Topic Extraction Analysis of Tweets About NYC Neighborhoods Using Latent Dirichlet Allocation

C. Gray, The Pennsylvania State University, *Eberly College of Science*

Abstract– The purpose of this project is to explore topic modeling through tweets related to the opinions of New York City neighborhoods. The tweet data used was retrieved from Twitter API using snscrape (SNS), a scraper for social networking services [5]. The primary method being deployed in this research is Latent Dirichlet Allocation (LDA). LDA is one of the most popular topic modeling methods in which it assigns documents in the data to a specific topic [7]. The results of the model will be analyzed using word clouds, an intertopic distance map, and term relevance.

Keywords: crime, neighborhood, quality, community, topic modeling, LDA

INTRODUCTION

In real estate markets around the world, there are numerous factors that influence the price of properties. The most obvious factor is the house quality and attributes. Things such as number of bedrooms and bathrooms, square footage, age, and whether it has a basement or garage can greatly affect the value of a property. Another factor that has a huge impact on the value of real estate is the quality of the area. Is there a high availability of jobs in the area? Is commercial real estate highly developed or is it an up-and-coming area? Does the area have adequate resources for children such as good schooling, playgrounds, after school programs, etc.? The answers to these questions will determine how much a consumer is willing to pay for residential property in any particular neighborhood.

Although real estate markets around the world are highly dependent on facts and statistics, this industry can be subjectively influenced as well. The way in which a particular neighborhood or city is perceived by the public can have great influence on its public attraction and in turn, its monetary value. For example, Dubai, United Arab Emirates is thought to be one of the most luxurious and modern cities in existence. Because of this, coupled with its actual luxurious and modern nature, the city of Dubai is a highly sought experience with an extremely lucrative real estate market. On the other hand, Philadelphia, Pennsylvania is a beautiful, historic city; home to illustrious educational institutions including Drexel University, University of Pennsylvania, Temple University, etc. However, the quality of the northern areas of the city paired with the negative perception of the city makes Philadelphia undesirable.

In order to be successful in their practice, real estate professionals should be aware of the public perceptions of their respective work area. This paper proposes using tweet data to extract public perception of New York City neighbors using Latent Dirichlet Allocation (LDA) topic modeling. Text data was gathered from the Twitter API using a collection of keywords related to the research topic. A series of visualizations will be explored to analyze the topic extraction generated by the presented LDA model. The following sections will discuss the data explored, methods deployed, and results yielded.

METHODOLOGY

DATASET

As mentioned above, the data set utilized in this research consists of real tweet data from users located in NYC. The data was gathered using snscrape (SNS), a scraper for social networking services [5]. About one thousand tweets including any number of specified keywords were collected. The keywords given to the scraper include "crime" OR "neighborhood" AND "safety" OR "police" OR "dangerous" OR "quality" OR "community" OR "schools" OR "city" OR "housing". The first set of data gathered did not require any combination of words to be included in the tweets, ie. the word “AND” was not used. This command resulted in too many unrelated tweets so the second set of data gathered required the tweets to include both “crime” AND “neighborhood”. After browsing through the second set of tweets, I found that this data was highly biased given the negative connotation of the word “crime”. The final data set was gathered using the combination of keywords specified above.

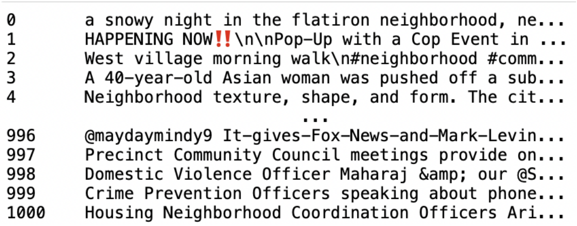
The dataset consisted of one thousand rows and seven columns. The features include Datetime, Text, Username, UserLocation, LikeCount, ReplyCount, and RetweetCount. The tweets collected were posted between May 2019 and April 2022. All columns except Text and Datetime were ignored as they are irrelevant to the experimentation.

EXPERIMENTAL SETUP

The topic modeling experiment presented was conducted using Python 3.8.8 via Jupyter Notebook [4].

