Assignment 1 - Tweet Volume Analysis MSE 231

Teammates: David Lang, Nikolas Martelaro, Tres Pittman

Due Date: October 12, 2016

Filtered Tweet Analysis: `SNL`

We used the Twitter API to investigate what phenomena are centralized versus localized to a particular event in a particular region. We attempted to do this by looking at tweet volume for Saturday Night Live. In particular, we looked at the phrase `SNL` during the 24 hours in which Saturday Night Live premiered its new season (October 1, 15:45 PST - October 2, 16:30 PST).

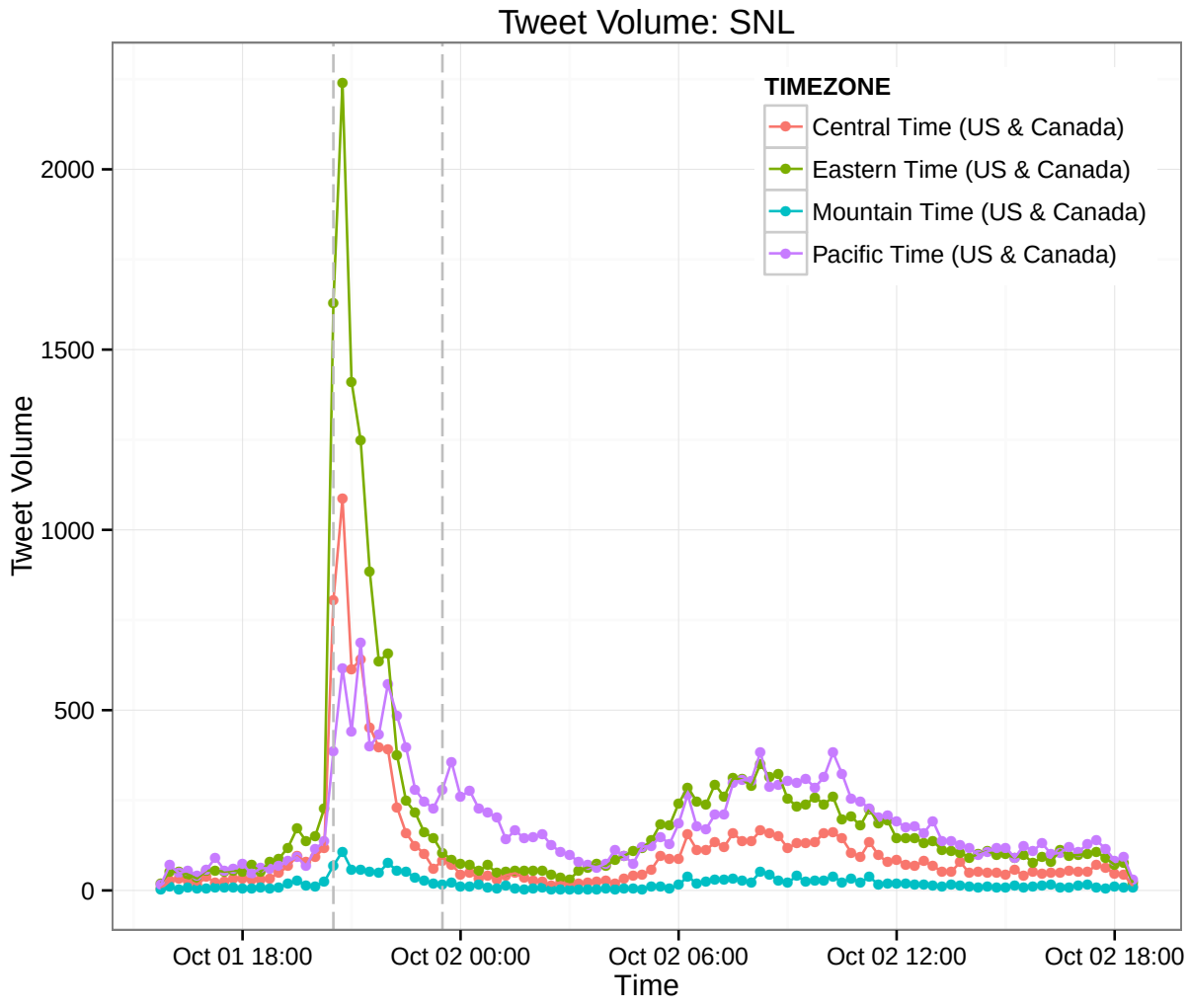

Figure 1: 24 hours of tweets using the phrase `SNL`

Our initial assumption was that the time series would spike at different points in time. Our rationale was that the television show broadcasts concurrently in Central and Eastern Time. On

the West Coast, Saturday Night Live airs approximately three hours later. For these reasons, we anticipated that tweet volume would be staggered, spiking first in the Eastern and Central time zones and then in the Pacific time zone. Figure 1 shows the tweet volume over our 24-hour data collection period. There is a spike in tweets in all regions just after 20:30 PST (8:30PM) indicated by the first vertical grey line. There is a small spike at 23:30 (11:30 PM) in the Pacific time zone but not nearly as large as one would expect based on the earlier spikes. The plot is suggestive that the SNL phenomenon is not driven by regional television broadcasts.

