What kind of disasters make us more perplexed?

#🤨 Chris Fleisch, Dylan Kenny, Krista Mar

Abstract

Our research is focused on using social media data from reddit and twitter to explore how language changes during natural and manmade disasters. Our first research question explores how much language changes relative to baseline by looking and data before and after disasters and using snapshot language models to look at changes in perplexity. Our second research question uses topic modeling to explore what people are talking about and see the similarities and differences between different types of disasters using Latent Dirichlet Allocation (LDA) and non-negative matrix factorizations (NMF). We found that in certain communities there is a clear spike in perplexity around the time of a disaster. We saw relevant topics appear during these events that were not present in the time before the event took place.

Introduction/Background

Using written text to measure social psychological responses during disasters has a long history in the fields of psychology but has always faced methodological challenges. Many standard research methodologies such as randomized controlled trials, random assignment, and repeated assessments are not possible in studying unpredictable events such as natural disasters and terrorist attacks. Traditional psychological study using retrospective self report was subject to distortions of memory. (Cohn, 2004) Social media has opened up a wealth of opportunity for researchers to obtain real time data in online communities.

Natural language processing is one approach in the interdisciplinary subfield of affect detection, which can take advantage of the large quantities of data available from social media. As the field has evolved, many different approaches have been suggested. Early approaches include looking at how text triggers different emotions and using multidimensional scaling to create visualizations of affective words based on similarity ratings of words. (Oswald, 1975; Oswald 1990). The more popular approach today is a lexical analysis of text to identify affective states of the writers. One way of doing this is the Linguistic Inquiry and Word Count, which identifies particular words that reveal the effect context of the text (2007). Currently, sentiment analysis has gained traction within the natural language processing field. (Pang, Lee 2008). However, after originally thinking to go down this route, we realized some obstacles: topical shifts can drown out changes in sentiment, expertly labeled corpora are sparse on social media data and the ones found did not neatly fit our topic area. Unsupervised or semisupervised models such as sentiment140 are promising, but still have shortcomings. Finding a good evaluation metric for unsupervised sentiment analysis is difficult. Our research adds to this body of literature, but we took a different approach. We looked at lexical innovation in online communities using changes in perplexity against baseline as a proxy. We thought that perplexity would increase during disasters relative to baseline and then decrease again.

To further explore the topics being discussed, we used latent dirichlet allocation as introduced by (Blei, Ng, Jordan, 2003) to perform topic modeling. This gave us a crisp way to see what kind of topics people were talking about pre, during, and post disaster. An out of the box latent dirichlet allocation models is what we used for our baseline. Further research that built on LDAs seemed promising, including research on author-topic modeling as suggested by (Rosen et al, 2004) or hierarchical LDAs. Since twitter data is special, these models needed to be adapted to this context. In (Hong and Davidson, 2010) they suggest aggregating short messages to get higher accuracy without having to adapt classic LDAs or author-topic modeling approaches. Another approach involves heterogeneous data sources like Wikipedia (Gupta and Ratinov, 2008), (Syed, Finin, and Joshi, 2008) to train an LDA that can then be used to find topics on new data sources like our comment data. Measuring the distance between the topics using the Jensen-Shannon divergence distance (Tong and Zhang, 2016) could also show us where topics change over time, but ended up not being very helpful for our research question.We also tried non-negative matrix factorization for topic modeling. While LDA is state of the art in many contexts and currently trendy in topic modeling, some researchers have found that LDA has convergence issues on small corpus and that NMF works better in some instances(Saha, A., & Sindhwani, 2012). NMF was considered along with LDA in the meta review, which focused on topic modeling using social media data in (O'Callaghan et al, 2013). We explore limitations of our work in the Appendix.

To summarize, our main two research questions are the following:

(1) How much does language change relative to baseline around disasters? (methodology: snapshot language models; eval metric: perplexity)
(2) Are the types of topic that people talk about post disasters different between natural disasters and manmade disasters? (methodology: LDAs, NMF, eval metric: qualitative coherence of topics)

Data

Our initial data sets from twitter and reddit geographically and chronologically centered around disasters in 2017: Hurricane Harvey, Hurricane Irma, Hurricane Maria, and the Mandalay Bay Las Vegas shooting. We also looked at Reddit data from historical disasters: 2012's Hurricane Sandy, 2013's Boston Bombing. And we take a quick look at a large event like 2013's Boston's World Series for comparison.

- Twitter (16,000-50,000 geotagged tweets total, 3 weeks prior and 2 weeks after disaster)
- Reddit (relevant city subreddits, 30 days prior and 30 days after the disaster)

The Reddit comment data is publicly available (http://files.pushshift.io/reddit/comments/) in compressed format. Using grep we parsed out the subreddits we were interested in: r/boston (bombing), r/boston (World Series), r/florida (Irma), r/houston (Harvey), r/Miami (Maria), r/nyc (Sandy), r/PuertoRico (Maria), r/vegas (shooting).

The popularity of each subreddit and the number comments varied. The vegas community had the lowest with about 12,000 comments during the 60 day period around the event. Houston had over 185,000 comments during the 60 day period we analyzed. The r/houston dataset was also randomly sampled to about ¼ its original size which helped with our analysis. See Figure 1. The vocab size for each subreddit ranged from about 10,000 to over 58,000 over the same 60 day period. See Figure 2. Preprocessing the reddit data involved tokenizing it into sentences and words using the nltk library and padding it with beginning and ending sentence tokens when generating perplexity. We also limited the length of the comments for both perplexity and LDA to reduce noise.

The Twitter data was scraped using a tool, GetOldTweets. The sample size of each community ranged from 16,000 to 55,000 during the entire capture period. San Juan was an outlier and proved to be a challenging dataset to use at only 1800 tweets. The team suspects that twitter is not as popular in Puerto Rico as it is in the continental US. This dataset was kept in the study as an experimental case. The team intended to measure how effective these techniques would be on such a small dataset. See Figure 3. The vocab size for each location's Twitter data ranged from 30,000 to 87,000, excluding San Juan which didn't have many tweets at all. See Figure 4.

Daily snapshot language models

Our methods are similar to those used in psychological literature in using online journal or social media to look at baseline and post disaster use of language. We borrow the frame that (Danescu-Niculescu-Mizil et al, 2013) develop to look at language change in online communities over time using language snapshot models. We are adjusting the snapshot frequency to a 3 day window because of the time sensitive nature of our work. Different snapshot frequencies have been tested by the team, but 3 days seems like a stable benchmark. Disasters occur quickly and more observations are needed than the month model used by(Danescu-Niculescu-Mizil et al, 2013), since linguistic change in online communities might happen more slowly.

