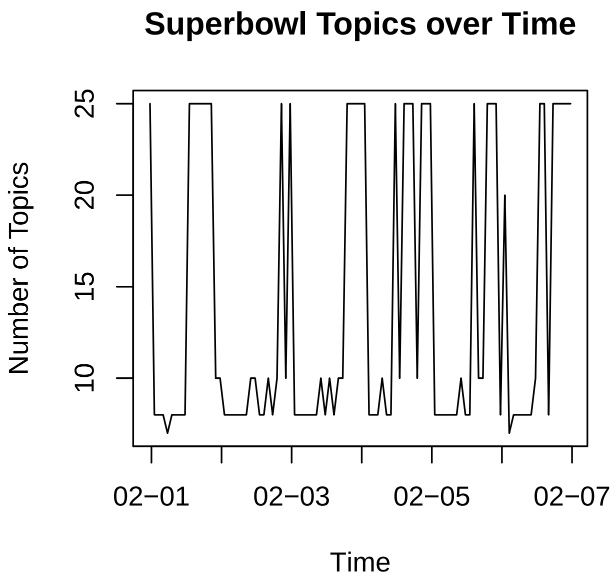
Paper Summary

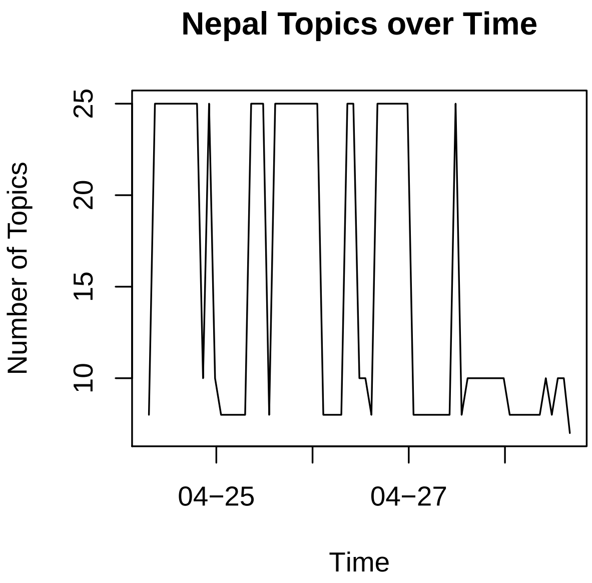
Robert Turner

February 28, 2016

For this project we used topic models to analyze Twitter's response to major events. More specifically, used the maptpx package to fit Bayesian topics models to successive time intervals before and after the 2015 Superbowl and the Earthquake in Nepal. In this way we hoped to see Twitter react to these events and also how exactly they reacted. For example, Superbowl discussions could potentially be about the game, the players, or even the commercials. Similarly, Twitter users could talk about the Nepal Earthquake happening, the damage done by it, or ways to help. By fitting successive topic models we also wished to see how long topics related to these two events could be seen. We also hoped to compare the response to these two events, as although they are both popular events they both have a different nature. The Super Bowl is popular due to its entertainment value, while the Nepal earthquake is talked about due to the severity and impact of the event. The number of topics for each time interval was chosen independently and automatically by choosing the model with the highest Bayes factor. This resulted in models ranging from 8 to 25 topics. The results were then visualized using the LDAVis package. The Tweets were downloaded with no keyword or location filtering using the streamR package and any non-English Tweets were removed before analysis.

