Julian Gutierrez, Molly Wang, Aryaman Vir
May 3, 2015
NETS 150 Homework #5

Analyzing Public Sentiment on Twitter

Project Goal

Examine and analyze public sentiment on Twitter following occurrence of a major event.

Methodology

1. Collect all public tweets that are geotagged within time period of 5 days following the the event and filtering for those that contain the "Irene" query term.

   Note: Twitter's privacy setting has changed to block people from readily downloading public tweets, so we had to write custom code to scrape all tweets from within a certain time frame and then to parse all those tweets based on whether they contained our chosen keyword.

2. Assign a sentiment (1 - positive or 0 - negative) to each tweet, based on all of the words it contains

3. Aggregate tweets by the state with the closest geographic center based on longitude and latitude

Find overall talkativeness:

- Return the percentage of tweets containing a given term against all tweets in a given time span.

Find most talkative state:

- Return the state containing the most tweets containing a given term.

Find public sentiment:

- Find the cosine similarity of the body of tweets when run against two documents of positive and negative words respectively, and return the one with a higher cosine value
- Return (1 - positive) or (0 - negative) depending on whether the overall public sentiment about the event was more positive or negative

Variables

Tweet type:

- Public
- Geotagged

Time frame:

- Begin tweets date: 2011-08-28 19:02:28
- End tweets date: 2011-09-03 14:06:46

Event:

- Hurricane Irene
    - Keywords: “Irene”, “Hurricane”, “Hurricane Irene”
- Hurricane Irene was the first hurricane of the 2011 Atlantic hurricane season, occurring on August 20th and ending on August 28th. The hurricane had a widespread impact affecting Lesser Antilles, Greater Antilles, Turks and Caicos, Bahamas, Hispaniola, Eastern United States (Landfalls in North Carolina, Connecticut, New Jersey, and New York) and Eastern Canada. Irene caused 49 direct and 7 indirect fatalities. However, the exact number of fatalities varies from source to source due to political pressures and because not all fatalities were reported. An estimated damage of $16.6 billion was also caused. Irene was a massive storm – with hurricane force winds extending nearly 100 miles from the center – which brought a larger impact to land areas. Approximately three million people in New England lost power, severe flooding occurred in Vermont and New Jersey and North Carolina experienced strong isolated tornadoes and winds that caused damage in isolated areas.
- Sources:
    a. http://earthsky.org/earth/the-history-of-hurricane-irene
    b. http://en.wikipedia.org/wiki/Hurricane_Irene
    c. http://mceer.buffalo.edu/infoservice/disasters/Hurricane-Irene-2011.asp

Geographical bounds:

- 50 states of America & the District of Columbia

Hypothesis

We believe that states closest to the location of Hurricane Irene will be the most talkative and have a generally more negative sentiment about the event.

These states will be:
Southeast region - Florida, South Carolina, North Carolina, Virginia.
Mid Atlantic region - Maryland, Delaware, Pennsylvania, New Jersey, New York, New England, Connecticut, Massachusetts, Vermont, Maine.

States further away from Hurricane Irene will display a more positive sentiment or have no sentiment at all.

Results & Conclusions

Overall, we examined a dataset of 1,685,265 public and geotagged tweets within the five day period following Hurricane Irene.

Find overall talkativeness

3,411 of all tweets contained the keyword “irene,” which is approximately 0.2% of all tweets. The breakdown of the files we parsed is below.

File	Total # of tweets	# of tweets containing “irene”	Percentage
“tweets_1.txt”	575528	2431	0.4%
“tweets_2.txt”	575528	831	0.1%
“tweets_3.txt”	534209	149	0.02%

Note: “tweets_1.txt” contains tweets originating closest to the begin date/time of 2011-08-28 19:02:28. “tweets_2.txt” contains the second closest time set, while “tweets_3.txt” contains the furthest, ending at date/time 2011-09-03 14:06:46.

It makes sense that the further removed you are from the time of an event, there will be less people talking about it. We see this demonstrated as the percentage of total tweets about Hurricane Irene dips from 0.4% of total tweets on Twitter from the day after down to 0.02% up to 5 days after Hurricane Irene.

Popular events and subjects can often reach up to 3% of total tweets on twitter. The total tweets for hurricane Irene peaked at a significantly lower amount (0.4%). However, this number may not directly translate the concern of the people as the hurricane was not an event which affected countries globally. Americans only form 7% of the total active members on twitter and hence for a localized event, 0.4% is relatively high. The hurricane also caused widespread damage of power and telecommunication lines due to which people were unable to tweet. Therefore the number of total tweets about Irene is not an accurate measure of the concern of the people.