Topic modeling

We used topic modeling to see if we could detect the various disasters in the twitter and reddit data. As social media data is messier and often contains oddities like misspellings and abbreviations, we knew there would be some limitations with the efficacy of topic models. Using preprocessing methods suggested by literature, we originally tested using gensim's out of the box LDA model as our baseline. A challenge of LDAs is picking the correct number of topics k. With our baseline LDA, we tried a few different values of k, but didn't find coherent topics. After trying our baseline LDA, we tried a couple of additional models in the gensim package. One was the hierarchical dirichlet process (HDP). The advantage over LDA is that the number of topics is not needed beforehand. While the theory isn't quite solidified, HDP also had problems with coherence so we decided to try other models. We have included the NMF model as our final model for twitter data and an LDA pretrained on Wikipedia data as our final reddit model.

Results

Baseline result on reddit data

We evaluated 30 days before and after events in a range of subreddit topics to see if the language of online communities changes during natural disasters or terrorist attacks. Using the Add K smoothed trigrams to evaluate perplexity using the 3 methods outlined above on the r/miami data. The first two implementations of the snapshot model did not reveal the large spike in perplexity around the event we were looking for. However, the 3rd implementation of the 3 day train/eval snapshot model showed a spike after hurricane Irma hit Miami.

We also found new words that were used in the communities during the events. Words like 'hurricane', 'harvey', and 'explosion' quickly entered the vocabulary. Here's an example of the Miami subreddit that has an increase in perplexity using Add K around the time of the hurricane and associated words that were introduced during shortly after September 1st. See Figure 5. The Kneser-Ney smoothing trigrams showed very similar results and spikes, but with overall lower perplexity which is what we ended up using for the rest of the results. See Figure 6. We then focused on implementation 3 of our snapshot model and tried shortening the comment length to 15 words to try and reduce noise around the event spikes in our plots. See Figure 7. The 15 word limit reduces the noise in the perplexity, but still allows for a significant spike in the perplexity when the hurricane hits the Miami area. Our final analysis will use the Kneser-Ney model with shortened 15 word comments for calculating the snapshot perplexity. See Figure 8.

Final Perplexity Analysis on Reddit

For the final analysis of the reddit perplexity we focused on the Kneser-Ney trigram model, using 15 words from each comment, and 30 days before and after each event. Each event starts around the middle of each plot (all subreddit perplexity result plots available in the Appendix).

One example is from the Boston subreddit. The bombing took place on April 15th and we see an immediate spike following that time. The language of the community has a distinct change that is evident in the plot of the perplexity. We also see a significant spike in words that were not used before in the community. After a couple weeks the perplexity returns to a much lower level and the new words that were introduced also begin to leave the vocabulary of the community. See Figure 9. Some of the perplexity changes saw much bigger changes like in the florida subreddit that went from a perplexity of 50 to 175. While others showed changes that were not as significant, but still a visible change.The perplexity measure has helped us to identify when language changes in online reddit communities. They seem to change significantly when a large event takes place in their associated city. The hurricanes show a slightly wider spike probably due to the lingering effects of the hurricane. The spikes for the terrorist events seem to be a little more narrow and return to lower levels a little quicker (see all results in Appendix).

Baseline Results on Twitter Data

The twitter data yielded similar results to the reddit analysis. In this case, we used 14 days of training data and 3 day test windows. Additional tests were run with 1 day evaluation windows to see if test window size made a significant difference. There was no appreciable difference; however the 1 day windows gave provided a little more granularity in the plots, but at the expense of smoothness. Essentially, there was more noise with 1 day windows. For the baseline, we used Add K for our smoothing of the twitter data; however the final model used Kneser-Ney smoothing. See Figure 10, 11.

The Vegas shooting happened 8 days after the training data, at the initial spike in perplexity. The team has investigated the second spike (10/18/2017) and it turns out there was a conference, Adobe MAX, in Vegas that day and people were talking quite a bit about that and the traffic it caused. Although it's a bit hard to tell, Houston made landfall at the time of the initial spike in the graph. Switching to KN smoothing should make some of these plots a little easier to read. Unsurprisingly, the noise of the results seems related to the lack of data. San Juan did not have a lot of twitter activity prior to Hurricane Maria, so it's not particularly surprising that there isn't a big spike at any one point.

Final Perplexity Analysis on Twitter

The perplexity analysis of Twitter data was valuable; however there were certainly plenty of challenges. The first obvious challenge was in trying to overcome the noisiness of tweets. We performed various types of sanitation to try to whittle down the vocabulary to something more useful and consistent. This was an anticipated challenge. We did *not* anticipate having trouble accessing historical geographically tagged tweets. This became a limiting factor in our analysis. The tools we used to gather the data worked intermittently, partially due to rate limiting, but eventually provided us enough data to get meaningful results. The team performed a series of different training methodologies and sizes. The sliding training window technique that performed well for the reddit data produced results that were not particularly impressive with the Twitter data. This is likely due to the limited sample size and relative noisiness of the tweets. Results of the sliding window attempts are shown in our iPython notebook to avoid taking up too much room in this document.

Using a larger 14 day static training period with 1 day test periods seemed to perform best and yield very obvious spikes around the time of the event. Unfortunately it was also prone to suffering drastically in the event of scarce data. Using a larger 3 day window helped balance out these problems. A 3 day test period was helpful and would show cleaner spikes. It also helps smooth out a lot of day to day changes in the model. This made the results more reliable and ensured it was less likely to be caused by random chance. Both test sizes tended to show roughly the same trends though and could be used to infer when events took place. Tweets from Miami demonstrate the effect we'd expect to see. The below plot shows perplexity measurements across 1 day test windows. It's easy to see how much perplexity fluctuates on a day to day basis. Despite this, the spike on the day of the event is still very pronounced. See Figure 12.

The perplexity has a sharp spike around the point at which Hurricane Irma reached Florida. This particular test case doesn't see an increase in subsequent days; however the volatility and noisiness of the Twitter data may account for that phenomenon. We do notice an appreciable growth in the perplexity as the hurricane approaches. This supports the notion that things like hurricanes, which have warning signs and are discussed prior to the actual event, are more likely to see more gradual peaks as opposed to a surprise event like a mass shooting. This would be a situation in which long static training periods are better than sliding ones. A sliding training period, depending on its size, would likely consume the slow changes of gradual events and therefore mask their magnitude. Contrarily, the Las Vegas data supports the notion of sliding training windows. Las Vegas is intrinsically a very dynamic location with an ever changing atmosphere. The shooting saw a spike in perplexity; however the Adobe MAX conference also created a large spike. For a location like this, having wider sliding training windows is likely the correct approach to strike a balance between a stale model and one that is so volatile that it's worthless. The results of all static window analyses are included in the Appendix, while all other results (including sliding windows) are included in the Twitter iPython notebook. See Figure 13.

