PRINCIPLES OF COMPLEX SYSTEMS

MORPHOLOGY OF TOPICS ON TWITTER

January 11, 2019

David W. Landay
University of Vermont
Graduate Student, Comp. Systems and Data Sci.

Abstract

We propose a method to analyze inter-temporal topic dynamics of collective discussion on Twitter. The large twitter pipeline – the **Decahose** – located on the Vermont Advanced Computing Core (VACC), serves as our dynamic representative of global discussion across fifteen minute time intervals. In this analysis, tweets are considered documents, and the collection of tweets, or corpera, at any given time interval represents a global conversation at time τ. We then extract topical information at each time step using Latent Dirichlet Allocation (LDA) to model the aggregate driver of discussion on twitter, for any current moment in time. Given the quantity and robustness of the available dataset, we argue that the propagation of topics through time can be described by changes in entropy between subsequent per-topic distribution over words. We hope that the findings in this preliminary investigation will encourage further study into a method for near real-time analysis of collective interest on twitter.

0.1 Introduction

The internet serves as the hub for our collective consciousness. Never before has the ability to share our thoughts and feelings with other humans been so readily available. Online social media services offer users a platform for not only interacting over a virtual connection, but collaborating, meeting, and organizing in real life. Increasingly, platforms like Facebook and Twitter are being used to organize protests and politically charged meetups in the real world. But, what happens when these events lead to mass violence? Given information prior to an event about unified conversation, can we anticipate the impact of a social gathering in real life? Tragic events like those which unfolded at the "Unite the Right" rally in Charlottesville, VA on August 12_{th}, 2017, motivate the search for answers to questions like these. In this preliminary report, we explore an experimental method to find out whether an analysis of such a complex system is possible.

For this study, the availability of large quantities of tweets located on the Vermont Advanced Computing Core (VACC) affords an opportunity to seek out latent topic structures over national conversations around the time that the events in Charlottesville were unfolding. Using Latent Dirichlet Allocation, we attempt to find important topics over time specific collection of tweets, and then determine when a subject becomes a point of unified discussion via the amount of entropy expressed between two "per-topic" distributions over words. Specifically, we will use the Jensen-Shannon Divergence, as a metric for how similar a topic is between two consecutive corpora.

0.2 LDA: Latent Dirichlet Allocation

Latent Dirichlet Allocation, LDA, is a Bayesian statistical model introduced by [3] that learns a "Dirichlet" prior on the per-document topic distributions, via a prior on the per-topic word distribution that gets randomly assigned at initialization. The model makes a few essential assumptions. Namley, that words hold a good deal of symantic information whereby documents discussing similar topics contain similar distributions of words, and every document contains a certain number of topics. One limitation of the model however, is that the number of topics present within a corpus must be known, and serves as a input to the system. In our case, it is impossible to directly assess the performance of each model via some metric, because we have no prior knowledge of the topics that define each collection of

tweets. But, One saving grace is that tweets are limited to 240 characters and are of narrow scope topically. Hence, the need for tunable parameters is out of the question, but may not hinder proper classification of topics.

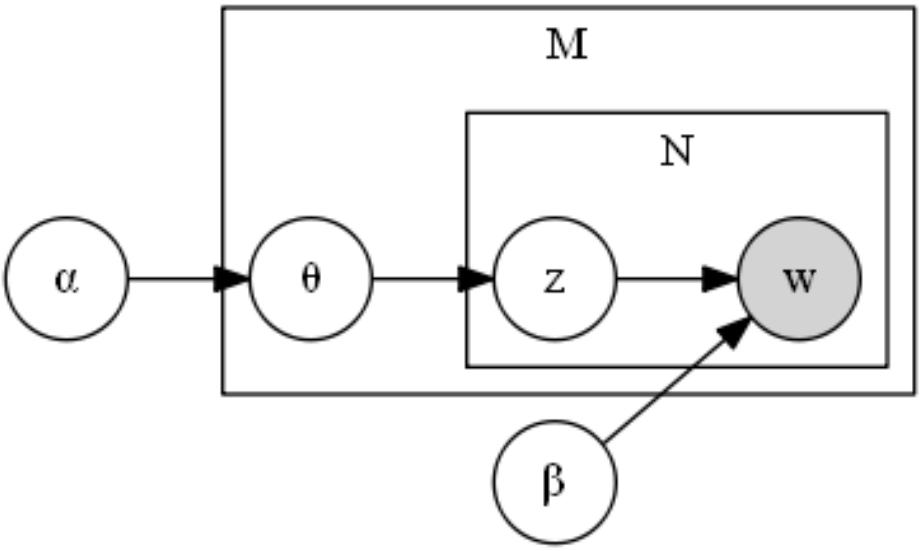


Figure 1: A plate notation visualization of the LDA model. [3]