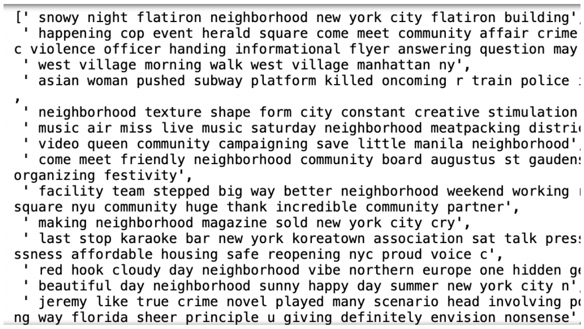
METHODS

When applying machine learning methods to text data, the first step is to perform natural language processing (NLP). NLP is the process by which computers translate text data to numerical data to allow for analysis. This research utilizes the Natural Language ToolKit (NLTK) module in Python [9]. This module is responsible for performing any tokenization, stemming, and removal of stop words in the data. Figure 1 and Figure 2 below display the text data before and after undergoing NLP, respectively. After performing NLP on the data, the tweets are free of username tags, hashtags, emoticons, punctuation, articles, and all non-alphanumeric characters. All characters have also been converted to lowercase. Finally, the keywords used to scrape the tweets were removed to reduce bias in the topic modeling.



```
0       a snowy night in the flatiron neighborhood, ne...
1       HAPPENING NOW‼️\n\nPop-Up with a Cop Event in ...
2       West village morning walk\n#neighborhood #comm...
3       A 40-year-old Asian woman was pushed off a sub...
4       Neighborhood texture, shape, and form. The cit...
                              ...
996     @maydaymindy9 It-gives-Fox-News-and-Mark-Levin...
997     Precinct Community Council meetings provide on...
998     Domestic Violence Officer Maharaj &amp; our @S...
999     Crime Prevention Officers speaking about phone...
1000    Housing Neighborhood Coordination Officers Ari...
```

Figure 1: Tweets Before NLP



```
[' snowy night flatiron neighborhood new york city flatiron building'
 ' happening cop event herald square come meet community affair crime
c violence officer handing informational flyer answering question may
 ' west village morning walk west village manhattan ny',
 ' asian woman pushed subway platform killed oncoming r train police
,
 ' neighborhood texture shape form city constant creative stimulation
 ' music air miss live music saturday neighborhood meatpacking distri
 ' video queen community campaigning save little manila neighborhood'
 ' come meet friendly neighborhood community board augustus st gauden
organizing festivity',
 ' facility team stepped big way better neighborhood weekend working
square nyu community huge thank incredible community partner',
 ' making neighborhood magazine sold new york city cry',
 ' last stop karaoke bar new york koreatown association sat talk pres
ssness affordable housing safe reopening nyc proud voice c',
 ' red hook cloudy day neighborhood vibe northern europe one hidden g
 ' beautiful day neighborhood sunny happy day summer new york city n'
 ' jeremy like true crime novel played many scenario head involving p
ng way florida sheer principle u giving definitely envision nonsense'
```

Figure 2: Tweets After NLP

The method proposed in this work is Latent Dirichlet Allocation (LDA), a popular machine learning model used to identify topics within text documents. LDA is an unsupervised learning model that analyzes text as a bag of words. Let us consider each row of data to be a document; the LDA model will handle each document as a bag of words and identify k topics in the data based on their term frequencies relative to each document. The model will then determine which topic each document corresponds to [6].

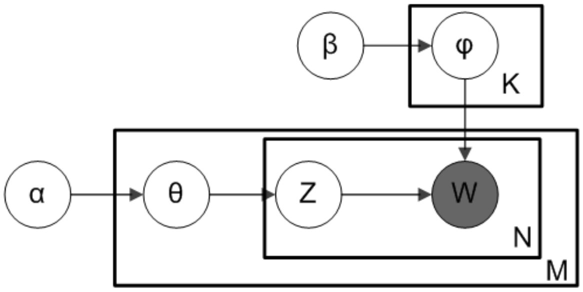


Figure 3: Plate Notation for LDA [6]

$$\begin{cases} \boldsymbol{\theta}_d \sim \mathrm{Dirichlet}_K(\boldsymbol{\alpha}) & d \in [M] \\ \boldsymbol{\phi}_k \sim \mathrm{Dirichlet}_V(\boldsymbol{\beta}) & k \in [K] \end{cases}$$

Figure 4: Dirichlet Priors [7]

Taking a look at Figure 3, we can observe the plate notation for a general LDA model. Here, *M* represents the total number of *i* documents while N signifies the total number of *j* words in a document *i*. The variables α and β represent the parameters of the Dirichlet priors on the per-document topic and per-topic word distributions, respectively. Finally, φ represents the word distribution of topic k, θ represents the topic distribution for document *i*, and *Z* represents the corresponding topic to the *j*th word in document *i*. W is the only variable that can be observed which are the words , or in this case, the bag of words.

