of gyration was able to be found by taking the square root of the sum of squares of the distances between the user's center and all of their locations divided by the number of locations the user mentioned. The radius of gyration of a data set about an axis of rotation is defined as the average radial distance from the center of mass of an object to all data points within the object (Dangol, n.d.).

$$Radius\ of\ Gyration = \sqrt{\frac{\sum_{i=1}^{N} m_i (r_i - r_{CM})^2}{\sum_{i=1}^{N} m_i}}$$

The calculated radii were all in units of degrees and had to be converted to a more recognizable unit. In the United States geographical area, a typical latitude or longitude degree distance is equivalent to about 69.172 miles. With this number, we were able to convert all of the radii into units of miles.

Results

RQ1: On average, a user mentioned about 1.3 different locations (standard deviation = 1.014 locations). A majority of users had only one cluster of locations with the rest of users having their number of clusters less than or equal to 14 (Figure 2A).

RQ2: The average distance between each location cluster of each user was 174.93 miles with a standard deviation of 585.17 miles. Among the users, the average radius of gyration was about 116 miles with a standard deviation of 309 miles. Plotting this data within a cumulative distribution function revealed that 99.8% of users had a radius of gyration that was less than or equal to 2000 miles (Figure 2B).

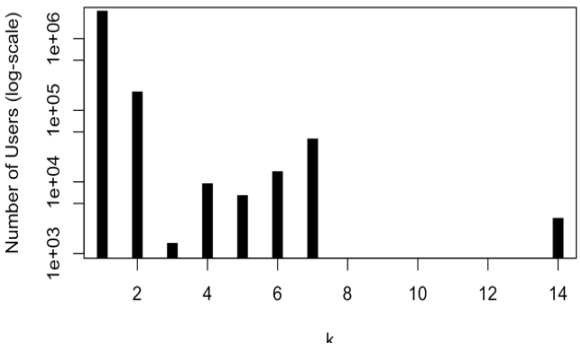


Figure 2A

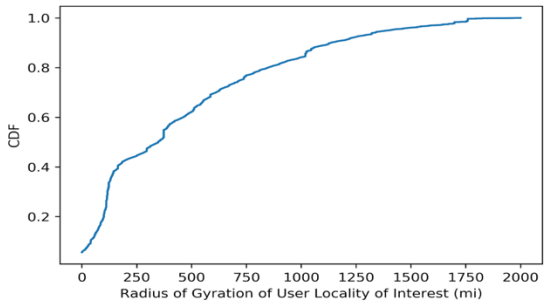


Figure 2B

References

Dangol, T. (n.d.). *Radius of gyration*. Science Topia. https://www.sciencetopia.net/physics/radius-gyration

Garbade, M. J. (2018, September 12). *Understanding k-means clustering in machine learning*. Towards Data Science. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Shearer, E., & Matsa, K. E. (2019, December 31). *News use across social media platforms 2018*. Pew Research Center. https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/