Find the most talkative state

Below is a table with a ranking of the states that had the most tweets containing the word “irene” to the states with the least tweets with the keyword.

State	# of tweets containing “Irene”

1. New Jersey	1112
2. Connecticut	405
3. Massachusetts	241
4. Vermont	172
5. District of Columbia	136
6. Rhode Island	126
7. Maryland	117
8. California	106
9. Florida	96
10. Texas	74
11. Maine	69
12. North Carolina	68
13. New Hampshire	65
14. Louisiana	61
15. Pennsylvania	58
16. New York	56
17. Georgia	48
18. Delaware	47
19. Illinois	44
20. Ohio	42
21. Virginia	42
22. South Carolina	38
23. Washington	21
24. Oklahoma	15
25. Oregon	15
26. Arizona	14

27. Nevada	13
28. Alabama	11
29. Tennessee	10
30. West Virginia	10
31. Kentucky	9
32. Indiana	8
33. Mississippi	8
34. Arkansas	7
35. Iowa	7
36. Utah	7
37. Colorado	6
38. Minnesota	5
39. Idaho	4
40. Montana	4
41. New Mexico	4
42. Wisconsin	3
43. Kansas	2
44. Missouri	2
45. Michigan	1
46. Nebraska	1
47. Wyoming	1
48. Alaska	0
49. Hawaii	0
50. North Dakota	0
51. South Dakota	0

We see that the most talkative states are in fact located in the region where Hurricane Irene occurred, proving our original hypothesis. The states listed in the hypothesis (most affected) also are in the top 16 in the above table. The results obtained are consistent with our hypothesis. The states that were affected the most, had the maximum number of tweets about the event. The number of tweets from the top 5 states (also the most affected states) is almost double of all the other states combined. The top 5 states account for approximately 45% of the total tweets. Some states had negligible amount of tweets which may have been due to widespread power cuts or because Twitter may have not been the social platform for communication. Apart from this, it is important to take into account that only 7% of Americans are active on Twitter and hence these results are not entirely reflective of the general behaviour of the people.

Find public sentiment

Note: We want to clarify what positive and negative construe. Examples of tweets with a more positive sentiment include those like "Cozied up with the fire and tea #irene" whereas negative sentiments include "Still out of power...Debri everywhere...Thanks Irene #irene."

State	Sentiment
Alabama	Negative
Arizona	Negative
Arkansas	Positive
California	Positive
Colorado	Positive
Connecticut	Negative
Delaware	Positive
District of Columbia	Positive
Florida	Negative
Georgia	Negative
Idaho	Positive
Illinois	Negative
Indiana	Negative

Iowa	Negative
Kansas	Positive
Kentucky	Positive
Louisiana	Negative
Maine	Negative
Maryland	Positive
Massachusetts	Negative
Michigan	Unknown
Minnesota	Negative
Mississippi	Positive
Missouri	Positive
Montana	Positive
Nebraska	Negative
Nevada	Negative
New Hampshire	Positive
New Jersey	Positive
New Mexico	Positive
New York	Positive
North Carolina	Positive
Ohio	Positive
Oklahoma	Positive
Oregon	Negative
Pennsylvania	Positive
Rhode Island	Positive
South Carolina	Negative
Tennessee	Positive

Texas	Negative
Utah	Negative
Vermont	Positive
Virginia	Negative
Washington	Negative
West Virginia	Negative
Wisconsin	Positive
Wyoming	Unknown

Surprisingly enough, we saw that there was a mix of sentiment toward Hurricane Irene across all regions of the state, disproving our original hypothesis. We actually see that states along the Northeast corridor and Mid-atlantic region where Hurricane Irene were overwhelmingly positive whereas states in the Midwest to West coast have more negative. A possible explanation for this is that the regions affected most heavily do not have access to power or electronics, and therefore cannot communicate their dismay or negative sentiments about Hurricane Irene on Twitter. Those with access are safer from the impact and are using the opportunity to speculate on the weather and how they're (happily) spending time with family and neighbors inside or going for a morning jog after and surveying the calm after the storm (these are example of common Tweet content themes that we've seen).