EXPERIMENTATION

In order for the text data to be understood by the LDA model, it must first be converted into a bag of words. The next step was carried out using the scikit-learn module (sklearn) of Python [10]. The CountVectorizer function from this module was used to convert each document of text in the data into a matrix of token counts. Now that the data has been transformed into a bag of words, we can now initialize our LDA model using the gensim module [2] . The default α and β parameters (1/n_topics) have been used to initialize this model because there is not prior information about the per-document topic and per-topic word distributions. The results from the experimentation are discussed in the following sections.

RESULTS

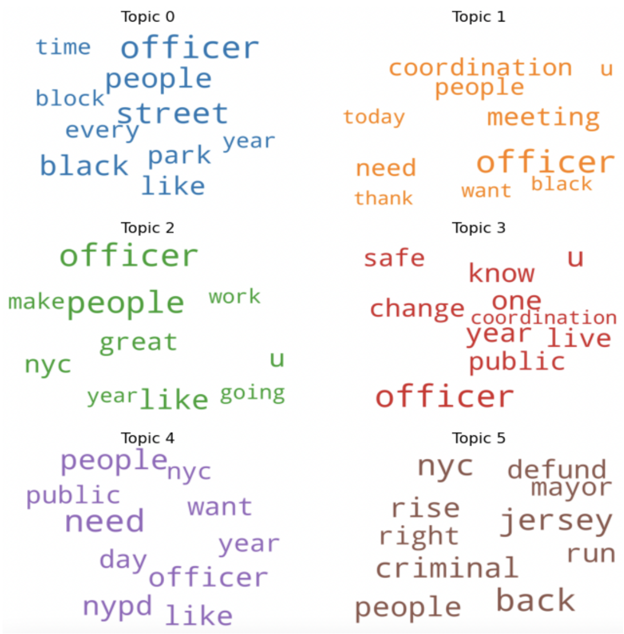


Figure 5: WordCloud of Frequent Words per Topic

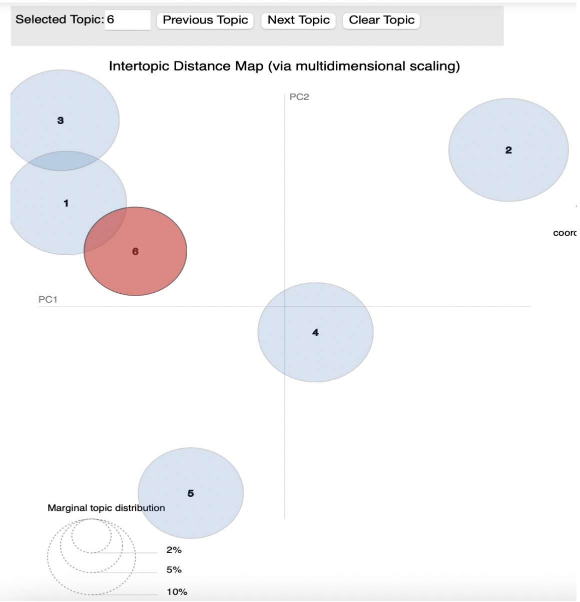
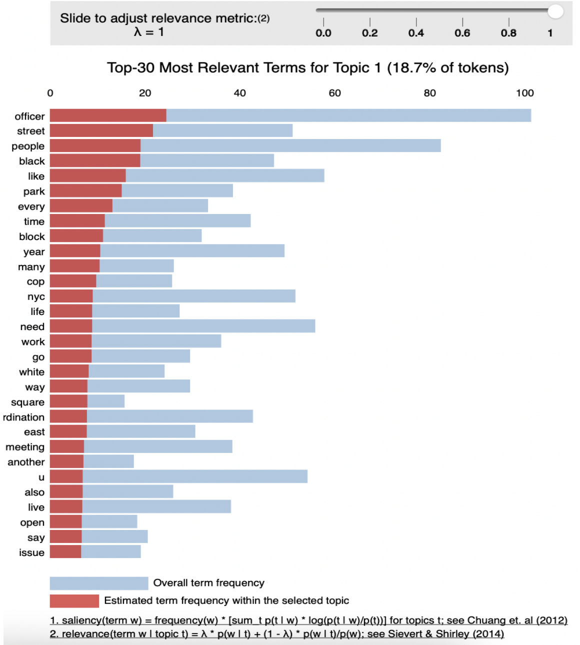