0.3 K-L Divergence

Kulbeck-Liebler Divergence, is a measure of informational cost, and is the "relative entropy" between two distributions. In our case, "How well does topic p translate to topic q". This is an asymmetric measure, thus, is not a statistical metric that we will use to measure the similarity between two per-topic distribution over words. Jensen-Shannon Divergence, on the other hand, sets a lower bound of 0 and an upper bound of 1 over our comparison of two distributions, and is thus used a measure of distance between two distributions. By comparing each distribution to the mean of the two, instead of each other directly, we arrive at the calculation for JSD, which is given by

$$\text{JSD}(\,P \,||\, Q\,)\,\frac{1}{2}D_{KL}(\,P \,||\, M\,) \,+\, \frac{1}{2}D_{KL}(\,Q \,||\, M\,) \tag{1}$$

where $M = \frac{1}{2}(\,P + Q\,)$ is the average between two distribution vectors, P and Q, and D_{KL} denotes the Kulbeck-Liebler Divergence over $P \,||\, M$ and $Q \,||\, M$ respectively.[2]

0.4 Finding the Weighted Importances of Topics

To determine the focus of discussion at any time step, we wish to know the topic outputs that best describe the collection of tweets for a given fifteen minute interval. By finding an association between each tweet in a current corpus and each of the topic outputs from the LDA model, we may weight the importance, or relevance, of each output by the number of tweets that contribute to it. By doing this, we also generate the distribution over words that best represents the topic.

We associate tweets to topics using maximum-likelihood estimation. The likelihood that a tweet belongs to a topic is determined by it's largest per-document topic distribution. Specifically, given a sparse vector representation of each document, we can transform the data according to the LDA model that has been trained over an entire corpus. **Fig.2** shows the per-document topic distributions for several time specific corpora.

