Spatial Aggregation Facilitates Discovery of Spatial Topics

Aniruddha Maiti
Temple University
Philadelphia, PA-19122, USA
aniruddha.maiti@temple.edu

Slobodan Vucetic
Temple University
Philadelphia, PA-19122, USA
vucetic@temple.edu

Abstract

Spatial aggregation refers to merging of documents created at the same spatial location. We show that by spatial aggregation of a large collection of documents and applying a traditional topic discovery algorithm on the aggregated data we can efficiently discover spatially distinct topics. By looking at topic discovery through matrix factorization lenses we show that spatial aggregation allows low rank approximation of the original document-word matrix, in which spatially distinct topics are preserved and non-spatial topics are aggregated into a single topic. Our experiments on synthetic data confirm this observation. Our experiments on 4.7 million tweets collected during the Sandy Hurricane in 2012 show that spatial and temporal aggregation allows rapid discovery of relevant spatial and temporal topics during that period. Our work indicates that different forms of document aggregation might be effective in rapid discovery of various types of distinct topics from large collections of documents.

1 Introduction

Social microblogging sites such as Twitter generate large volumes of short documents through the activity of hundreds of millions of users around the world. This provides an unprecedented access to the pulse of the global society. Due to the sheer volume and diversity of the generated content, topic discovery has been an invaluable tool in an effort to make sense of this data. Regardless of a precise definition of a topic and a particular topic model, topics discovery is used to describe pertinent themes in a document corpus and serve to identify events, trends, and interests at the global, local, or a social group level.

Among the most popular topic modeling techniques are Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-negative Matrix factorization (NMF). When applying those techniques for topic discovery from microblogs, there are three main challenges: (1) how to improve computational speed, (2) how to extract useful topics, and (3) how to deal with short texts. Many papers were published that address one or more of these challenges and most of them propose to modify the original topic models.

In this paper, we are focusing on *aggregation* (also referred to as *pooling*) (Alvarez-Melis and Saveski, 2016) (Hong and Davison, 2010) (Weng et al., 2010) (Steinskog et al., 2017), a particular document preprocessing technique that has been empirically shown to be useful for topic discovery from microblogs. The main idea of aggregation is to combine multiple documents into a single document according to some external criterion and to apply a topic discovery algorithm on the aggregated documents. The earliest mentions of aggregation (Mehrotra et al., 2013) (Hong and Davison, 2010)(Weng et al., 2010) are motivated by the difficulty when applying NMF and LDA to very short text documents (Hong and Davison, 2010). This difficulty in finding useful topics is often attributed to the sparseness of the document-word matrix (Yan et al., 2013) (Cheng et al., 2014), which fails to provide confident counts of word co-occurrence and information about the shared context (Phan et al., 2008). Microblogs often come with metadata such as hashtags, author name, time stamp, or location. By aggregating the microblogs according to such metadata, the intuition is that the resulting aggregated documents contain a sufficient number of words for topic modeling schemes to identify meaningful topics. In addition, the authors of those early papers observe that aggregating microblogs that are similar in some sense (semantically, temporally) enriches the content present in a single document and results in better topics (Mehrotra et al., 2013). Finally, due

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 252–262
Florence, Italy, July 28 - August 2, 2019.

(a) Common Topic (b) Distinct Topic

Figure 1: Examples of common and distinct topics. a) Common topic: a work-related topic. b) Distinct temporal topic: presidential debate

to reduction in a number of documents, aggregation also leads to computational savings.

While aggregation has received interest in the research community and there are several empirical studies illustrating its benefits, we are not aware of a study that manages to provide, beyond brief intuitive arguments, an insight into why aggregation works and what are its advantages and limitations. In this paper we attempt to provide such an insight from the perspective of discovering spatially specific topics. As will be evident, our insights extend to other means of aggregation.

Our argument will be given in the context of matrix factorization, where a document-word matrix X is represented as a product $W \cdot H$, where j-th row of matrix H represents word distribution in j-th topic and i-th row of matrix W represents a distribution of topics in i-th document. We adopt the terminology from (Kim et al., 2015), which distinguishes between *common* and *distinct* topics (see Figure 1), where distribution of common topics within the corpus is not impacted by the aggregation metadata such as location, time, or author of a microblog, and distribution of distinct topics is correlated with the metadata. We show that factorization of the aggregated matrix X_a, obtained by merging documents based on metadata (e.g., location), allows its low rank approximation as $W_a \cdot H_a$, where the resulting topic matrix H_a retains the distinct topics from H (e.g., spatial topics) and where the common topics from H are merged into a single topic in H_a. We will show empirical results confirming this observation both on synthetic and real-life data. In particular, we will demonstrate this behavior in case of spatial and temporal aggregation.