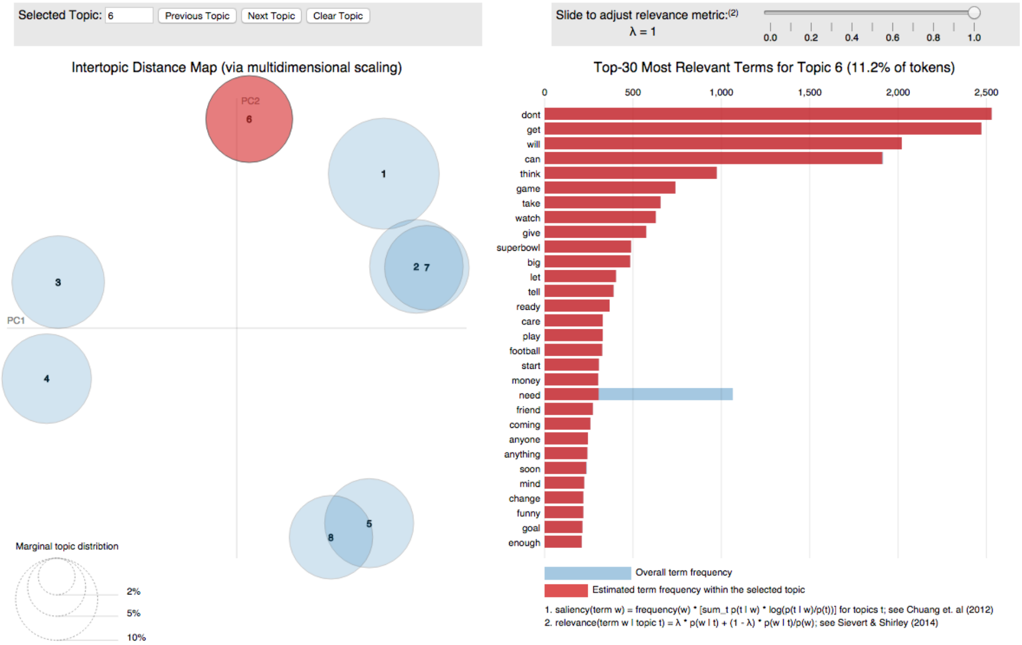
Considering only the number of topics, there seems to be a weak trend with the time of day and the number of topics. Models fit later in the day typically had more topics than those fit early on. This matches up with Twitter activity as most Tweets are sent after noon local time. However, this relationship appears to be weak as the time of day when the number of topics increases seems to vary across the days.





The topic models fit showed a response to both of these events immediately after they happened. For the Superbowl, some topics relating to it were seen before the game started, but these topics simply referenced the game happening. During the game not only were there more topics in general, there were more topics referencing specific aspects of the Superbowl. For example, highlighted below is a topic discussing the half-time show of the game. Finally, after the game ended the topics began to condense again, resulting in 1 or 2 topics referencing the Superbowl as a whole. An example model for each of these time periods is shown below, and more can be found on the referenced github page.

Pre Superbowl 8 Topic Model:



Selected Topic: 6

Slide to adjust relevance metric:(2) λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 6 (11.2% of tokens)

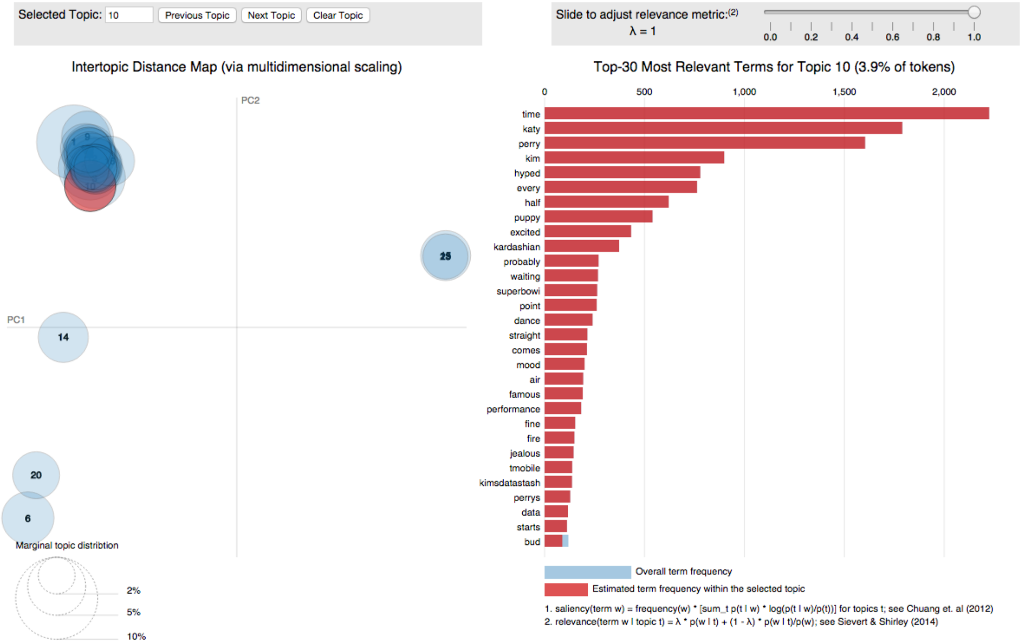
Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Mid Superbowl 25 Topic Model:



Selected Topic: 10

Slide to adjust relevance metric:(2) λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 10 (3.9% of tokens)