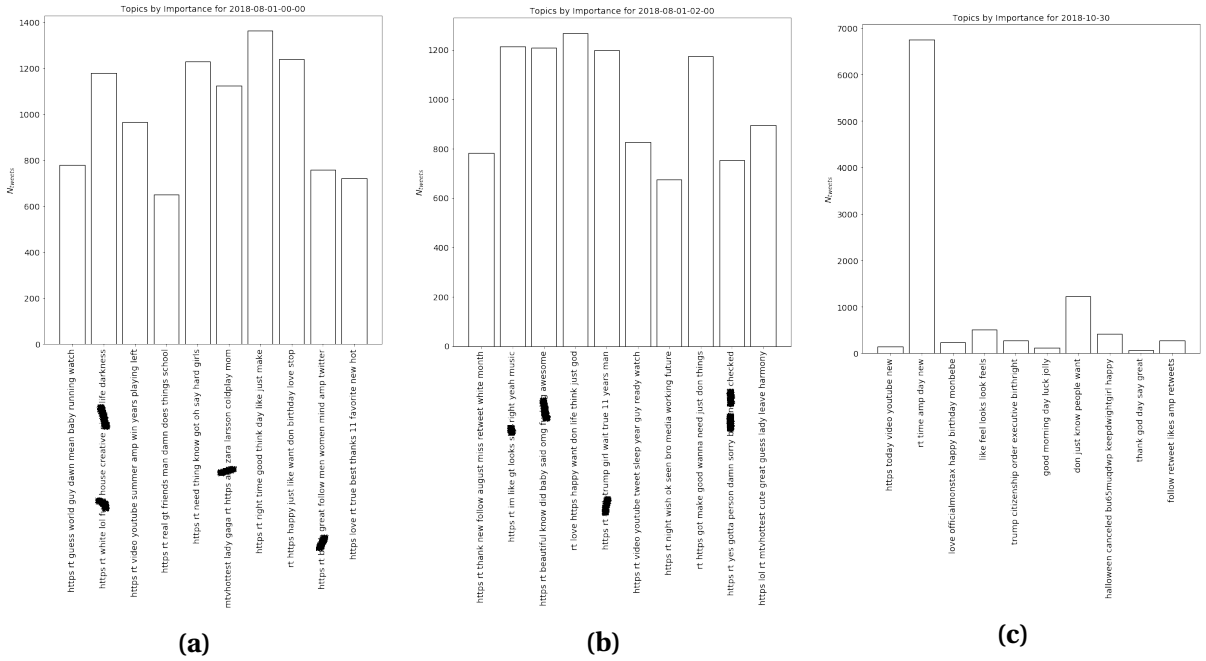


Figure 2: Examples of topic outputs from LDA models trained over time specific tweet corpora. **(a)**, **(b)**, and **(c)** show frequency counts of the number of documents (tweets) that each topic is referenced by. Each set of tweets is the distribution over words that the topic represents. The top 5 topic features are displayed for each plot.

The topic distributions are generated from a sample of all tweets in each corpora. For this initial study, we argue that a sample size of 10,000 tweets from which to build per-

document topic distributions, is representative of the entire corpora. **(a)** and **(b)** are two distributions looked at in this preliminary study. The inclusion of **(c)** is to demonstrate that distribution over topics are not always uniform. Perhaps this shows that one topic dominates the conversation over that time period, but likely indicates that not enough data cleaning has been performed over each set of documents. Notice how the word "amp" appears as a feature in second and last topics. Furthermore, we might expect that the topic second from the last, on the right hand side of the x-axis, would best represent the highest ranked topic because "halloween" is the event that occurs the following day. The number of hyperlinks and retweets, 'rt', that appear as topic features is also a concern and should be accounted for before further investigation. However, hyperlinks that point to specific articles may shed more light onto the overall focus of discussion. For a future experiment, an analysis of the hyperlinks might prove useful towards understanding latent topic features.[4][4]

0.5 Results & Discussion

A preliminary experiment was conducted over a two hour span of twitter data. The data covers eight corpora; from midnight on August 1_{st}, 2017 to $2_{a.m}$ on the same day. The sparsity in the preliminary data set is due to technical complications with the VACC. However, the infrastructure for programatically preprocessing data is developed and in place for future trials.

In the previous section, we generated the per-topic word distributions for each corpora, and rank-ordered the topics based on their relative importance, or the number of documents that contribute to the creation of the distributions. Recall that every corpora is defined by 10 distinct topics, and the per-topic word distribution is extracted from a sample of 10,000 tweets from each corpora. Below, **Fig. 3** looks at the similarity between the distributions corresponding to the i_{th} ranked topic for the midnight corpus, and i_{th} ranked topics corresponding to every subsequent corpus out to $2_{a.m}$. For example, we compare the top ranked topic at 2017-08-01-00-00 to the top ranked topic at 2017-08-01-00-15, 2017-08-01-00-30, and so on up to 2017-08-01-02-00, by measuring Jensen-Shannon divergence between the two word distributions. We then do this for next topic in the rank hierarchy, the next, and so on up to the 10_{th} topic. Recall that the format for the per-corpus time-stamp is `%Y-%m-%d-%H-%M`, or "year-month-day-hour-minute".