The main contribution of this paper is in demonstrating that applying standard topic discovery algorithms such as NMF and LDA on aggregated documents results in discovery of topics related to the aggregation method. Moreover, since the aggregated matrix X_a can be orders of magnitude smaller than the original matrix X, the computational cost can also be reduced by orders of magnitude. Finally, as observed in the previous work, aggregation also alleviates the problem of sparsity when discovering topics in microblogs.

2 Related Work

Topic modeling from microblogs has a vast amount of literature (Steiger et al., 2015). Early work includes using NMF on term correlation matrix (Yan et al., 2013) and ncut-weighted NMF (Yan et al., 2012). Recent work includes NMijF (Nugroho et al., 2017), which takes into account tweet-to-tweet interactions. Location recommendation model based on topic modeling was proposed in (Hu et al., 2013). NMF is used in DiscNMF (Kim et al., 2015) and STExNMF (Shin et al., 2017) to identify spatio-temporal topics. Pairfac (Wen et al., 2016) employs tensor decomposition accounting for location, time, and venue. In TopicOnTiles (Choi et al., 2018), the entire space-time is divided into small tiles and NMF is performed on each tile separately. LDA (Blei et al., 2003) has also been used for topic detection. In (Zhao et al., 2011), LDA is used to categorize and summarize tweets. In (Weng et al., 2010), LDA is used to find influential users in Twitter.

Traditional topic modeling techniques such as LDA, LSA, and NMF are sensitive to sparsity (Hong and Davison, 2010). Different types of document aggregation schemes have been suggested to overcome this issue (Alvarez-Melis and Saveski, 2016). One example of an aggregation scheme is the author-topic model (Weng et al., 2010), in which multiple tweets from the same user are aggregated to construct documents representative of the user. In (Hong and Davison, 2010), it was observed that document aggregation endows the resulting dataset with interesting properties, where aggregation based on authors has been reported to produce topics which are different from topics discovered on non-aggregated dataset. User level aggregation was also found to be useful in related papers (Giorgi et al., 2018). Similar results were also observed for aggregation based on hashtags (Steinskog et al., 2017). These papers did not attempt to explain the mechanism behind changes in the discovered topics and this is where our current paper makes a contribution.

3 Methodology

3.1 Problem Setup

Let us assume we are given a corpus of documents $D = \{d_1, d_2, d_3, .., d_N\}$, where N is the total number of documents. Let V be the vocabulary of unique words in the corpus. By using the bag of words representation, the corpus can be represented by a document-word matrix X of dimension $N \times V$, where element $X_{i,j}$ is the count of j-th word in i-th document. We will also assume that each document d_i is associated with a time stamp $t(d_i) \in 1, ...T$, where T is the number of time steps, and location $l(d_i) \in 1, ...L$, where L is the number of locations.

We will make an assumption that there are K topics $t_1, ...t_K$, where topic t_k defines probability that word w_j will be generated by the topic as $p(w_j|t_k)$, and that each document in a corpus is represented by a single topic. Our simplifying assumption that each document is generated by a single topic is acceptable when dealing with short documents such as microblogs. In addition, it will make it easier to describe the main effect of document aggregation.

Among the K topics, we will assume that the first K_d are spatially distinct topics and the second K_c are common topics. For common topics, the probability or their occurrence does not depend on location of the document. In other words, $p(t_k|l) = p(t_k)$, where l is location. Conversely, for spatially distinct topics, the probability of their occurrence is dependent on the document location. We illustrate such a setup in Figure 2, where there are 4 spatially distinct topics generated within 4 different circular regions and 2 common topics occurring equally likely over the whole square region. In this example, the probability that a distinct topic is generated within its assigned circle is constant and is zero outside.

Given D, the objective is to find the distinct topics. In the following we will argue that document aggregation enables computationally efficient discovery of the distinct topics.

3.2 Effect of Spatial Aggregation on Rank

In this section we will explain why spatial aggregation of documents facilitates discovery of spatially distinct topics. If we select a subset X_k of all documents from X generated by topic t_k, the best rank-1 approximation of X_k is proportional to $n_k \cdot h_k$, where n_k is a column vector of length N whose i-th element is the sum of all words in i-th document and h_k is a row vector of length V whose j-th element h_{kj} equals $p(w_j|t_k)$. Let us denote this rank-1 approximation as ${X_k}^1$. If we sort the document-word matrix X by topics, we can approximate it by vertically concatenating rank-1 matrices ${X_k}^1$. The rank of the resulting matrix X^1 will be less than or equal to J.

We observe that the rank of matrix X can be as high as $V >> J$ and that matrix factorization of X into product $W \cdot H$ cannot guarantee successful topic discovery. On the other hand, we observe that factorization of X^1 can easily result in discovery of the underlying J topics. Unfortunately, generating matrix X^1 is as difficult as the topic discovery problem itself. We argue in the following that aggregation based on location results in generation of a matrix closely related to X^1. As such, we demonstrate that spatial aggregation is very useful for discovery of spatially distinct topics.