Selected Topic: 6
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
2
1
6
4
5
coorc
Marginal topic distribution
2%
5%
10%

Figure 6: Intertopic Distance Map



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)

Figure 7: Topic 1 Relevant Word Percentages

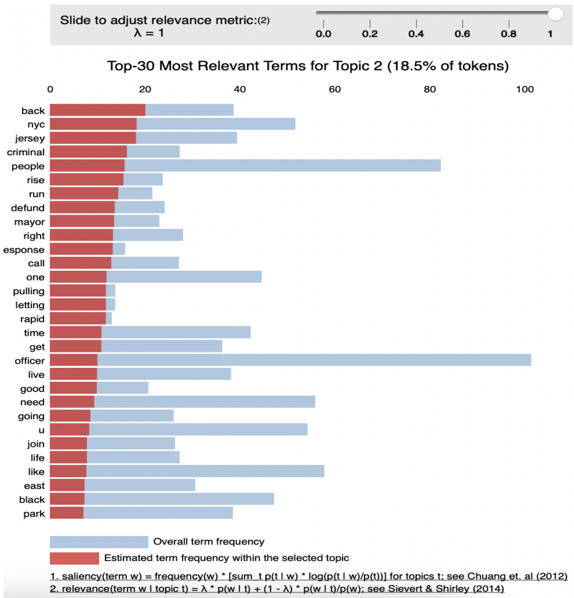


Figure 8:Topic 2 Relevant Word Percentages

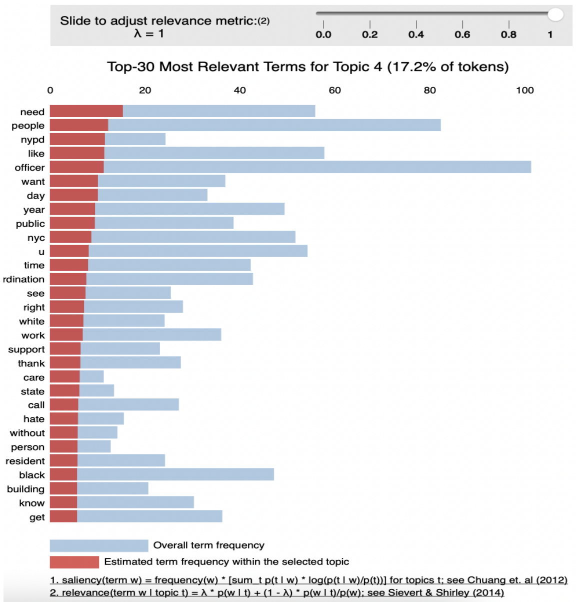


Figure 10: Topic 4 Relevant Word Percentages

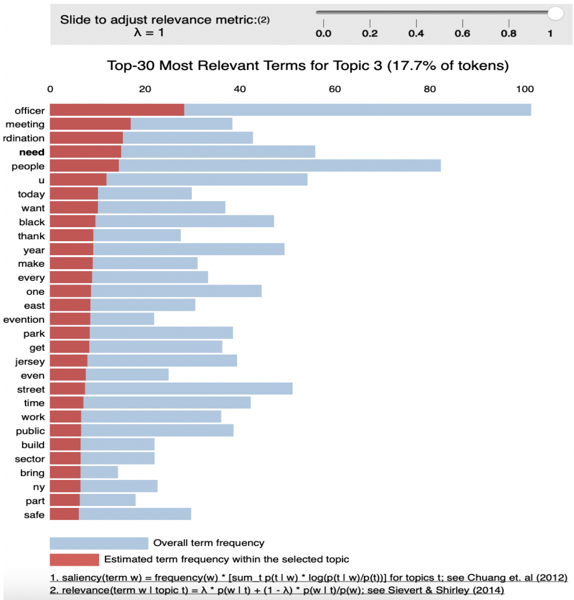


Figure 9: Topic 3 Relevant Word Percentages

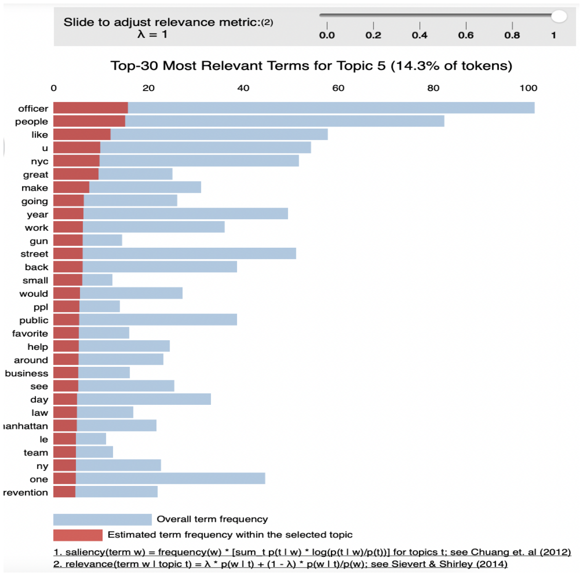
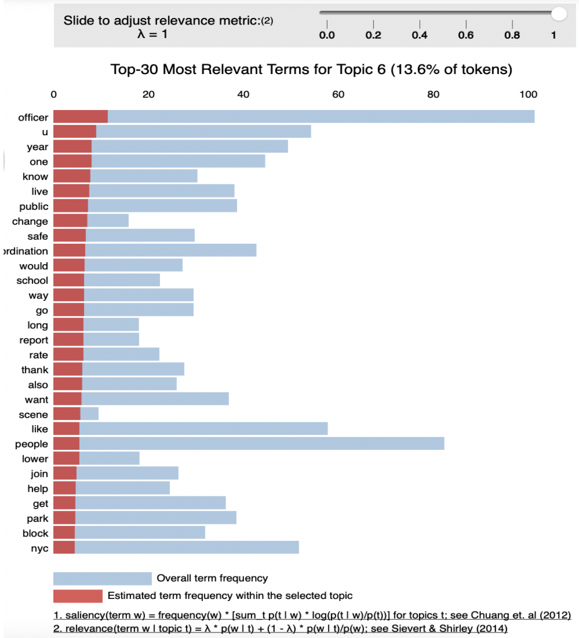


Figure 11: Topic 5 Relevant Word Percentages



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)

Figure 12: Topic 6 Relevant Word Percentages

DISCUSSION

The figures depicted in the results section above show us frequent words in each topic, as well as each topic's similarity to each other. *Figure 5* displays a word cloud of the six topics identified. The word clouds are labeled topic {0,5} but should be regarded as topic {1,6}.Each word cloud is color-coded and includes ten of the most frequently seen words in each topic. The larger the word, the more frequent it shows up in that topic. We can observe that "officer" is one of the common words amongst all topics besides topic 6. This observation is supported by *Figure 7*, *9*, *11*, and *12*.

Figure 6 displays the intertopic distance map for the model. Each bubble represents a topic; the bigger a bubble is, the more documents there are in the data that correspond to that topic. By this logic, we can identify topics 1, 2, and 3 as the most common topics throughout the data. The distance between each bubble represents the similarity between each topic. Since the bubbles for topic 1, 3, and 6 are fairly close to each other, we can assume that those three topics are very similar. On the other hand, topic 2 appears to be very different from topic 3.

Overall, the topics identified by the model provided valuable insight into what people are tweeting about their neighborhood. It would be very difficult to summarize each topic solely based on the result depicted above. To further this research, it would be wise to apply sentiment analysis methods to extract sentiment of each topic.

References

[1] Doll, Tyler. “LDA Topic Modeling.” Medium, March 11, 2019. https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd.

[2] “Gensim: Topic Modelling for Humans.” *Radim Å˜EhÅ¯Å™Ek: Machine Learning Consulting*, 2 May 2022, https://radimrehurek.com/gensim/.

