Twitter Feeds From Here to Across the Pond
A study of where the majority of tweets come from and their content
James McArver

EES201 – Introduction to Geographic Information Systems – Spring2012, Furman University, Greenville, SC

Abstract

This project used collected data from geocoded, public twitter feeds to map the origin of these tweets as well as their content. The data was collected from both the United States and the United Kingdom. Online programs displayed the most common words used in these tweets to give an aggregate picture of the majority of twitter posts collected over varying time intervals.

I. Introduction

There were two goals for this project. The first was to find out from where the majority of tweets sent from the United States of America and the United Kingdom were sent. The second was to analyze the content of the collected tweets and see which words and phrases were used most frequently. These goals were achieved using code written specially for this project and online services such as the Twitter Application Programming Interface (API). An example of some of the code used is as follows:

```
if (is_null($testname)){
    print 'null' . $delimeter;
} else {
    print $data['user']['screen_name'] . $delimeter;
}
$testdata = $data['text'];
if (is_null($testdata)){
    print 'null' . $delimeter;
} else {
    print urldecode($data['text']) . $delimeter;
```

This code allows us to grab the unique username, as well as the text of the tweet that follows. All of this information was printed with delimiters, to help the transition of turning this data into mappable points. Also written in the code was the functions that searched for geocoded tweets only, so that the information could be plotted on using its XY coordinates.

II. Literature Review

Dhiraj Murthy's article (July 2011), asserts that Twitter is breaking into the news reporting scene, as it allows ordinary civilians to suddenly become the people who report the news first. He uses heavily the example of the reports of the downing of the US Airways flight 1549, which were first captured by an eyewitness who tweeted a picture of the plane in the water. Murthy includes this example as well as the Mumbai bomb blasts in order to show how Twitter can help to break a story first, and how it can be used when normal reporting methods are either absent, or impossible. His article covers Twitter from the standpoint that Twitter is a good way to get eyewitness stories and views Twitter from a positive perspective.

Brian Smith's article, "Socially Disturbing Public Relations: Twitter, Haiti, and Interactivity in Social Media (Nov 2010), explains the concept of Twitter, as well as the impact the micro blogging site can have on society. He includes as a case study the earthquake in Haiti in order to illustrate the spread of information about a topic, the town hall type discussion generated by the tweets, and the different varieties of conversations that arise. While his article mostly proclaims the positive nature of Twitter during times of emergencies, he also points out the controversies that arise, and the duplicity of users in real life versus their tweets. Smith's article shows how Twitter, and discussions on Twitter, can simply turn into pandering, comment wars, or mindless dribble, despite the importance of the topic itself.

Laura Bly's article (Nov 2009) demonstrates the usefulness of twitter in situations that would otherwise be very time consuming and inconvenient, as well as the power the Twitter community may have over companies. The article is rather lighthearted and explains two stories about a writer who was stuck in the airport of a foreign country, and a songwriter whose guitar was smashed by United Airways and would not be reimbursed. These stories are told in order to affirm the main point of the author's article, which is that companies can no longer afford to mess up things, as they have the potential to go viral and gain a large amount of unwanted attention.

Palen et. all's article in Bulletin of the American Society for Information Science and Technology, documents the use of Twitter before and during the response of the threat of flooding near the Red River Valley. Palen et. all critically analyzed over 20,000 of the over 2,000,000 tweets that related to the topic in order to illustrate the topic's trend, the response to the flood threat via Twitter, and how others in the surrounding area reacted. While the response on Twitter to this event was relatively small compared to other events such as the earthquake in Japan, it nevertheless shows the communal reaction to threat. This article is similar to the research I will be attempting: geocaching tweets relating to different phenomena and events in order to visualize trends and patterns.

III. Methodology

After collecting data over several spans of time at semi-random time intervals, a base map of the world was created. This base map allowed the plotting of both the United States and the United Kingdom on the same map. The data collected consisted of all public, geocoded tweets sent from both countries over certain time intervals using the Twitter API program. These tweets were then added to a spread sheet, and logged by their respective latitudes and longitudes. By logging them as such, the resulting data could be added to the basemap as sets of XY coordinates. These points were plotted using the WGS1984 coordinate system.

After plotting the points, the kernel density function was used, to show areas of high twitter traffic, i.e., high amounts of tweets sent from the same relative area. This allowed for an easier understanding of the plotted points, as when the number reached a certain point, it was impossible to see if the number of points were increasing, simply because they continued to overlap.