Let us define binary matrix Q with L rows and N columns as spatial aggregation matrix which merges the N original documents into L aggregated documents, where $Q_{l,i} = 1$ if document x_i belongs to l-th location and $Q_{l,i} = 0$ otherwise. We construct the aggregated document-word matrix of size $L \times V$ as $X^a = Q \cdot X$. The expected value of l-th row of matrix X^a equals:

$$E(X^a_l) = \sum\nolimits_k (n_{lk} \cdot h_k), \tag{1}$$

where, n_{lk} is a scalar equal to the number of words generated from topic t_k in documents from l-th location and h_k is a row vector defined in the first paragraph of this subsection. If the number of documents at l-th location is large, the observed X^a_l will be close to $E(X^a_l)$. Since based on equation (1) each row of X^a can be approximated as the linear combination of K topic vectors h_k, it follows that matrix X^a is approximately of rank K or less. We can thus closely approximate X^a as product $W^a \cdot H^a$, where k-th row of matrix H^a equals h_k and (l, k)-th element of matrix W^a equals n_{lk}.

We will now show that $W^a \cdot H^a$ has rank lower than K. Since the K_c common topics are assumed to be location independent, the number of documents generated by k-th common topic is approximately the same at every location. Thus, we can approximate $n_{lk} = n_k$ for each of the K_c common topics. Therefore, the last K_c columns of matrix W^a are constant. As a result, the rank of matrix

$W^a \cdot H^a$ is $K_d + 1$ or less, where the K_c common topics increase the rank by only one. As a result, we can replace the last K_c columns of W^a with a single column equal to the sum of the last K_c columns of W^a and replace the last K_c rows of H^a with a single row equal to the sum of the last K_c rows of H^a. The resulting topic matrix H^a is of dimension $(K_d + 1) \times V$, where the last row is a sum of word probabilities over all common topics, while the first K_d rows are reserved for each of the K_d spatially distinct topics. This is a significant result showing that spatial aggregation facilitates discovery of spatially distinct topics while it collapses all documents generated by the common topics into a matrix that can be closely approximated by a rank-1 matrix.

3.3 NMF and LDA on Aggregated Data

In the previous section we did not specify a particular algorithm for matrix factorization and topic discovery. NMF is a popular matrix factorization algorithm for nonnegative matrices such as document-word matrices. NMF finds nonnegative and sparse matrices W and H whose product approximates the original matrix. It solves the following optimization problem:

$$F(W,H) = \frac{1}{2}||X - W \cdot H||^2_{Fro} + \alpha \cdot \rho \cdot ||W||_1 + \alpha \cdot \rho \cdot ||H||_1 + \frac{1}{2}\alpha(1-\rho) \cdot ||W||^2_{Fro} + \frac{1}{2}\alpha(1-\rho) \cdot ||H||^2_{Fro}. \quad (2)$$

Here, the Frobenius norm of a matrix A is denoted by $||A||_{Fro}$ and α and ρ are regularization parameters. Rows of W of size $N \times K$ represent the topic mixture within a particular document where K is the number of topics. Rows of H of size $K \times V$ represent the word distribution within a particular topic. The NMF optimization problem is typically solved iteratively and the algorithm becomes expensive for large data sets. NMF is also sensitive on collections of short documents such as microblogs. NMF favors commonly occurring topics and commonly ocurring words, which makes finding rare spatially distinct topics very difficult. Document aggregation based on metadata such as location directly addresses the aforementioned NMF issues.

The arguments in the previous sections demonstrate the benefit of aggregation through matrix factorization. However, our assumptions made in 3.1 closely resemble the generating process used in LDA, where each document is a mixture over latent topics, and each topic is characterized by a distribution over words. From the corpus, LDA learns the topic distribution over documents and word distribution over topics. While, in theory, LDA should be able to discover topics directly from the original matrix X, it suffers from the same shortcomings as NMF: it is slow, fragile, and sensitive to sparse documents. As will be demonstrated in the experiments, document aggregation has very similar effects on both NMF and LDA.

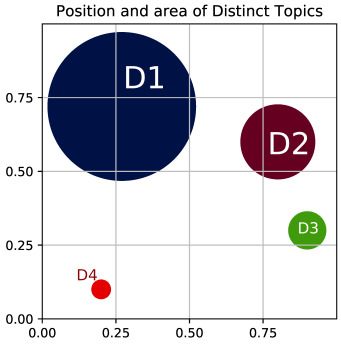

Figure 2: Spatially distinct topics on simulated data

To summarize, the resulting distinct topic discovery procedure has the following steps:

1. Construct document-word matrix X.
2. Construct spatial aggregation matrix Q from metadata.
3. Perform NMF on aggregated matrix $Q \cdot X$ to find spatially distinct topics.