[3] Hu, Yingjie, Chengbin Deng, and Zhou Zhou. “A Semantic and Sentiment Analysis on Online Neighborhood Reviews for Understanding the Perceptions of People toward Their Living Environments.” *Annals of the American Association of Geographers* 109, no. 4 (July 4, 2019): 1052–73. https://doi.org/10.1080/24694452.2018.1535886.

[4] “IPython — Jupyter Documentation 4.1.1 Alpha Documentation.” Accessed May 2, 2022. https://docs.jupyter.org/en/latest/reference/ipython.html.

[5] JustAnotherArchivist. *Snscrape*. Python, 2022. https://github.com/JustAnotherArchivist/snscrape.

[6] “Latent Dirichlet Allocation.” In *Wikipedia*, April 5, 2022. https://en.wikipedia.org/w/index.php?title=Latent_Dirichlet_allocation&oldid=1081147949.

[7] “Lecture - OneDrive.” Accessed May 2, 2022. https://pennstateoffice365-my.sharepoint.com/personal/xkz5224_psu_edu/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fxkz5224%5Fpsu%5Fedu%2FDocuments%2Fstat440%5Fsp22%2Flecture%2Flatent%5Fdirichlet%5Fallocation%5Flecture%2Epdf&parent=%2Fpersonal%2Fxkz5224%5Fpsu%5Fedu%2FDocuments%2Fstat440%5Fsp22%2Flecture&ga=1.

[8] Machine Learning Plus. “Topic Modeling Visualization - How to Present Results of LDA Model? | ML+,” December 4, 2018. https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/.

[9] “NLTK :: Natural Language Toolkit.” Accessed May 2, 2022. https://www.nltk.org/.

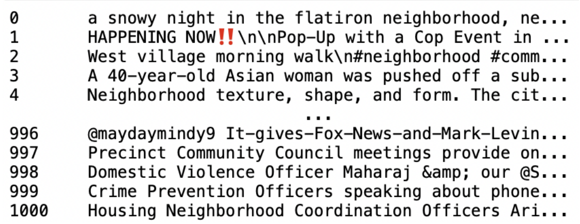
[10] “Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.0.2 Documentation.” Accessed May 2, 2022. https://scikit-learn.org/stable/.

[11] Tran, Author Khuyen, et al. “Pyldavis: Topic Modeling Exploration Tool That Every NLP Data Scientist Should Know.” *Neptune.ai*, 15 Nov. 2021, https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know.

Appendix

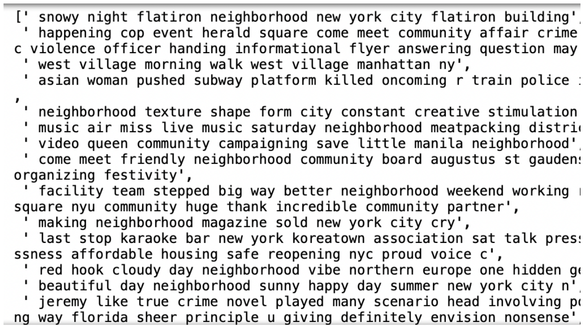
Code: STAT440_Project.html

Figure 1: Tweets Before NLP



```
0       a snowy night in the flatiron neighborhood, ne...
1       HAPPENING NOW‼️\n\nPop-Up with a Cop Event in ...
2       West village morning walk\n#neighborhood #comm...
3       A 40-year-old Asian woman was pushed off a sub...
4       Neighborhood texture, shape, and form. The cit...
                              ...
996     @maydaymindy9 It-gives-Fox-News-and-Mark-Levin...
997     Precinct Community Council meetings provide on...
998     Domestic Violence Officer Maharaj &amp; our @S...
999     Crime Prevention Officers speaking about phone...
1000    Housing Neighborhood Coordination Officers Ari...
```

Figure 2: Tweets After NLP



```
[' snowy night flatiron neighborhood new york city flatiron building',
 ' happening cop event herald square come meet community affair crime
c violence officer handing informational flyer answering question may
 ' west village morning walk west village manhattan ny',
 ' asian woman pushed subway platform killed oncoming r train police
,
 ' neighborhood texture shape form city constant creative stimulation
 ' music air miss live music saturday neighborhood meatpacking distri
 ' video queen community campaigning save little manila neighborhood',
 ' come meet friendly neighborhood community board augustus st gauden
organizing festivity',
 ' facility team stepped big way better neighborhood weekend working
square nyu community huge thank incredible community partner',
 ' making neighborhood magazine sold new york city cry',
 ' last stop karaoke bar new york koreatown association sat talk press
ssness affordable housing safe reopening nyc proud voice c',
 ' red hook cloudy day neighborhood vibe northern europe one hidden ge
 ' beautiful day neighborhood sunny happy day summer new york city n',
 ' jeremy like true crime novel played many scenario head involving po
ng way florida sheer principle u giving definitely envision nonsense',
```