In addition to creating a visualization of tweets geographically, the content of each individual tweet was analyzed using two online programs, Wordle and Worditout. These programs projected the most common words used in the tweets as larger than the others. Only the top 150 words, which were used a minimum of 500 times were projected.

IV. Results and Discussion
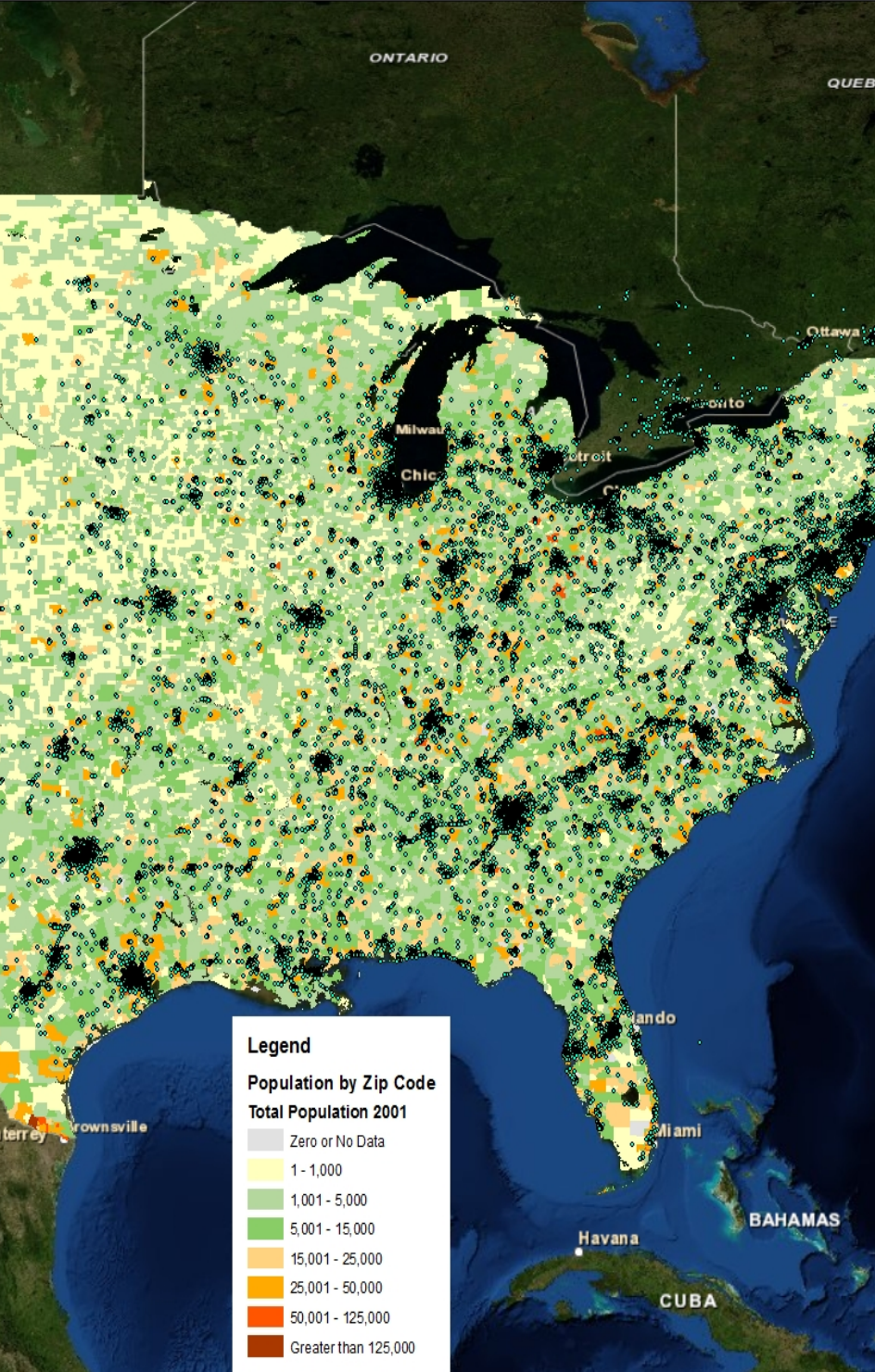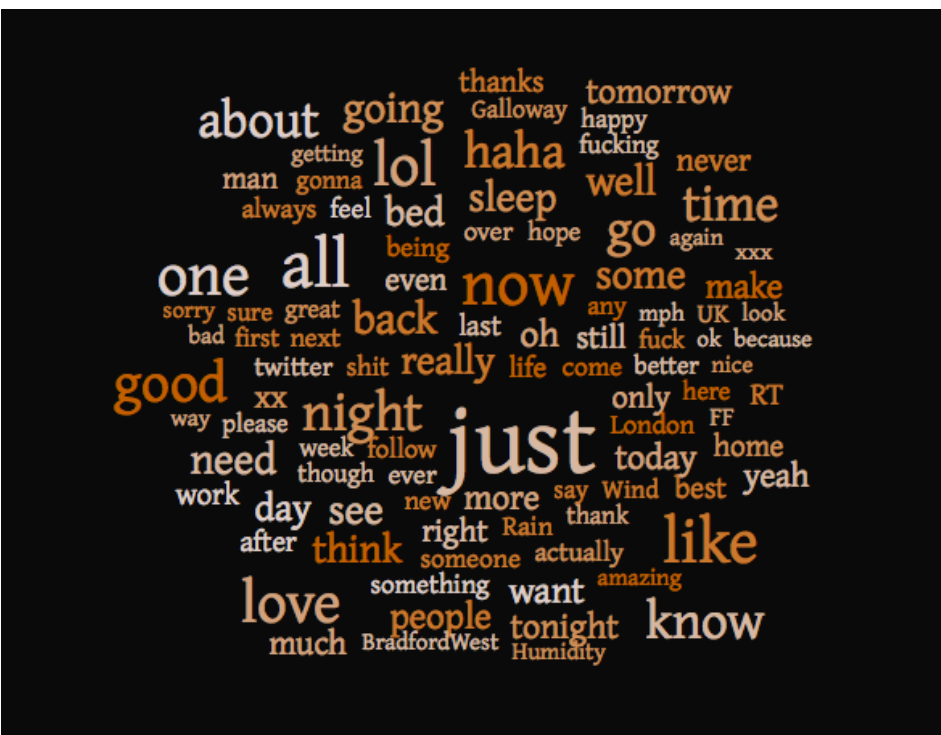
Maps showing part of the US in detail with the population of each zip code mapped, showing the clustering of tweets in these areas, and a kernel density of areas in the UK where tweets were the highest.