If we wish to identify spatial-temporal topics, we may additionally aggregate the data based on time. First, the entire time span can be divided into smaller intervals. Then, all documents in each space-time cell are aggregated into a single document. Although we do not show it in our experiments, our major insight about the effect of document aggregation extends to other forms of aggregation such as author- or hashtag-based.

4 Experiments on Simulated Dataset

In this section, we use synthetic data to study the effect of document aggregation on topic discovery.

Following the setup provided in Section 3.1, we created a dataset using a simplistic generative model. Words in each document in the dataset are generated from two common topics ($C1$ and $C2$)

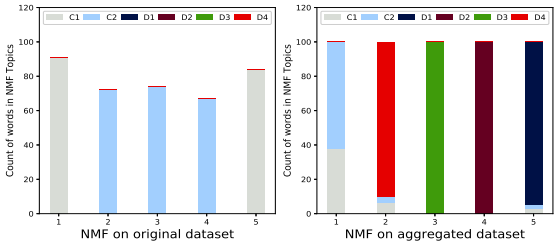

Figure 3: Five topics discovered by NMF on non-aggregated and aggregated documents

and four spatially distinct topics ($D1$, $D2$, $D3$ and $D4$). Each common and distinct topic uses a vocabulary with 100 words. Each document is associated with a single topic. To generate a document, a topic is chosen first, then 10 words are sampled randomly from the 100 words associated with that particular topic. Documents generated from the common topics are distributed randomly within the square. For each distinct topic, a circular region is defined within the square and the documents associated with that topic are placed by uniformly sampling within the circle. The placement of the circular regions is shown in Figure-2. A total of $10,000$ documents are generated for each common topic and $1,000$ documents for each spatially distinct topic. We call this dataset the *non-aggregated dataset*. To demonstrate how aggregation affects the topic discovery, we divided the entire region in 4×4 small squares. Then we merged all the documents from each small square into a single aggregated document. In this way, we constructed 16 aggregated documents. We call this dataset the *aggregated dataset*.

NMF set to find 5 topics was applied to the non-aggregated and the aggregated datasets. In Figure 3, we show the distribution of words in each of the 5 identified topic. For example, the first bin in the left subplot shows that discovered topic 1 has 91 unique words, all belonging to common topic $C1$. On the other hand, the first bin in the right subplot shows that discovered topic 1 has 100 unique words, 38 belonging to common topic $C1$ and 58 to common topic $C2$. We can see that none of the spatially distinct topics are discovered when we apply NMF on the non-aggregated data. All five identified topics contain words from the 2 common topics. On the other hand, in the aggregated dataset, the first identified topic contains a mixture of words from the 2 common topics, while the remaining 4 are almost entirely comprised of words from the 4 spatially distinct topics. This result experimentally supports our insight about the impact of spatial aggregation presented in section 3.2.

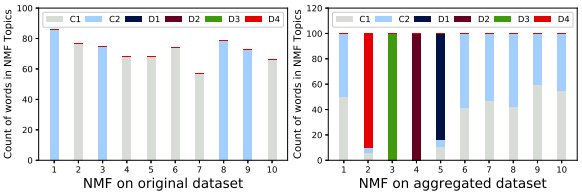

Figure 4: Ten topics identified by NMF on non-aggregated and aggregated documents

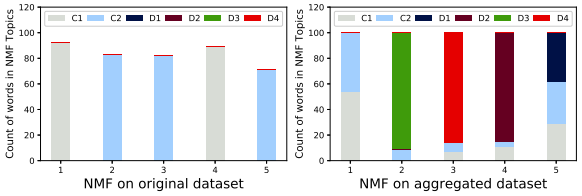

Figure 5: Five topics identified by NMF on original and aggregated data using a smaller set of documents

4.1 Effect of Number of Topics in NMF

We repeated the NMF experiment, but this time we set the number of NMF topics to 10. We can see from Figure 4 that all 10 topics found on the non-aggregated data are still one of the two common topics. On the other hand, after applying NMF on the aggregated data, 4 of the discovered topics directly correspond to the 4 spatially distinct topics, while the remaining 6 discovered topics are a mixture of the 2 common topics.

4.2 Effect of Number of Documents

We repeated the experiments on a smaller corpus to see its effect on topic discovery. We generated $1,000$ documents for each common topic and 150 documents for each distinct topic. The result is summarized in Figure 5. As compared to Figure 3, we can see a slight deterioration of the quality of discovered spatially distinct topics from the aggregated data. In particular, all of the 4 discovered spatial topics are corrupted with more words from the common topics, which is particularly visible from the rightmost bin containing and an almost equal mixture of words from topics $D1$, $C1$, and $C2$. We observe that topic $D1$ corresponds to the largest circle.