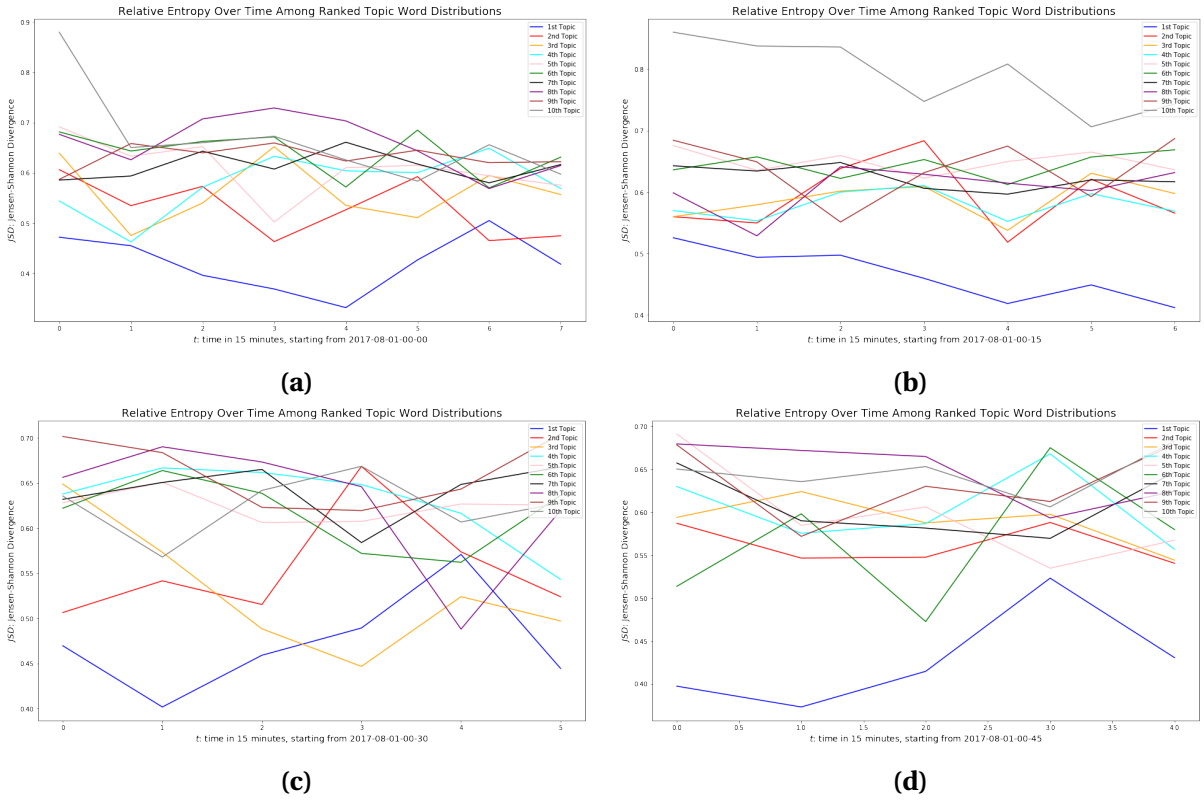
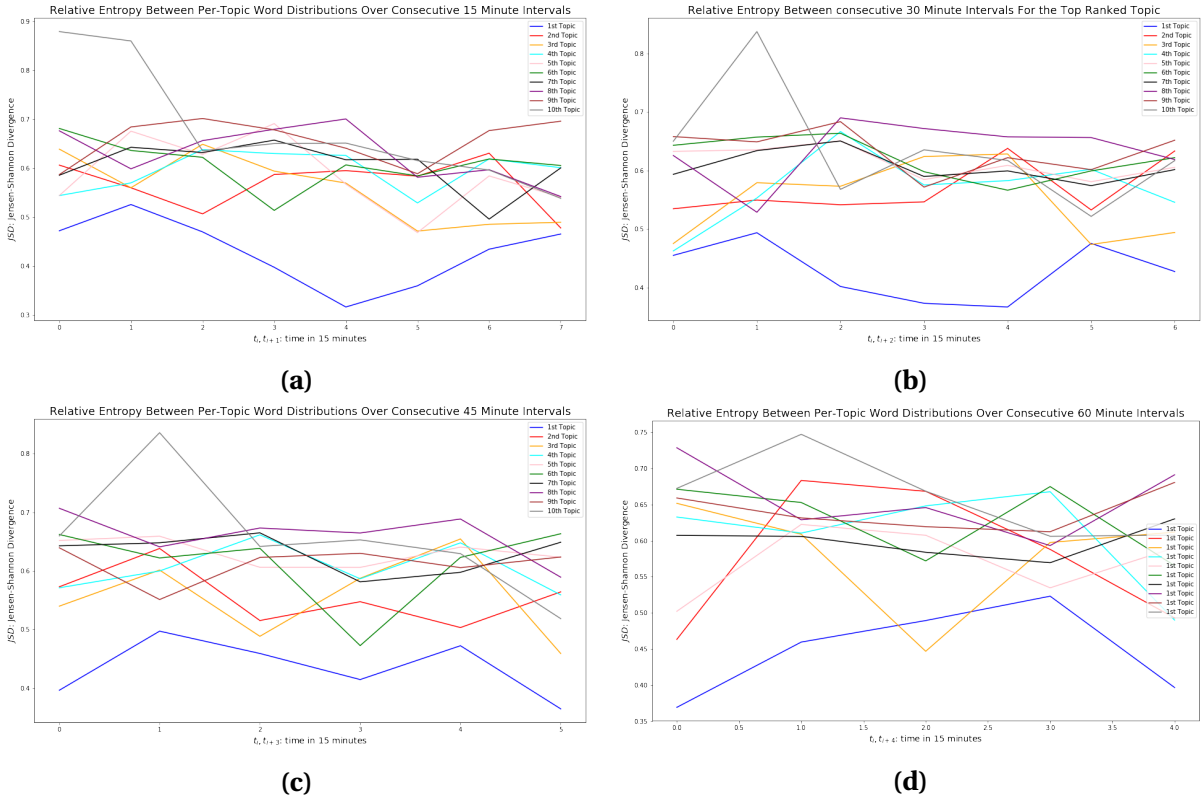