States further removed from the incident are potentially tweeting about the news coverage of the aftermath of Hurricane Irene, which is usually overwhelmingly negative. Media outlets will survey the damage done and report on cases of homelessness and lost friends and family as a result of Hurricane Irene. Because this is their only opinion of the event, it will make sense that it is more negative as a result.

Potential Sources of Error

Error could have risen from the fact we used the Euclidean algorithm to determine the tweet origin state. The Euclidean distance is calculated along a flat plane but Earth is spherical. Nevertheless, the Euclidean algorithm provides a very accurate calculation.

The dataset of tweets are all public and geotagged in the five day period following Hurricane Irene; therefore, we might not have a completely accurate reading of how talkative a state was about Hurricane Irene. There might have been many more tweets about the hurricane that we could not access either because they were from private Twitter accounts or because they did not contain a geotag. It's estimated that approximately 15-20% of all tweets are geotagged.

We assumed that in the five day period following the hurricane, an extremely large majority of people tweeting anything with the mention of “Irene” will be about the hurricane. However, there is a very small probability that we accounted for tweets that mention “Irene” but are not about the hurricane and instead about a person or other object/event.

Code Implementation

Finding the state a tweet originated from

1. All scraped tweets are stored in the format [longitude, latitude, text]
2. Use the coordinates of the center of each state (according to the input file)
3. Assume that the tweet originated in the state to which its coordinates are closest.
4. For simplicity, assume the Earth is flat and calculate distance by implementing Euclidean algorithm. This approach is not completely accurate, but it provides a relatively accurate formula

Source: http://en.wikipedia.org/wiki/Euclidean_algorithm

Scraping the Twitter database

1. Remove lines that don't start with coordinate system (find and replace with "")
    a. `^(?!\[).+`
2. Remove tweets that are not in the US (second coordinate positive) (find and replace with "")
    a. `^([0-9]*\.[0-9]+),\ ([0-9]*\.?[0-9]+).+`
3. Since all the deleted lines remain empty, get rid of them (find and replace with "")
    a. `^\n`
    b. Source: http://stackoverflow.com/questions/12008986/sublime-text-2-how-to-delete-blank-empty-lines
4. Delete number and timestamp for all lines (find and replace with "") (there will be a tab as separation from the original file)
    a. `\]( |\t)+[0-9]( |\t)+(19|20)\d\d[- /.](0[1-9]|1[012])[- /.](0[1-9]|[12][0-9]|3[01])( |\t)+(([0-1]?[0-9])|([2][0-3])):([0-5]?[0-9])(:([0-5]?[0-9]))`
    b. See *breakdown* below for details on how this value was determined
5. Delete opening square brackets
    a. `^\[`

6. Comma between two coordinates (losing 1 digit in floating point) replace with " -"
    a. `[0-9], \-`
7. Split in 3 equal parts to accommodate for limited memory space
    a. `tweets_1.txt, tweets_2.txt, tweets_3.txt`
8. Done! Now we have a text database of all public tweets within a defined time frame

---

Breakdown of how 4a was determined

>

Bracket in the beginning, replace with ""

```
^\[
```

>

Comma (“,”) between two coordinates (losing 1 digit in floating point) replace with " -"

```
[0-9], \-
```

>

Closing bracket tab in the middle and number

```
\]( |\t)+[0-9]( |\t)+
```

>

Hour xx:xx:xx replace with ""

```
^(([0-1]?[0-9])|([2][0-3])):([0-5]?[0-9])(:([0-5]?[0-9]))
```

Source:
http://regexlib.com/DisplayPatterns.aspx?cattabindex=4&categoryId=5&AspxAutoDetectCookieSupport=1

>

Date

```
(19|20)\d\d[- /.](0[1-9]|1[012])[- /.](0[1-9]|[12][0-9]|3[01])
```

Source: http://www.regular-expressions.info/dates.html

Other Sources

Original Set used (all_tweets.txt):
http://nifty.stanford.edu/2013/denero-muralidharan-trends/data/

Task Flow

Molly

Aryaman
Julián

1. Get states coordinates and code to separate tweets into respective states by calculating Euclidean distance from origination of tweet by longitude & latitude
2. Find set of negative & positive keywords
3. Project write-up and analysis

1. Separate negative & positive keywords
2. Project write-up and analysis

1. Scrape Twitter for relevant dataset
2. Code from a JSON and get similarities from list of positive & negative keywords (use the cosine similarity against difference queries
3. Writeup on how the keywords were curated.

Comment on the findings with plots and data.