4.3 Effect of Grid Density

We repeated the previous experiment on the smaller dataset with $1,000$ documents for each

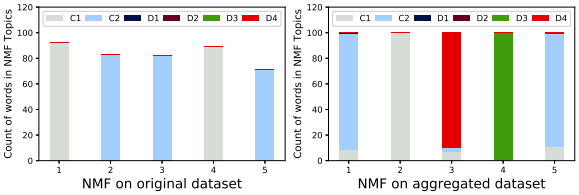

Figure 6: Five topics identified by NMF using dense spatial grid 64×64

common topic and 150 documents for each spatially distinct topic, but this time with gradually increasing aggregation density. In Figure 6 we show results of applying NMF set to discover 5 topics for the spatial aggregation scheme with a grid size 64×64. As expected, the results look more similar to topic discovery from the non-aggregated dataset. Interestingly, despite the vary coarse aggregation (many spatial blocks were empty or with a single document), we still discovered topics $D3$ and $D4$, which correspond to the smaller circles.

5 Experiments on Real Life Data

Identifying spatially distinct topics in a real life dataset is a challenging task. As we will demonstrate, we found that the aggregation scheme is quite successful in identifying distinct topics. We performed our experiments on Hurricane Sandy Twitter corpus downloaded through Twitter search API[1] using the tweet IDs released in (Wang et al., 2015). The downloaded corpus contains 4.7 million tweets that temporally span 12 days surrounding the Hurricane Sandy and a few other distinguishable events between October 22nd, 2012 and November 2nd, 2012. Every tweet in the dataset is also geotagged to one of 13 states along the East Coast of the U.S. During preprocessing we transformed all characters to lowercase and removed stopwords and special characters. We also excluded repetitive letters that convey enthusiasm (e.g., birthdayy, birthdayyy, birthdayyyy). Finally, TF-IDF document-word matrix is constructed using the $20,000$ most frequent words in the corpus.

Since the spatial distribution of tweets is highly imbalanced, we decided not to use a regular spatial grid. Instead, we employed k-means clustering on the latitude and longitude information for each tweet to identify 200 cluster centers in space. Each tweet is assigned to its nearest cluster center for spatial aggregation. Figure 7 shows different clusters on $50,000$ tweets randomly sampled

[1] https://developer.twitter.com/

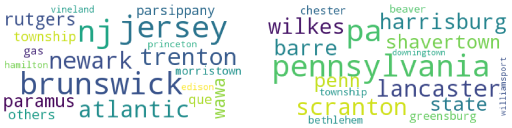

Figure 8: State specific distinct topics

from the corpus. We can observe that the density of clusters is much larger within heavily populated urban areas along the East Coast.

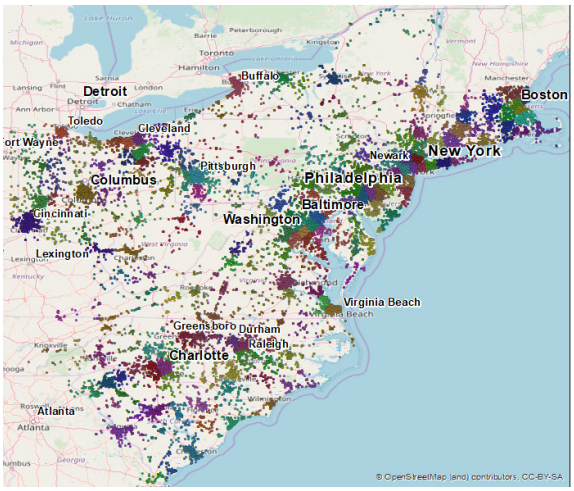

Figure 7: K-means cluster for spatial aggregation

NMF was employed to find 500 topics with $\alpha = 0.1$ and $\rho = 0.5$. Only 107 rows of H were found to have at least one nonzero entry. Application of NMF on the 200 aggregated documents identifies some spatially distinct topics covering regions of varying size. Figure 8, shows word clouds for two large state-specific distinct topics. We also found that large metropolitan areas such as New York City, Philadelphia, and Pittsburgh are represented as separate spatially distinct topics. One such example is shown in Figure 9. Almost all the words in this topic are related to New York City airports.

In addition to spatial aggregation, we also performed experiments by aggregating data in space and time. In addition to the $k = 200$ spatial clusters we divided the time interval into 12 days, resulting in a total of $2,400$ spatio-temporally aggregated documents. As expected, this aggregation reveals distinct spatio-temporal topics.

We identified several purely temporal topics in this way, including the Halloween topic shown in Figure 10. It is interesting to observe that this topic also contains words related to the season opening

Figure 9: NYC airport-specific distinct topic

Figure 10: Halloween and CMA temporal topics

NBA game between L.A. Lakers and Miami Heat that occurred on the same day. Figure 10 also contains another temporally distinct topic associated with the 2012 Country Music Association (CMA) Award event that happened on the same day.