Though we cannot know exactly why the spike in tweets occurs simultaneously in all time zones following the start of the premiere, we can give some basic intuition or theoretical justification. For one, this national phenomenon suggests that Americans are following and engaging with SNL on Twitter as the show is being broadcast on the East Coast. Even if the show has not yet aired on the West Coast, Twitter users will see the Eastern or Central tweets as they happen and view clips of the show that other users post. Thus, West Coast tweeters seem likely to engage with those around the country who can view SNL as it is playing, underscoring the ability of Twitter to transcend traditional media outlets. Alternatively, the national surge in tweets could result from the prevalence of "cable cutting" and the gradual demise of traditional cable media. Live streaming programs may enable viewers throughout the country to watch the premiere when it is first available, in this case at 11:30PM EST. Of course, those Americans with the technological savvy to view popular television shows like SNL on the internet are much more likely to be active Twitter users than people who are unaware of web-based alternatives to television.

Another surprising period in Figure 1 occurs during the morning of October 2nd from about 06:00 - 12:00. Tweet volume consecutively rises and falls within these six hours, most likely due to people waking up and discussing the show. Interestingly, by midday the tweet volume for SNL has decreased substantially, and by about 15:00 the numbers of tweets have stagnated. This decline suggests that people were engaged with SNL in the morning but moved on to other topics later in the day. It is possible that overall tweet volume may drop after 15:00; however, our plot of non-filtered tweets over 24-hours suggests that tweet volume is fairly constant from 12:00 - 20:00.

Non-filtered tweet analysis

Our subset of non-filtered tweets, shown in Figure 2, suggests that as a proportion of the Twitter population, the SNL figure is even more dramatic in terms of the spike in user behavior. The East Coast clearly dominates in our SNL plot; however, East Coast tweets remain a small fraction of Pacific Time tweet volume. Other intriguing facts we note about the unfiltered data include that tweet volume tends to rise from 3AM to 9AM in all time zones. One possible explanation is that individuals tend to engage with Twitter to refresh themselves on news in the morning. We also note that that tweet volume remains relatively constant across all groups from 12PM to 9PM Pacific Time. One possible explanation is that much of Twitter behavior during this time is driven by bots or prescheduled posts rather than humans. We also note that tweet volume begins to drop at around 9:00 PM Pacific Time. We suspect that there is a substantial amount of cross-pollination across time zones (i.e., people in Pacific time zones have strong preferences to engage with users on the East Coast.)