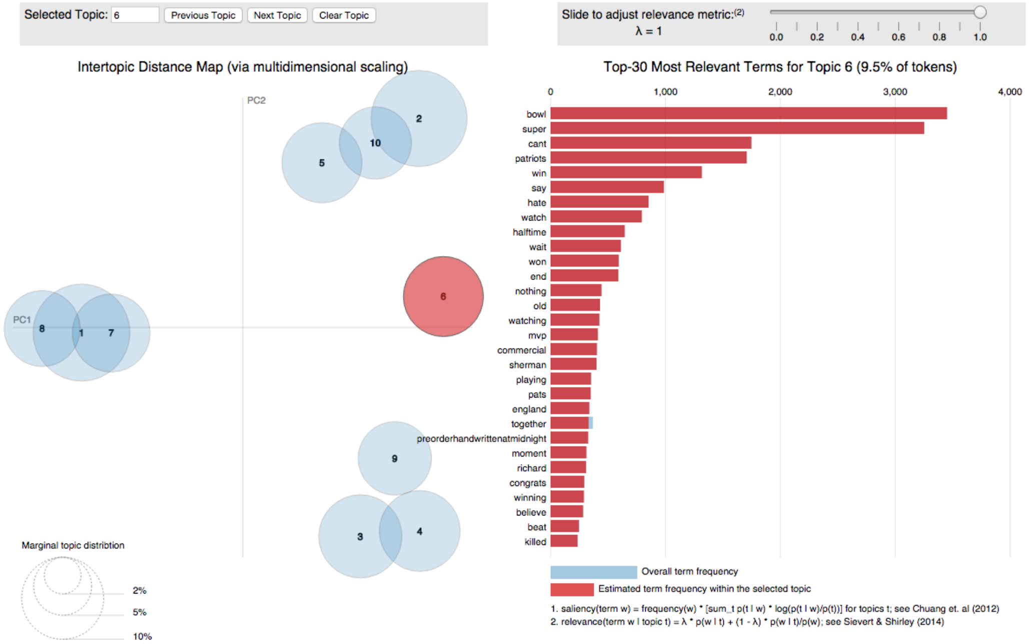
Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

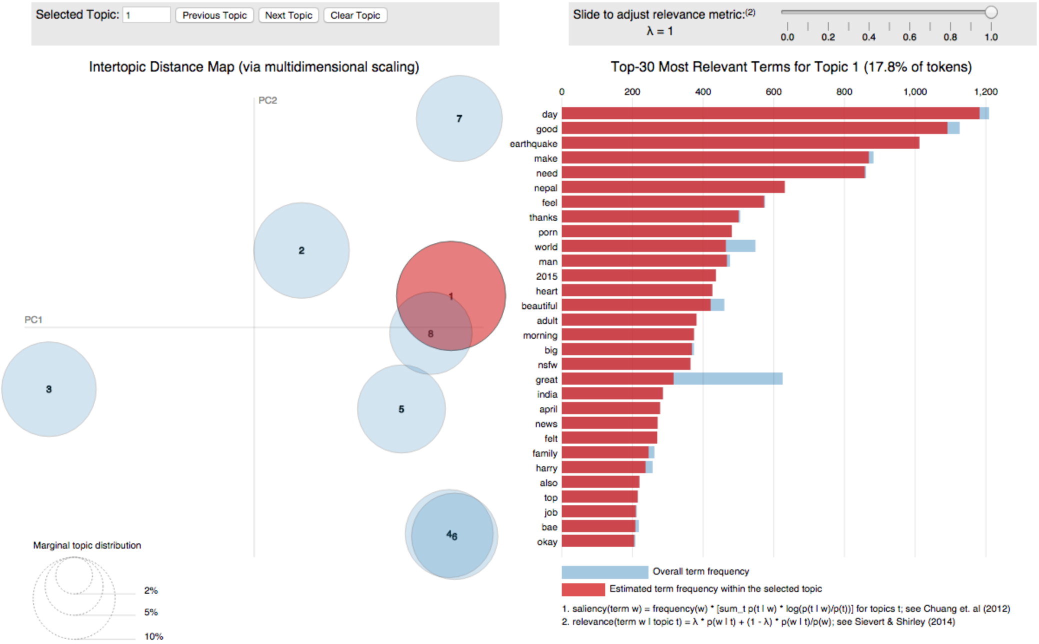
Post Superbowl 10 Topic Model:



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

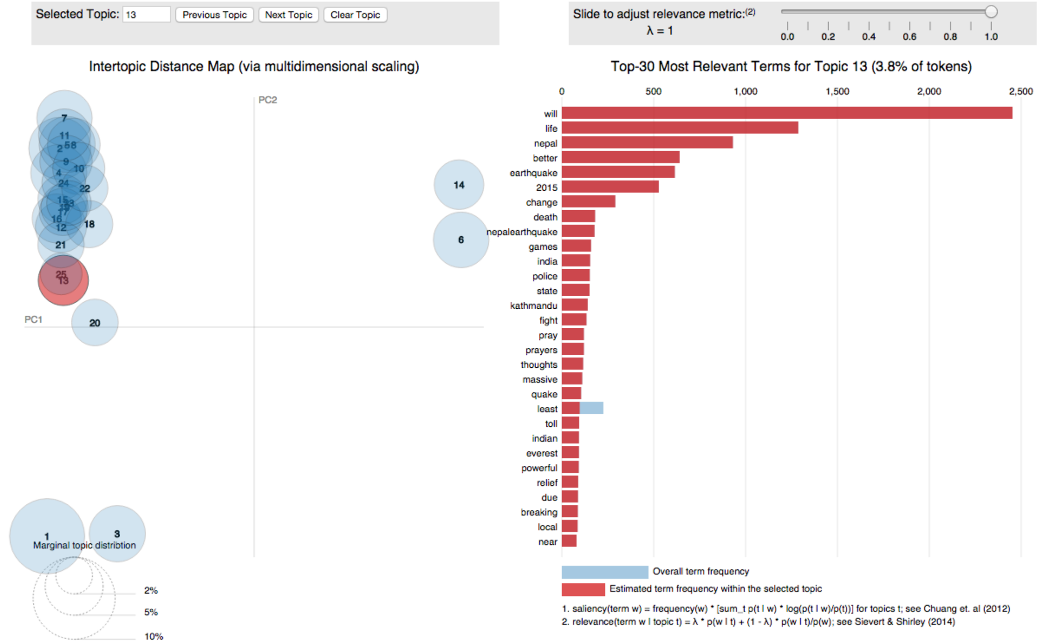
Tweets immediately following the earthquake showed a response to the event, but this response was contained in a topic with other irrelevant terms. This is likely due to the fact that the event took place in the very early morning for most Americans, and this analysis was limited to English Tweets. As time passes a more detailed Nepal topic forms, referencing the specific areas damaged. While there are not as many topics referencing Nepal as there were the Superbowl, there is still a noticeable response to this event shortly after it occurred. In addition, while topics referencing the Superbowl dissappear within the period analyzed, terms referencing Nepal are seen throughout the full time period. Examples of these topics are given below.

Immediately after Nepal Earthquake 8 Topic Model:



1. saliency(term w) = frequency(w) * [sum_t p(t I w) * log(p(t I w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w I topic t) = λ * p(w I t) + (1 - λ) * p(w I t)/p(w); see Sievert & Shirley (2014)

Days after Nepal Earthquake 25 Topic Model:



1. saliency(term w) = frequency(w) * [sum_t p(t I w) * log(p(t I w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w I topic t) = λ * p(w I t) + (1 - λ) * p(w I t)/p(w); see Sievert & Shirley (2014)

In addition to these results we also produced an R package containing all of the code necessary to format Tweets for analysis. This package was designed for files downloaded using the streamR package and produces a folder of text files containing the edited tweets. The package also selects desired variables, filters stop words, cleans up links and other Twitter artifacts, and can produce a data frame ready for model fitting. The package can also perform the same model fitting procedure carried out for this project. This package can also be found on the referenced github.

This project shows that topic models can be an effective method of event detection on Twitter and can also provide insight into how users are discussing these events. Both events were detected very shortly after they occurred, and the Superbowl generated numerous topics referencing unique aspects of the game. Not only do we know that people were talking about the game, we know that some were talking about the commercials, some were talking about the half-time show, and some were talking about the game or the players. This is a finer insight into Twitter's response to this event than simply noticing the response occurred. In addition this project produced an R package and code to recreate these results and to ease others into the topic modeling process, hopefully allowing facilitating further analysis in this area.

Github Links:
R package: https://github.com/rturn/parseTweetFiles
Paper and Results: http://github.com/rturn/Topic-Modeling-Twitter