To better illustrate this CMA-related topic, in Table 1 we show several representative tweets. These tweets were randomly selected from tweets containing at least one of the most frequent 10 words in the CMA-related topic.

5.1 Evaluation: Space-Time Scan Statistics

Looking at word clouds is a descriptive way to evaluate the quality of discovered topics. In this subsection we will present experimental results attempting to quantitatively evaluate the quality of the discovered topics. To achieve this we use the space-time scan statistics implemented in the SaTScan software (Kulldorff, 2010). We selected the 10 most frequent words in each discovered topic and labeled each tweet from the corpus based on the presence of these words. If a tweet contains any of the 10 words it is assigned to the corresponding topic. We call all tweets assigned to the given topic the positive tweets. If the topic is strongly spatial, we would expect the assigned tweets to be strongly spatially clustered. If the topic is strongly spatio-temporal, we would expect the assigned tweets to cluster within a particular spatio-temporal area. The space-time scan statistic is employed to measure enrichment by positive tweets of cylindrical windows covering a circular spatial region and a temporal interval. The cylindrical window is moved in space and time to search for the statistically strongest clusters (Kulldorff, 2010). The cylinder with the strongest enrichment of positive tweets (e.g., based on the ratio between positive tweets and all tweets within the cylinder) is a potential candidate for the significant spatio-temporal cluster.

Table 1: Tweets related to CMA awards

Anyone know what channel the cma is on?
Can't wait for the cma awards
Everyone get prepared for a bunch of cma awards tweets
Tomorrow is 46 cma awards so watching that!!
carrie underwood is amazing
Hunter hayes is perfect
Not sure why Taylor Swift is taking over the country charts...her music is more of a mix now between country and pop
Luke bryan on the CMAS omg omg !!!

Distributional properties of scan statistics can be used to evaluate the statistical significance of the strongest cylinder (Dwass, 1957). This is done by permuting the labels of tweets multiple times (999 times in this study) and calculating the score of the strongest cylinder in each permutation (Block, 2007). The p-value is then calculated by counting the fraction of the permuted scores larger than the score on the actual data. The p-value reported in this experiment can be thought of as a measure of the spatio-temporal distinctiveness of the identified topic.

Characterization of distinct topics using p-value has some limitations. We observed that many distinct topics discovered through aggregation receive p-value equal to zero, making it impossible to identify the strongest distinct topic. For this reason, we used *deviation* (Δ), which measures how many standard deviations apart is the score of the best cylinder observed on the actual data compared to the scores of the best cylinders observed on the permuted data.

Table 2: Evaluation of the topic quality using SaTScan

Topic General Theme	Deviation (Δ)	Topic Type
Power	26504.53	Temporal
NYC	25282.17	Spatial
NFL	12275.18	Temporal
Presidential Debate*	11089.34	Temporal
Snow	8624.95	Temporal
New Jersey*	8355.10	Spatial
Halloween*	7679.58	Temporal
Pennsylvania*	6728.94	Spatial
NYC Airport*	6424.54	Spatial
Weather	2220.64	Temporal

In Table 2, we show the strongest topics based on the deviation (Δ). In each case, the p-value was 0. For topics labeled with stars in Table 2,

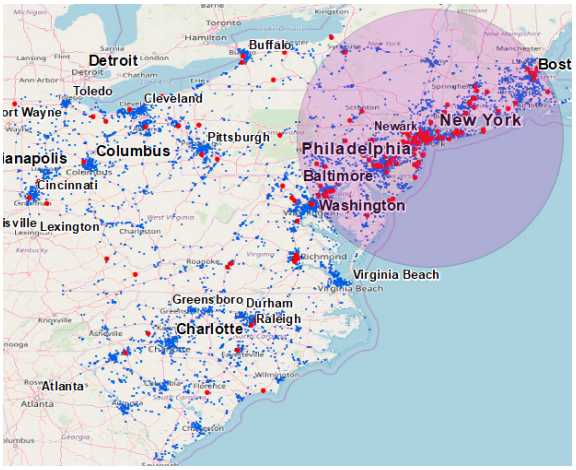

Figure 11: Positive (red) and negative (blue) examples and the position of the cluster identified by SaTScan for topic : Power Outage

the corresponding word clouds were shown in Figures 1, 8, 9, 10. For the remaining topics, the top ten words are presented in Table 3. It may be noted that New York City, being a very large metropolitan area, has multiple identified topics. One such topic, called *NYC airport*, was previously presented in Figure 9. Another such topic, called *NYC*, is presented in Table 2. The spatio-temporal region called *Power outage* is shown in Figure 11. 10,000 tweets in this figure are labeled as positive or negative based on the presence or absence of the keywords of this topic. This topic corresponds to multiple power outages in the aftermath of the Sandy Hurricane.