Figure 3: Plate Notation for LDA

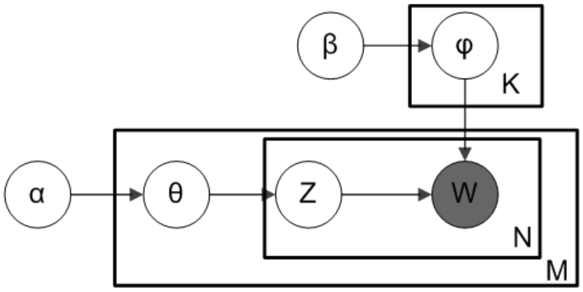


Figure 4: Dirichlet Priors

$$\begin{cases} \boldsymbol{\theta}_d \sim \mathrm{Dirichlet}_K(\boldsymbol{\alpha}) & d \in [M] \\ \boldsymbol{\phi}_k \sim \mathrm{Dirichlet}_V(\boldsymbol{\beta}) & k \in [K] \end{cases}$$

Figure 5: WordCloud of Frequent Words per Topic

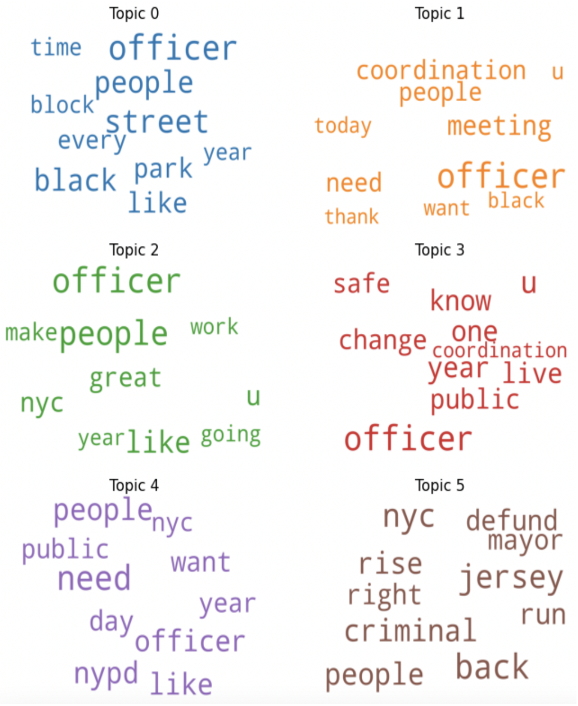


Figure 6: Intertopic Distance Map

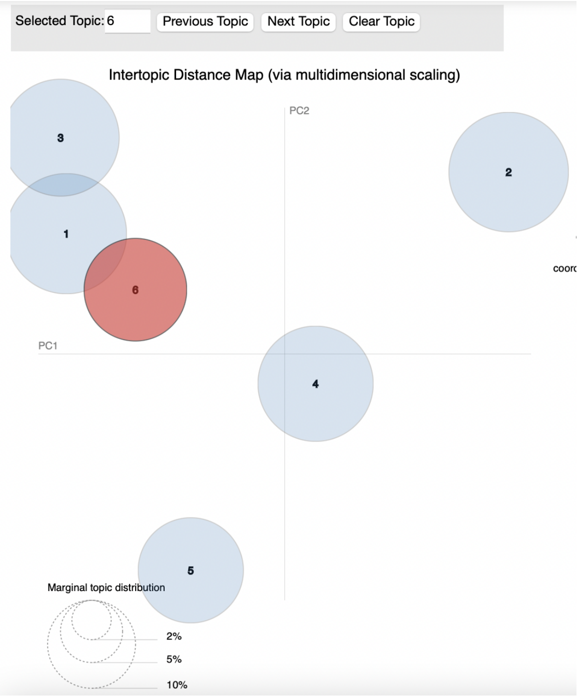

Selected Topic: 6
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
3
2
1
6
coorc
PC1
4
5
Marginal topic distribution
2%
5%
10%