Figure 3: Jensen-Shannon Divergence between each of the rank ordered i_{th} topic word distributions forward in time from a specific date. Word distributions are generated from time dependent corpora of tweets.

The relative entropy over each per-topic word distribution hints at the subjects that are talked about most, consistently over time. We must, once again, preface that the lack of continuous data hinders us from understanding any underlying story that may be emerging, but their is a convenient pattern among the trend lines that warrants further investigation. Specifically, from **(a)**, one can see that, with the exception of 4_{th} and 8_{th} topics, the hierarchy in topic rank is preserved by the measure of JSD. More specifically, from 2017-08-01-00-00 to subsequent corpora, the JSD between the distribution over words for the top ranked topic, remains the lowest overall, the second ranked topic appears second lowest etc... If there is little difference between the per-topic word distributions from 2017-08-01-00-00 onward, then it could be that there is little variation in global discussion of each topic. Note that this could be determined by the number of topics (components) k we parameterize the LDA model by. Also note that between each time-delta, most of the topics exhibit the same amount of entropy. We expect that entropy should increase as we compare distributions that are further, and further, away from each other in time. **(b)** actually indicates a trend

towards unified conversation, but the other comparisons do not. However, more data is clearly needed to see if these trends persist in time.

When we change the starting point of the sliding window, the date from which we measure the JSD between, the story changes. Consistently, **(a), (b), (c),** and **(d)** indicate that there is little variation in the top ranked topic, and several others, across time. However, we see different outcomes for other topics when we adjust our starting point. This begs the question *how long does a topic remain un-varied, and what is the threshold amount of entropy that indicates an important shift in conversation?*. More data is needed, but the results warrant further investigation of this question. At the very least, these plots allow us to see how the entropy "bandwidth" among distributions, changes given a starting point in time.



(a) **(b)**

(c) **(d)**

Figure 4: Jensen-Shannon Divergence measured between each of the rank ordered i_{th} topic word distributions, forward in consecutive time.

Fig. 4 shows JSD measured between consecutive per-topic distributions over words.

In this figure, we measure JSD between distributions that are 15 minutes apart **(a)**, 30 minutes apart **(b)**, 45 minutes apart **(c)**, and 60 minutes apart **(d)** respectively. Observing changes in Entropy from this perspective starts to get at the question posed in the previous paragraph. Here we see the JSD change in real-time, over different time-deltas. With few observations, it is hard to determine meaningful emergence from these plots. But given the outcome of **(a)**, and with more thorough investigation, we might like to look more closely at the distributions over words for the top topic between 2017-08-01-01-00 and 2017-08-01-01-15, based on the sharp decrease in JSD.

0.6 Future Work

In the future, we would like to use current infrastructure on the VACC to process more data; i.e. a significant number of corpora that that were published before and after the events of Charlottesville that may show a clear trend towards unified discussion on Twitter. Furthermore, we want to spend more time cleaning the data to prevent biasing the performance of the LDA models.

To remedy the fact that we cannot assess the accuracy of the LDA model without prior knowledge of the per-document topic distribution, we would like to explore adopting Bayesian Information Criterion as a step in our preprocessing procedure, in order to choose the likeliest number of components (topics) for the model to fit over a subsequent corpus. An adaptive parameter sweep may ensure more accurate results.

0.7 Resources

[1]

Bibliography

[1] sklearn.decomposition.LatentDirichletAllocation — scikit-learn 0.20.1 documentation.

[2] Jensen–Shannon divergence, December 2018. Page Version ID: 872234878.

[3] David M Blei. Latent Dirichlet Allocation. page 30.

[4] Jonathan Huang. Maximum Likelihood Estimation of Dirichlet Distribution Parameters. page 9.