Table 3: General theme of topics and related words

Topics	Words
Power	power sandy generator trees electricity tree open lights safe hurricane
NYC	york brooklyn nyc park manhattan city square mta island halloween
NFL	cowboys steelers romo giants harden church redskins touchdown eagles party
Snow	snow snowing cold weather delay boone wind blizzard snowed outside
Weather	barometer humidity temperature mph wind rain blacksburg steady wnw rising

† Offensive words are removed

5.2 Comparison between LDA and LSA

Previous studies indicate that NMF on Twitter data works better than other available topic modeling techniques (Klinczak and Kaestner, 2015), (Godfrey et al., 2014). This may be attributed to a slightly better robustness of NMF to the short document lengths. This problem is ameliorated in this

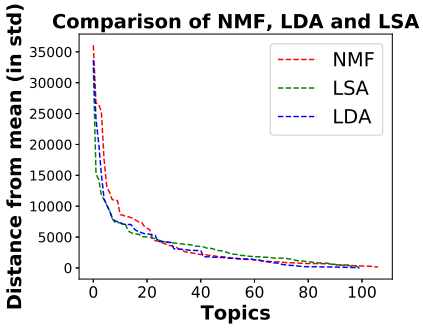

Figure 12: Comparison of NMF, LDA and LSA

study through aggregation. In view of this, it is expected that other topic modeling approaches are also able to identify distinct topics in the aggregated data. To verify this, we tried two other popular algorithms, LSA and LDA.[2] LSA is a truncated singular value decomposition technique. LDA is a generative probabilistic model. For LDA and LSA, the number of topics are taken to be 100 to be comparable to the number of topics identified by NMF. Document topic prior and topic word priors in LDA were set to 0.01.

We found that LDA and LSA identify distinct topics comparable to NMF when applied to the spatio-temporally aggregated data. Some of the similar topics are selected manually from the LDA and NMF topic lists and shown in Table 4 for comparison.

Table 4: Identified LDA topics similar to NMF

Topics	Words
NYC	york park brooklyn city pic nyc halloween st th center square street bar
Power	power sandy hurricane storm safe phone wind stay rain closed open
Weather	wind mb mph rain humidity cb barometer temp slowly cam midnight falling relative
Presidential Debate	romney obama debate class mitt president world vote talking week policy

† Offensive words are removed

It is difficult to draw one-to-one correspondence among all the topics identified by the three methods. We see from Table 4 that some topics are very similar in both NMF and LDA. However, while NMF discovers a topic related to the CMA, LDA and LSA do not. For this reason, instead of comparing the corresponding topics one at a time, the following strategy is applied. The topics in the

[2]python scikit-learn package is used for all three methods

three methods are first sorted based on the *deviation* (Δ) scores and plotted in Figure 12. The most significant topics identified by all three algorithms exhibit similar scores. For the top 20 topics, performance of NMF is only slightly better. The average score of the top 20 topics for NMF is 3,823, while the average scores for LSA and LDA are 3,638 and 3,390 respectively.

5.2.1 Common Topics in LDA and NMF

In Section 5.2, we mentioned that topic discovery algorithms such as LDA, NMF, and LSA are capable of finding distinct topics from aggregated documents. When non-aggregated data is used, these algorithms find common topics associated with day to day conversations. In Table 5, words associated with several common topics identified by LDA and NMF on a sample of the non-aggregated tweets are shown. It can be seen that the words in the identified topics do not correspond to a specific space or time.

Table 5: Common topics from non-aggregated data

LDA	NMF
cold shot dry blessed smoking wonderful	cold weather hot room hungry feet world
making sounds coffee running	fun sounds making lot times safe games looks
talking saw anymore west facebook	twitter goodmorning jail facebook instagram
guy past means throw start	guys girl safe play awesome stay

† Offensive words, informal words and internet short form of the words are removed

5.2.2 Influence of Aggregation Strategies and Randomization

Our experiments with the simulated data in Section 4.3 revealed that topic discovery is impacted by the aggregation grid density. To see if the behavior transfers to Twitter data, we varied the number of clusters from 100 and $1,000$. As the number of clusters increased, we observed that some of the distinct topics discovered by NMF for $k = 200$ disappeared when k was increased to 500 or $1,000$. For example, the CMA topic disappeared with those larger numbers of clusters. We also observed relatively small changes in discovered topics for different runs of the clustering for the same value of k. We conclude that clustering used for aggregation has a modest impact on topic discovery.