Topic Modeling

Overall topic modeling worked, but unevenly. The quality of results changed across disasters based on noisiness and quantity of the data. Extremely small data sets like twitter data for Puerto Rico performed poorly as well as datasets that had a relatively low incidence of keywords associated with the disasters. Our best topic model for reddit was a supervised LDA and for twitter our best model for twitter data was an unsupervised NMF model One of the options was pretraining our LDA model on Wikipedia data, which is much cleaner and is often used in topic modeling tutorials to create nice, clean topics. This would normalize what topics we found, which could make comparisons across disasters easier. For twitter, we use of non-negative matrix factorization models, since this allowed us to using tf-idf, not possible for LDAs, which we thought may improve the quality of our topics since twitter data was particularly noisy.

We've identified some larger trends. In the case of wikipedia LDA modeling, when there is a hurricane topics like 'storm', 'flood', 'water', 'damage' and 'recover' appear compared to 'guns', 'killing', 'dead' and 'attack' when there is a shooting. This seems to confirm previous research in psychology that people afraid of terrorist attacks most, then gun violence, and much further down natural disasters (Chapman, 2016) with a focus on death and killing in the case of manmade disasters and a focus on damage and recovery in the case of natural disasters. For twitter, we saw some topics as important before the disaster, but not after the disaster. In the case of hurricane harvey twitter data, there was a lot of topic about traffic before the disaster (#0 topic), but after the disaster, it completely disappears from what people are talking about. In the case of twitter data in Miami, the 5th topic was about going to beaches in Miami, but this disappeared from topics post disaster.

Where did topic modeling not work so well? The poorest performance was with Spanish language topic modeling on twitter data. All of the data from Puerto Rico was in Spanish as well as some of the data from Houston, Miami, and Florida. In the case of Houston, Miami, and Florida, the topic models were put Spanish into a topic, but the topic itself was incoherent. Language modeling is known to be not great in the case of code switching, which is very common in social media. While there are solutions such code switching language models such as those described in (Peng, N., Wang, Y., & Dredze), we did not attempt these as this wasn't our main goal, but it could implemented in future work. Polylingual topic modeling has found some success, but implementing polylingual models was beyond the scope of this project. As for topic modeling in only Spanish, out of the box topic modeling are optimized for English and secondly we had a problem of much smaller data sets from the Puerto Rico communities. In the case of our LDA Wikipedia model, there were coherent topics, but less coherent than with other disasters that had only English. The topic modeling on twitter data for Puerto Rico failed. This was due to both a paucity of Puerto RIco tweets and poor performance of our language models on Spanish.

Baseline Topic Modeling

We used the gensim package out of the box for topic modeling. The core code is based on the onlineldavb.py script by (Hoffman, Bach, Blei, 2010). Following the guidelines on the gensim package, we processed text before running it through the LDA. However, the topics that came out didn't pass the first sanity check suggested by(Blei, Ng, Jordan, 2003), looking 10 highest probability words for each topic and see if topics learned are meaningful. These topics were not coherent, so we tried other model. The results from the Baseline Topic Modeling are in the Appendix.

Topic modeling Final using NMF Results with Twitter

We used an unsupervised NMF model from scikit learn on twitter data. First, we first preprocessed the tweets to clean them and remove links, and filtered out stopwords. We used tf-idf to reduce the amount of noise in our data and try to pick out more important words. We then removed duplicates. We used different values of k number of topics for each disaster to try to improve results. We used NMF with generalized Kullback-Leibler, since Kullback-Leibler divergence tends to yield higher accuracy than other divergence measures and the NMF model with Frobenius norm did not perform as well for topic coherence. We looked at before and after the event for twitter. This proved to have more coherence than smaller windows, likely because of the brevity of tweets and paucity of geotagged twitter data.

In the the Vegas shooting twitter data most people seemed to be tweeting about going to parts of vegas they were visiting (topic #0), traffic (topic #1), enjoying time with their friends (topic #2), sports (topic #3), the vegas strip (topic #4), hotels (topic #5), etc. Most of the topics seem to be about enjoying their time in Vegas. After the shooting, the majority of topics seem to be about enjoying life in Vegas, but we were able to find a coherent topic, topic #10 that is clearly about the shooting. This parallels what we saw in the perplexity for the Vegas Twitter data. While there was a bit of a spike in perplexity, the spike wasn't as clear as it was in other disasters. People are likely more likely to geotag Vegas to talk about the good times they are having rather than the Las Vegas shooting. The entirety of the NMF results of the before and after twitter topic modeling is in the Appendix.

LDA Final (Wikipedia) Results with Reddit

Using the gensim library we trained an LDA using an archive of Wikipedia. We filtered out stop words. We removed any words that appeared in less than 20 articles or more than 10% of articles as described in the package documentation. Once we had defined our vocabulary we trained the LDA on 3 million documents from Wikipedia to create 1000 different topics. The training time took over 4 days. Earlier attempts to train on fewer articles and topics did not provide good results.

With our LDA model built we then passed in our reddit data in 3 day windows. This created a list of topics for each 3 day window. We then looked at those topics for each window to see if there were any interesting topic discoveries that would rise into the top topics during the event. We also attempted to look at Jensen-Shannon divergence distance of the topics by calculating the distance between each windowed segment. And also by calculating the windowed topic segment distance to the topic for the whole data set. However, both of these did not prove to provide any consistent results that would help us to identify interesting topic changes over time.

The first few topics always appear the same for all the different reddit datasets. However, relevant topics for each event were identified during the window of the event. These topics do not appear first, but further down in the topic classification. The data input to the LDA used the 15 word length comments to try and reduce the amount of noise in the data that worked well with the perplexity. In the Houston reddit comments the LDA model started to find topics related to the storm that had not appeared before. After these topics were introduced around the time of the event they would then appear further down in the list of topics or disappear from the topics by the end of the next 30 days. All results from the Wikipedia trained LDA topics for each subreddit available in the Appendix.

Removing some of the common topics found for each day might have helped make these result more impactful. Removing the common topics also might have made the divergence distance metric more useful as well. This idea has not been explored yet, but could be useful for future research. In our current setup we did a visual inspection to see which topics were produced for each windowed segment and we could see new topics being introduced when the event occurred. After some time depending on the event the topics that were introduced would descend from the top 10 topics list.