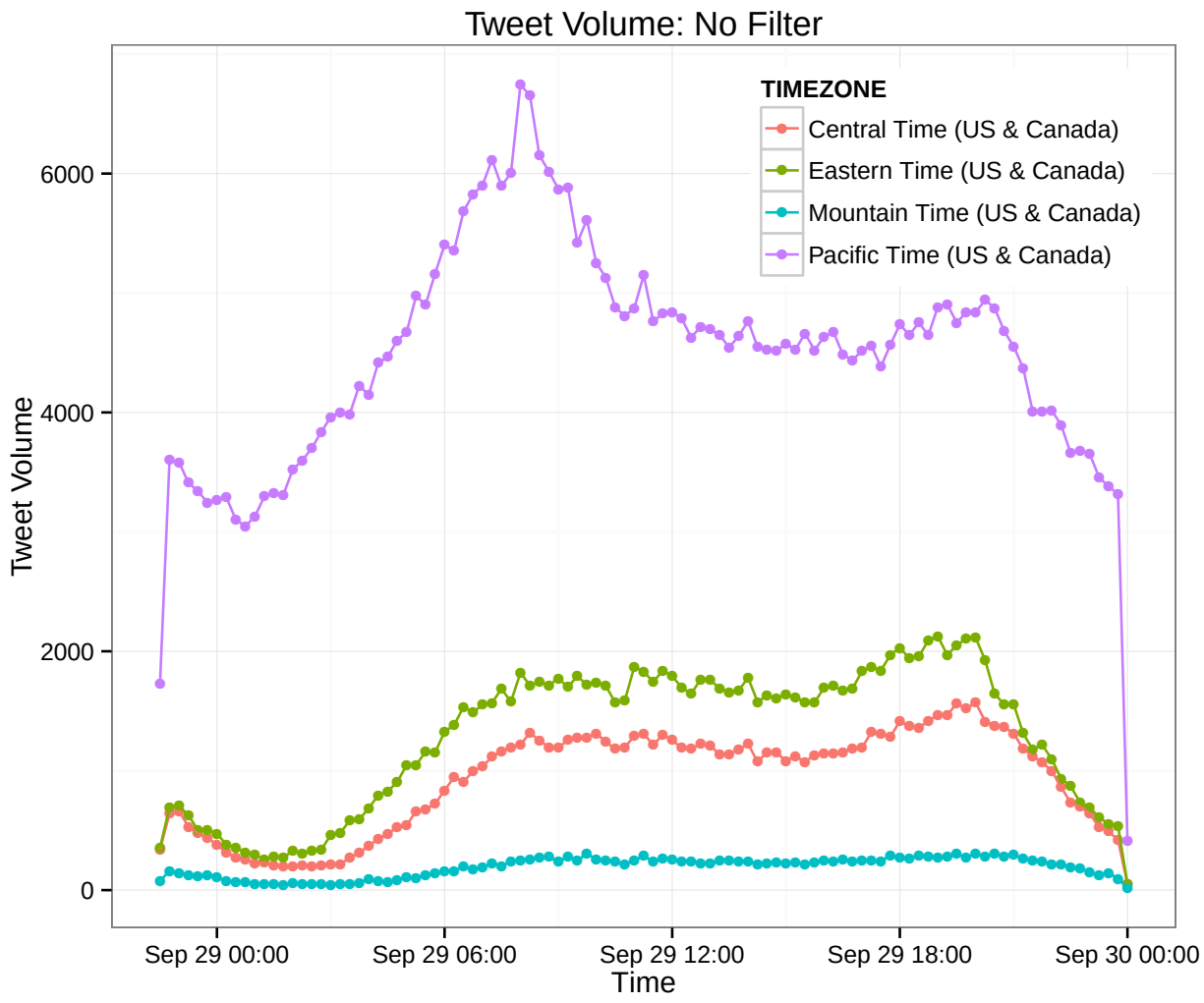

Figure 2: 24 hours of tweets using without filters

Possible Limitations to Our Analysis

Although tweet volume can generally be considered a good proxy of interest, there are reasons to doubt that that our search term was isolated to tweets pertaining to Saturday Night Live. SNL is an initialism associated with Sandia National Laboratories and other organizations. If any of these bodies had released a major press release or made an announcement, it may have skewed our results. Given the timing of our data collection, we doubt that this effect was meaningfully present. However, for terms that may have have double meanings or different contexts, caution should be taken.

In addition, we note that our unfiltered tweets show extreme drop-offs at both the beginning and end of our data collection session. This feature is likely an artifact of the way we binned the data insofar as our first and last 15-minute bins were incomplete.

More broadly, our methodology was limited by the parts of the JSON output that we ignored.

Because we focused exclusively on the date, time, and time zone of each tweet, we ignored other data - notably the actual content or text of the tweet. Although much of the JSON data was likely superfluous, by ignoring the text of the tweet, we may have been overweighting automated or Twitterbot-generated content. However, this effect was likely minimal, as a cursory review of the raw output suggests that the majority of the tweets are at least ostensibly authentic. Nevertheless, a more robust analysis might have attempted to clean the data of nonsensical or clearly artificial tweets.

It should also be noted that by ignoring text content, many of the same tweets are included multiple times in our dataset (retweets). Although a retweet may be considered a legitimate measure of the popularity of a subject like SNL or Twitter traffic in general, it requires much more effort to write an original tweet than to simply click the retweet button. At the very least, a further analysis might consider removing retweets with no additional text but leaving those where users retweet with a comment of their own. By treating all tweets as equal, our analysis may have overestimated tweet volume in general or overweighted timezones where many users passively retweet. (Of course, if this phenomenon were uniform across the US, then adjusting for retweets would likely be unnecessary). Comparing retweet to original volume could have lent insight to our investigation but would have required a serious computational effort since not every retweet is designated with RT in the JSON data.

Next, our analysis ignored the Arizona time zone. Our SNL data alone contained over 4,000 tweets from this time zone. Although considering it independently would have yielded negligible results, it would have been logical and instructive to reclassify tweets from Arizona to the Pacific time zone since Arizona Time is identical to Pacific Time during the autumn months.

Overall, the JSON data is quite rich and an analytical project could have gone much further. Filtering data by verification status and follower count, for instance, could have yielded additional if unsurprising results. Nevertheless, our basic timezone analysis was interesting while being relatively straightforward.