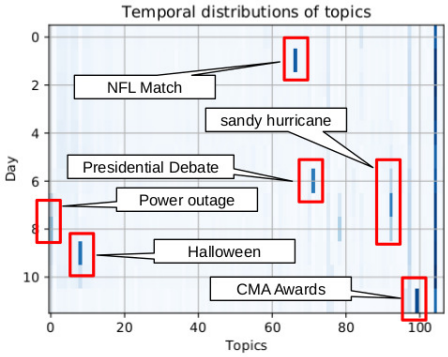

Figure 13: Visualization of temporal trends of topics

5.3 Temporal Trends in Topics

SaTScan reports the significant space-time cylinders for each topic. It is possible to categorize those cylinders as spatial or temporal by inspecting the their size. As an alternative, we can use matrix W obtained by NMF to identify temporal clusters. Let W^* be the matrix which is constructed from W by summing all the rows corresponding to the same time interval. W^* then represents a purely temporal description of topic distribution. By inspecting the columns of W^*, shown in Figure 13, we can obtain an additional insight into the nature of temporal topics. We can observe that only a small fraction of the identified topics are strongly temporal in nature.

6 Conclusion

In this work, we showed that spatial aggregation of documents leads to discovery of spatially distinct topics. We performed an extensive study on synthetic and real data and demonstrated that spatial and spatio-temporal aggregation indeed leads to discovery of spatial and spatio-temporal distinct topics. To evaluate the quality of the discovered topics we proposed a metric based on space-time scan statistics. Our results show that aggregation is a very powerful and computationally efficient method for discovery of distinct topics. While our study focused on spatial aggregation, aggregation on other types of metadata such as authors, hashtags, or communities is expected to work equally well and discover other types of distinct topics from large collections of documents.

References

David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. *ICWSM*, 2016:519–522.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Richard Block. 2007. Scanning for clusters in space and time:: A tutorial review of satscan. *Social Science Computer Review*.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Minsuk Choi, Sungbok Shin, Jinho Choi, Scott Langevin, Christopher Bethune, Philippe Horne, Nathan Kronenfeld, Ramakrishnan Kannan, Barry Drake, Haesun Park, et al. 2018. Topicontiles: Tile-based spatio-temporal event analytics via exclusive topic modeling on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 583. ACM.

Meyer Dwass. 1957. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, pages 181–187.

Salvatore Giorgi, Daniel Preotiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle H Ungar, and H Andrew Schwartz. 2018. The remarkable benefit of user-level aggregation for lexical-based population-level predictions. *arXiv preprint arXiv:1808.09600*.

Daniel Godfrey, Caley Johns, Carl Meyer, Shaina Race, and Carol Sadek. 2014. A case study in text mining: Interpreting twitter data from world cup tweets. *arXiv preprint arXiv:1408.5427*.

Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. ACM.

Bo Hu, Mohsen Jamali, and Martin Ester. 2013. Spatio-temporal topic modeling in mobile social media for location recommendation. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1073–1078. IEEE.

Hannah Kim, Jaegul Choo, Jingu Kim, Chandan K Reddy, and Haesun Park. 2015. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 567–576. ACM.

Marjori NM Klinczak and Celso AA Kaestner. 2015. A study on topics identification on twitter using clustering algorithms. In *Computational Intelligence (LA-CCI), 2015 Latin America Congress on*, pages 1–6. IEEE.

M Kulldorff. 2010. Satscan user guide for version 9.0. *Department of Ambulatory Care and Prevention, Harvard Medical School, Boston, MA*.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892. ACM.

Robertus Nugroho, Weiliang Zhao, Jian Yang, Cecile Paris, and Surya Nepal. 2017. Using time-sensitive interactions to improve topic derivation in twitter. *World Wide Web*, 20(1):61–87.

Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100. ACM.

Dear Sungbok Shin, Minsuk Choi, Jinho Choi, Scott Langevin, Christopher Bethune, Philippe Horne, Nathan Kronenfeld, Ramakrishnan Kannan, Barry Drake, Haesun Park, et al. 2017. Stexnmf: Spatio-temporally exclusive topic discovery for anomalous event detection. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 435–444. IEEE.

Enrico Steiger, Joao Porto De Albuquerque, and Alexander Zipf. 2015. An advanced systematic literature review on spatiotemporal analyses of twitter data. *Transactions in GIS*, 19(6):809–834.

Asbjørn Steinskog, Jonas Therkelsen, and Björn Gambäck. 2017. Twitter topic modeling by tweet aggregation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 77–86.

Haoyu Wang, Eduard H Hovy, and Mark Dredze. 2015. The hurricane sandy twitter corpus. In *AAAI Workshop: WWW and Public Health Intelligence*.

Xidao Wen, Yu-Ru Lin, and Konstantinos Pelechrinis. 2016. Pairfac: Event analytics through discriminant tensor factorization. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 519–528. ACM.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM.

Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xue-qi Cheng, and Yanfeng Wang. 2012. Clustering short text using ncut-weighted non-negative matrix factorization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2259–2262. ACM.

Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang. 2013. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 749–757. SIAM.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*, pages 338–349. Springer.