Figure 7: Topic 1 Relevant Word Percentages

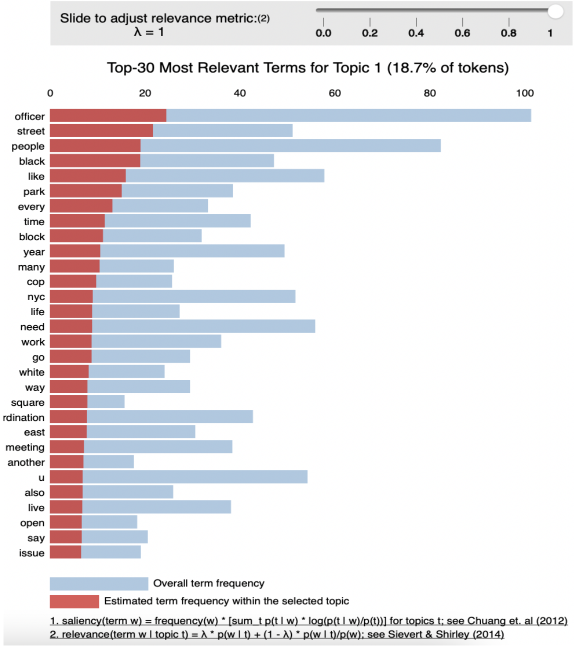


Figure 8: Topic 2 Relevant Word Percentages

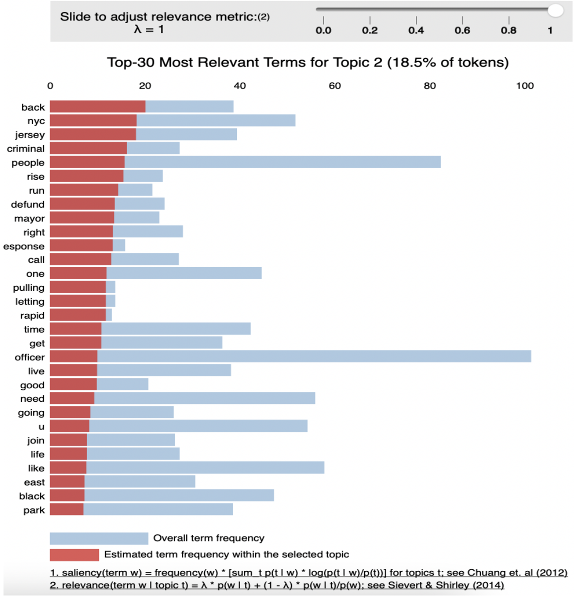
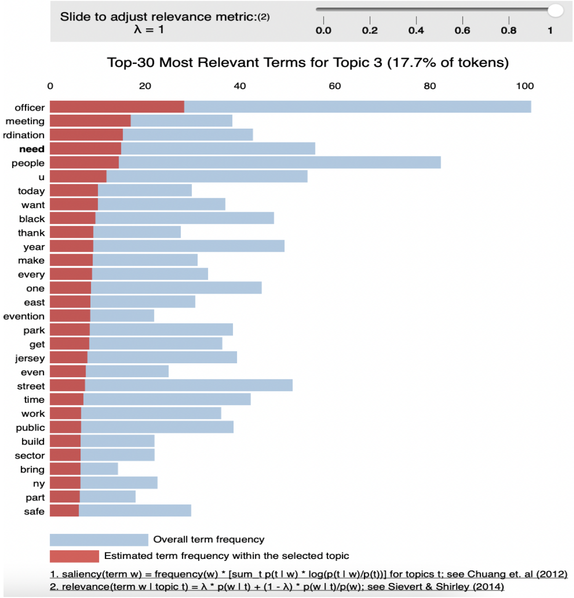


Figure 9: Topic 3 Relevant Word Percentages


Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Top-30 Most Relevant Terms for Topic 3 (17.7% of tokens)
0
20
40
60
80
100
officer
meeting
rdination
need
people
u
today
want
black
thank
year
make
every
one
east
evention
park
get
jersey
even
street
time
work
public
build
sector
bring
ny
part
safe
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)