Conclusion

We found that perplexity changes can be useful for identifying when a large event happens in online communities. Some of the changes are more extreme than others, which could be due to the activeness of those communities, but there were generally visible spikes in perplexity around large natural disasters and human caused events. To help with the analysis the data sometimes needed to be preprocessed or limited when there was too much. We were also able to see some of the new vocabulary that entered the online communities during these events. Topic modelling was more challenging and it was not always clear how to improve the results. We were overall able to pick up topics for the disasters. On the Wikipedia trained LDA topic models, we were able to identify topics that were relevant to each type of disasters or terrorist attack. When there is a hurricane topics like ‘storm’, ‘flood’, ‘water’, and ‘damage’ appear compared to ‘guns’, ‘killing’, and ‘attack’ when there is a shooting. This seemed to confirm our hypothesis that the psychological impact of natural disasters verses. man-made disasters is different. We saw the lowering of importance or disappearance of some topics like traffic in Houston or going to the beach in Miami. We saw a general trend that if perplexity spikes were cleaner or easier to detect in our language modeling section, topics were higher ranked among topics detected. This seemed to confirm our initial hypothesis that the changes in perplexity were about the topical shifts. Overall we were able to set out what we initially tried to test with varying success by disaster. We have ideas for future work in the Appendix.

Citations

Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A. (2016, February 25). 1. Reddit news users more likely to be male, young and digital in their news preferences. Retrieved November 16, 2017, from http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Chapman University. (2016, October 12). What do Americans fear?. *ScienceDaily*. Retrieved December 16, 2017 from www.sciencedaily.com/releases/2016/10/161012160030.htm

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, *15*(10), 687-693.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, *1*(1), 18-37.

C.E. Osgood, W.H. May, and M.S. Miron, Cross-Cultural Universals of Affective Meaning. Univ. of Illinois Press, 1975.

C.E. Osgood, Language, Meaning, and Culture: The Selected Papers of C.E. Osgood.
Praeger Publishers, 1990

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013, May). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 307-318). ACM.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856-864).
Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, 263-286.

Hong, L., & Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics (pp. 80-88). ACM.

Laylavi, F., Rajabifard, A., & Kalantari, M. (2016). A multi-element approach to location inference of twitter: A case for emergency response. *ISPRS International Journal of Geo-Information*, *5*(5), 56.

Nikolenko, S. I. (2016, July). Topic Quality Metrics Based on Distributed Word Representations. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 1029-1032). ACM.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2013). An analysis of interactions within and between extreme right communities in social media. In *Ubiquitous social media analysis* (pp. 88-107). Springer, Berlin, Heidelberg.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.

Peng, N., Wang, Y., & Dredze, M. (2014). Learning Polylingual Topic Models from Code-Switched Social Media Documents. In *ACL (2)* (pp. 674-679

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004, July). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487-494). AUAI Press.

Saha, A., & Sindhwani, V. (2012, February). Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 693-702). ACM.

Shearer, E., & Gottfried, J. (2017, September 07). News Use Across Social Media Platforms 2017. Retrieved November 16, 2017, from
http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

Sloan, L., & Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one*, *10*(11), e0142209.

Zareen Saba Syed, Tim Finin and Anupam Joshi (2008). *Wikipedia as an Ontology for Describing Documents.* Association for the Advancement of Artificial Intelligence

Rakesh Gupta, Lev Ratinov (2008), *Text Categorization with Knowledge Transfer from Heterogeneous Data Sources* Association for the Advancement of Artificial Intelligence

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011, April). Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval* (pp. 338-349). Springer, Berlin, Heidelberg

Zhou Tong and Haiyi Zhang (2016, August). A Document Exploring System on Lda Topic Model for Wikipedia Articles. In *The International Journal of Multimedia & Its Applications* (IJMA) Vol.8, No.3/4

Appendix

Limitations of our work:

Twitter and reddit users aren't representative of all people, not even all internet users. According to Pew research, 21% of all internet users user twitter. 4-6% of internet users use Reddit. (Shearer, Gottfreid, 2017). While twitter is roughly balanced in terms of gender of users, reddit has a ratio of approximately 2:1 men: women. (Barthel, Stocking, Holcomb, & Mitchell). People worst hit by disaster may lose access to power and thus can't complain about their problems on Twitter and Reddit.

We chose Twitter and Reddit because of public access to their data and APIs. We chose Reddit for the longer posts, which would perform better on language modeling tasks. We chose twitter since it is commonly used in disaster relief work that uses social media because of its real time nature and option to geotag. However, only 2% of all tweets are geotagged (Laylavi, 2016). There is reason to believe that people who geotag their location are not representative of the wider Twitter community(Sloan and Morgan, 2016). While there are known approaches to location inference, that was beyond the scope of the project. Language snapshot models haven't commonly been used for affect detection, but since we did not have a fix for getting good labeled data for sentiment analysis, this seemed like a feasible approach.

Future Work

In the future it would be interesting to extend analysis to other disasters to get better sense of meta-analysis of disasters. It would be interesting to try location inference on tweets, since we found geotagged tweets hard to come by. Delving deeper into to connect our work to the broader field of research would be a potential extension. While sentiment140 and other methods for sentiment analysis are flawed, they could be adapted and taken with a grain of salt. We did not explore applications of our discoveries, but could envision using these tools to look at large collections of past reddit comments looking for events that affected the community in a more automated way. It might also be interesting to see if a more real time analysis could be done to alert community managers to the changes that are happening on their social sites.