Visualizations of tweets from the UK (top) and the US (bottom)

Though the sample time is small, it is still possible to see differences in conversation between the United States and The United Kingdom. One example is that the word "God" is one of the prevalent words in the US tweets, but not the UK tweets. This can be explained most likely by the much lower religiosity rates in the UK than the US. Another example is the profanity use. For example, the US's use of the slang word "nigga," and prevalence of obscenities.

V. Conclusion

The majority of the tweets from both countries come from the more urban areas, such as London and New England. Upon comparing 2010 census data for the population of all zip codes, it was found that more tweets come from areas where the population and population density were higher. Unfortunately, This could be because of a number of factors, such as a prevalence of more "twitter-savvy" age groups, likely between the ages of 15-24, or a higher standard of living in these areas, allowing for the purchase or access to technology that offers the ability to tweet. Because of the limitations in the Twitter API itself, it is currently not possible to search for both a chosen subject and the location of the tweet itself, which severely limits attempts to narrow the focus of such data collection.

V.I. Future Research

I would like to conduct research in the area of why tweets come from these areas in a much higher number than other areas. To do this, I would like to look into other aspects such as median income, as well as other statistics of the populations. I also plan on accumulating data for 24 hours for both countries, so as to get a better picture of where these tweets come from, as currently my data is flawed by the difference in time zones. For example, much of the data for the United Kingdom includes words that reflect the time of day that this when this data was accumulated, such as sleep and evening. It was evening in the United Kingdom when I collected this data. Twenty four hours of data would eliminate this problem. It would also be interesting to examine the prevalence of words such as "just" from a linguistic standpoint.

VIII. Acknowledgements

VII. References

Palen, L., Starbird, K., Vieweg, S., & Hughes, A. (2010, June/July). Twitter-based Information Distribution during the 2009 Red River Valley Flood Threat. Bulletin of the American Society for Information Science and Technology, 36(5), 13-17. Retrieved from http://search.proquest.com/docview/746750137?accountid=11012

Bly, L. (2009, November 13). Tweets to the Rescue! USA Today, p. 8D.

Smith, B. (2010). Socially Disturbing Public Relations: Twitter, Haiti, and Interactivity in Social Media . Public Relations Overview, 36(4), 329-335. doi:10.1016/j.pubrev.2010.08.005

Murthy, D. (2011, July). Twitter: Microphone for the masses? Media Culture Scoiety, 33(5), 779-789. doi:10.1177/0163443711404744