Exploratory Data Analysis

Figure 1

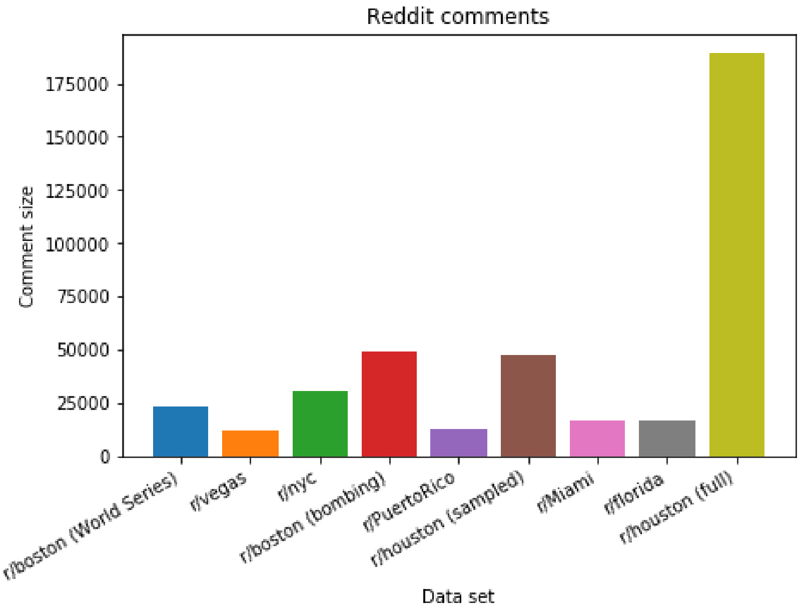


Figure 1

Figure 2

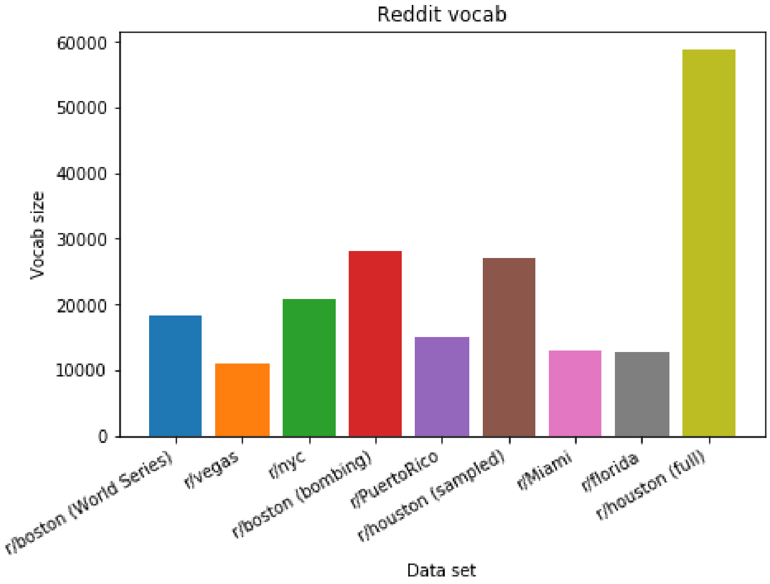


Figure 2

Figure 3

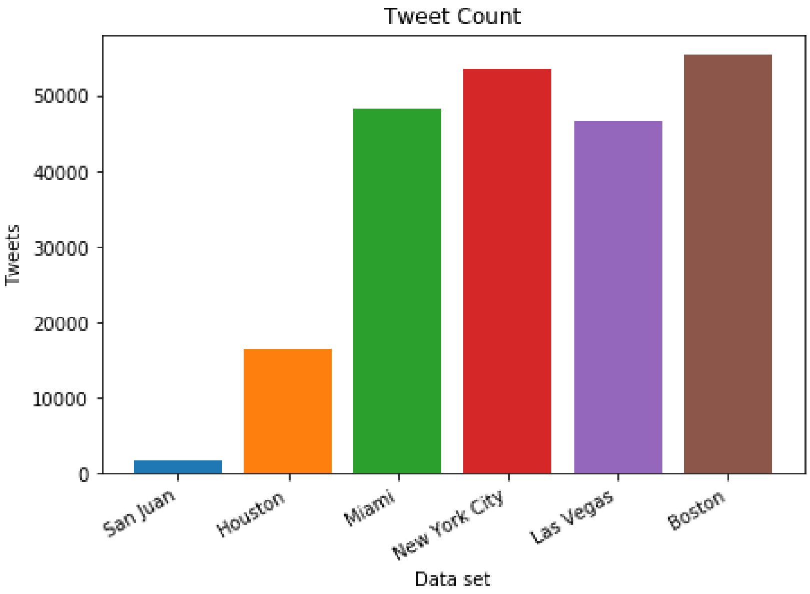


Figure 3

Figure 4

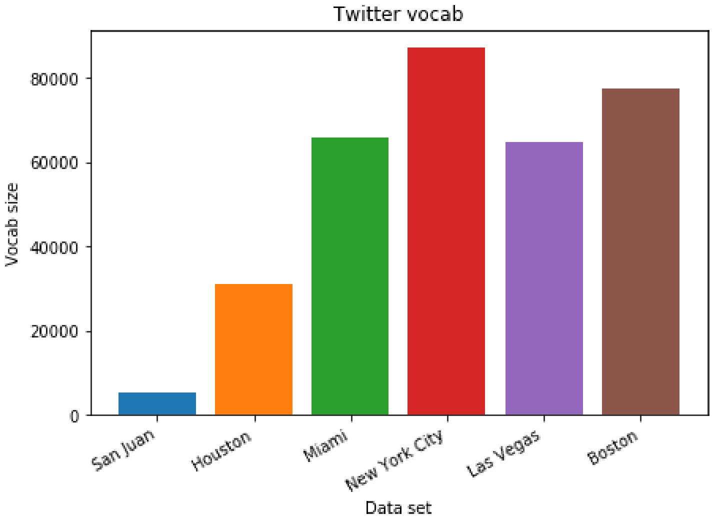


Figure 4

Figure 5

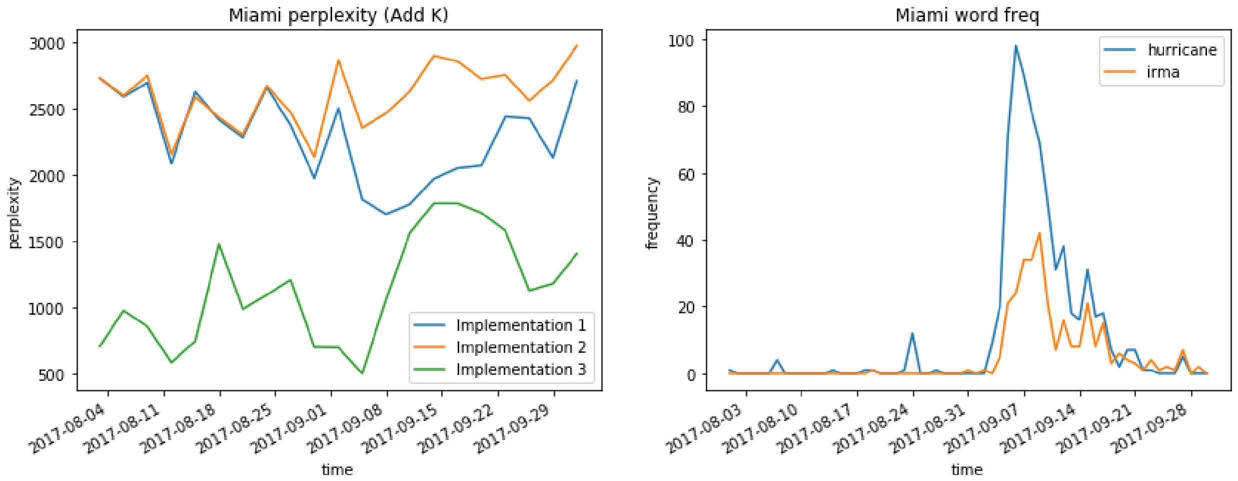


Figure 5

Figure 6

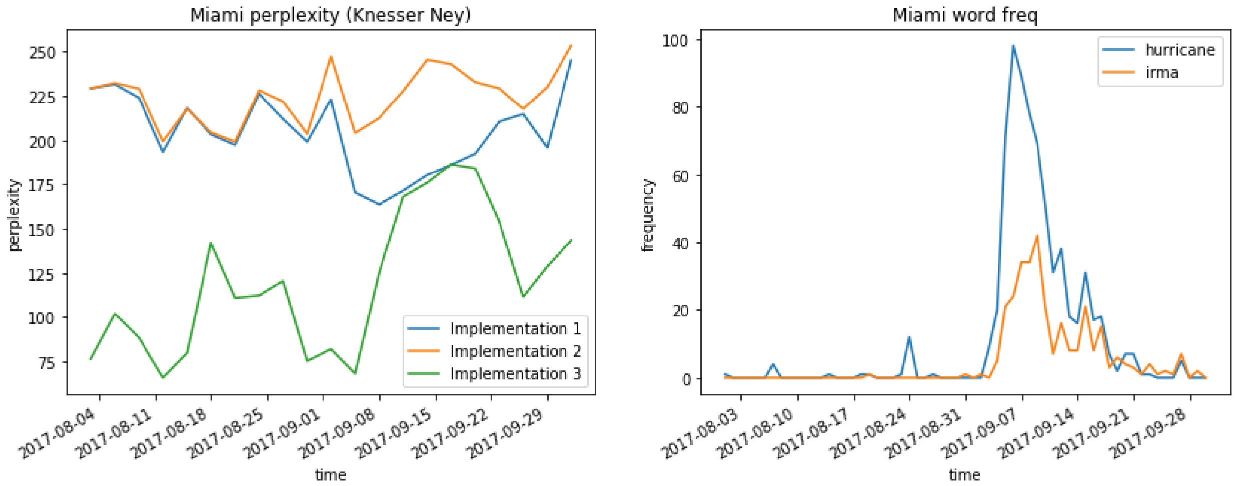


Figure 6

Figure 7

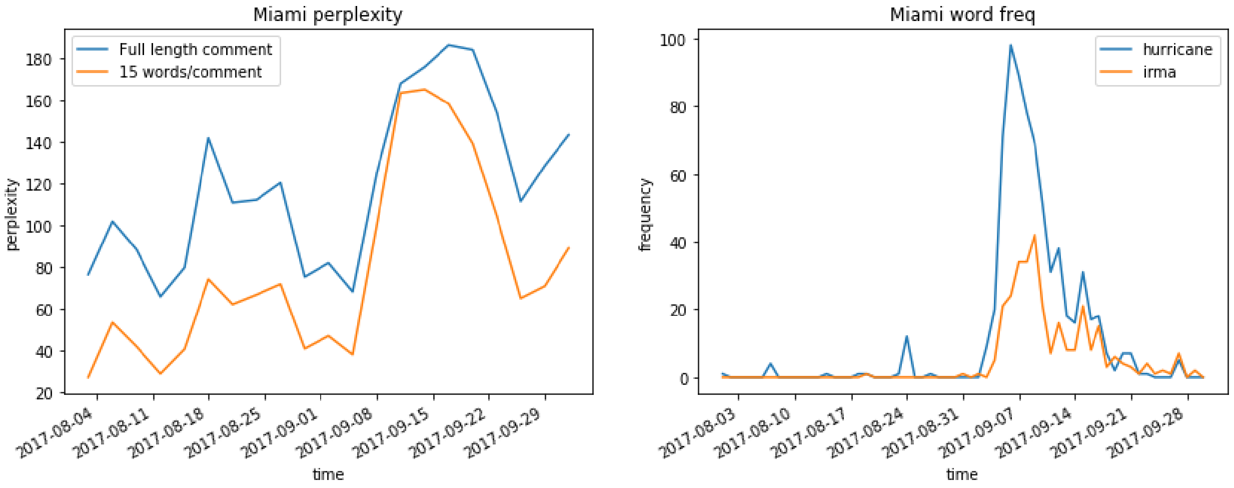


Figure 7

Figure 8

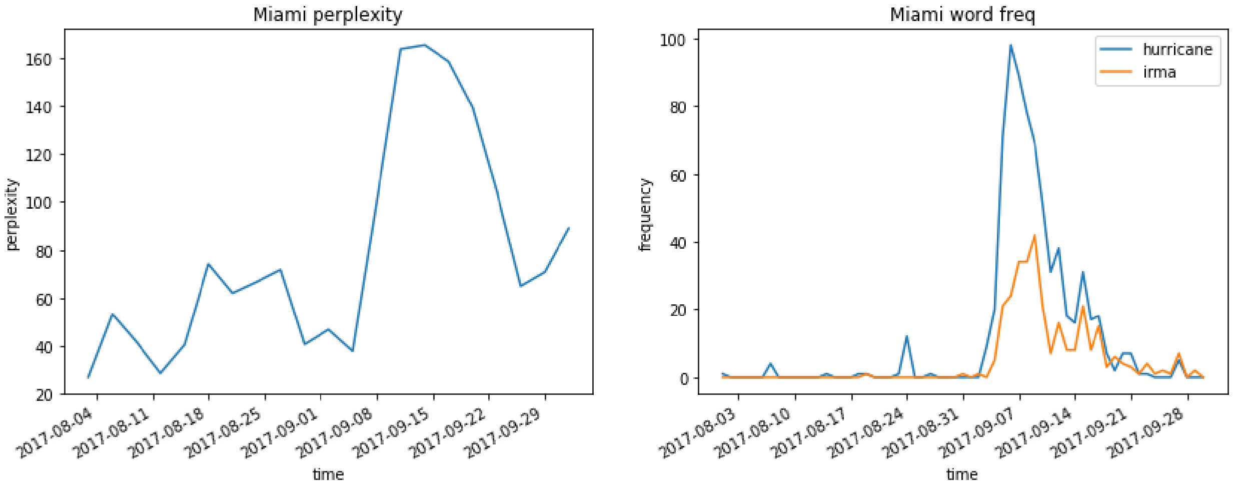


Figure 8

Figure 9

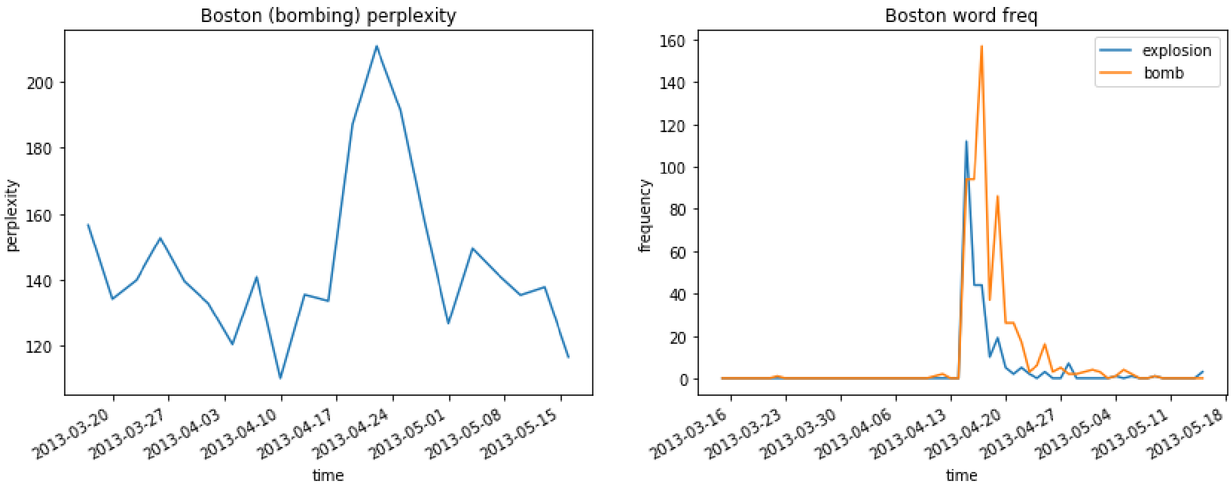


Figure 9

Figure 10

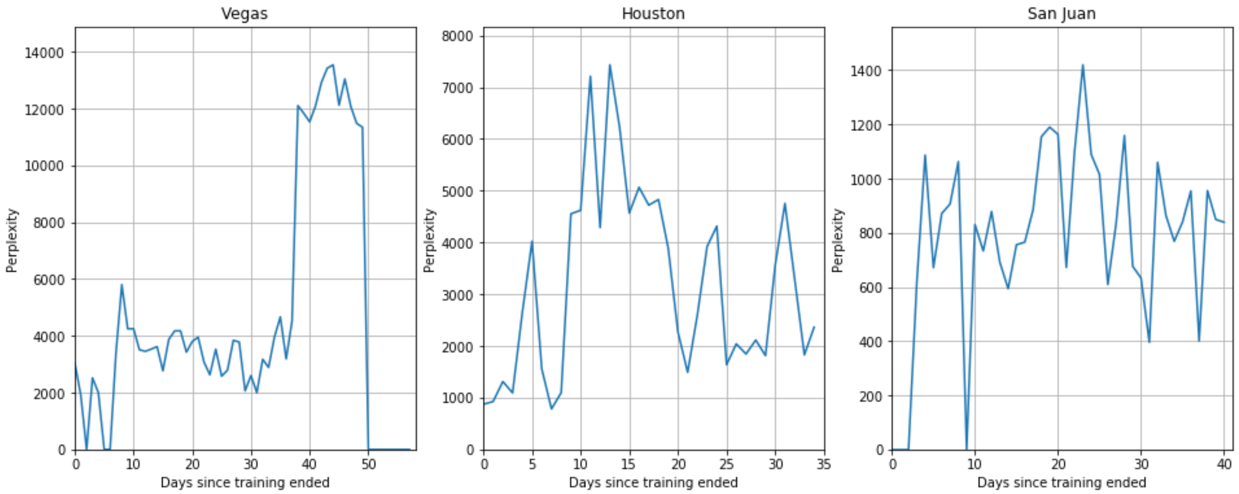


Figure 10

Figure 11

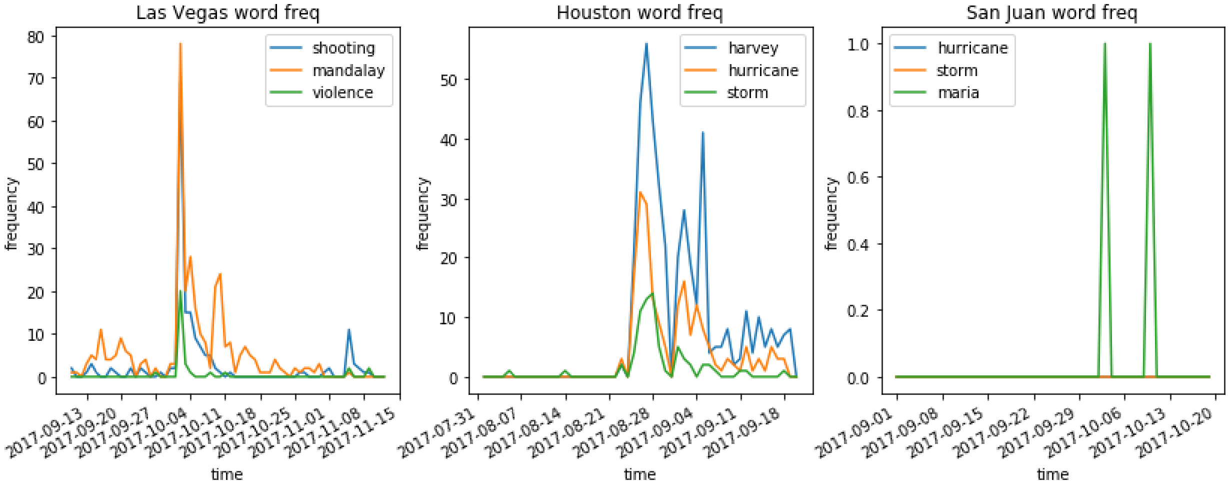


Figure 11

Figure 12

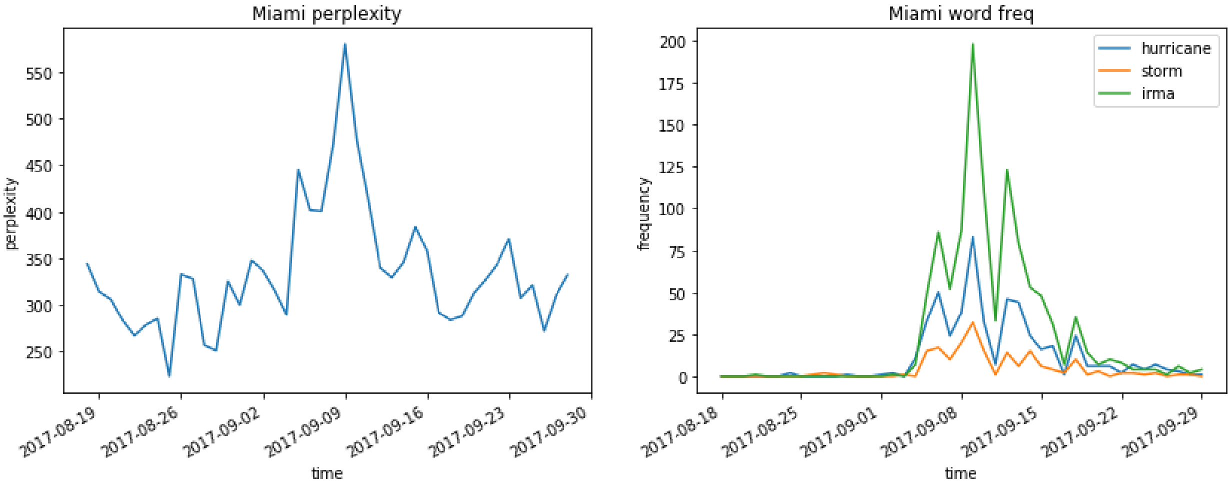


Figure 12

Figure 13

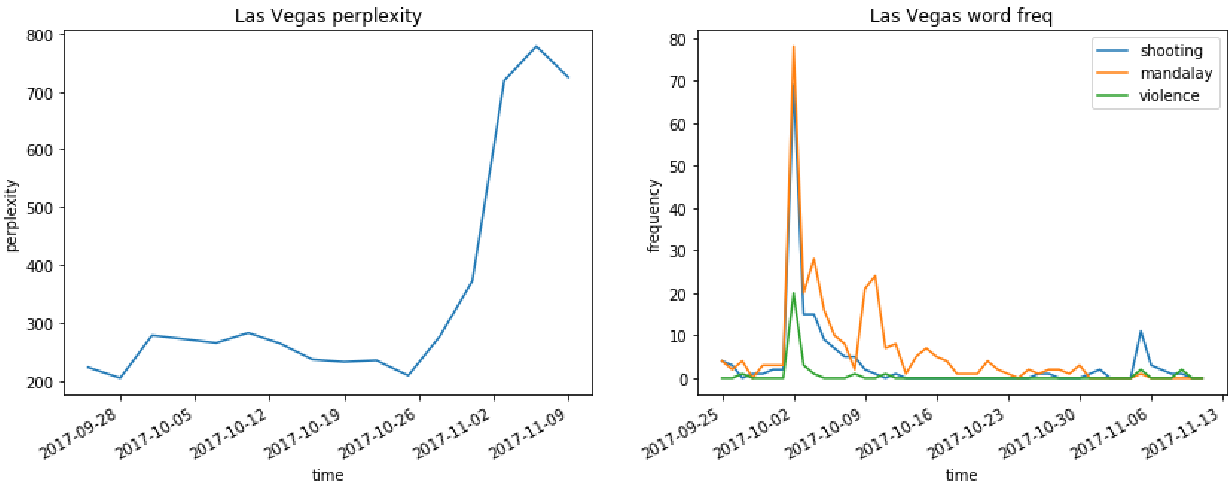


Figure 13

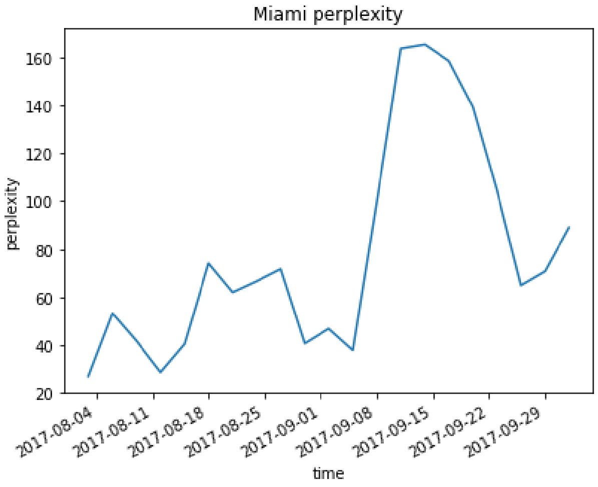
Baseline Topic Modeling Results

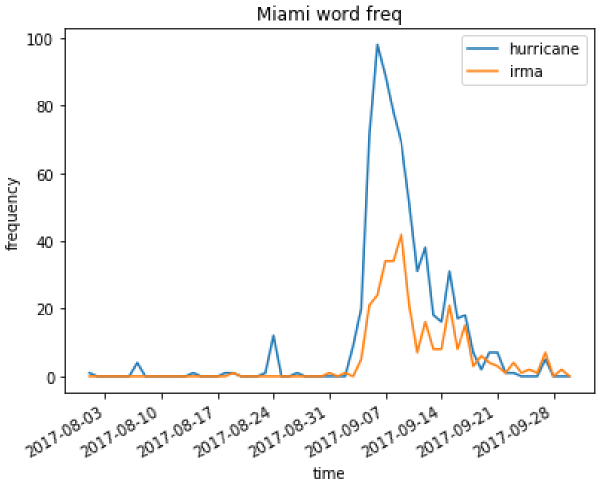
Using a the basic LDA algorithm with k=50, the top 10 topics were for Hurricane Irma (Florida/Miami)

- 0.006*"Miami" + 0.006*"It" + 0.005*"like" + 0.005*"people" + 0.005*"one" + 0.004*"power" + 0.004*"get" + 0.004*"would" + 0.004*"deleted" + 0.004*"know"'
- u'0.008*"get" + 0.007*"people" + 0.006*"like" + 0.006*"Miami" + 0.005*"com" + 0.005*"You" + 0.005*"It" + 0.005*"one" + 0.005*"deleted" + 0.005*"power"'
- u'0.006*"The" + 0.006*"like" + 0.006*"people" + 0.005*"Miami" + 0.005*"get" + 0.005*"It" + 0.005*"power" + 0.005*"one" + 0.005*"would" + 0.004*"know"

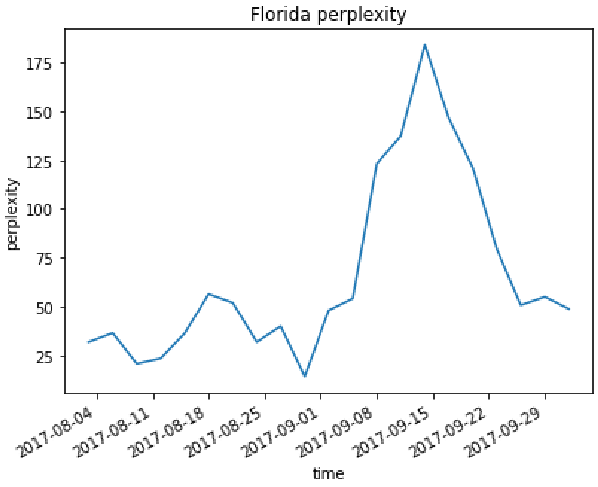
Reddit Perplexity Results

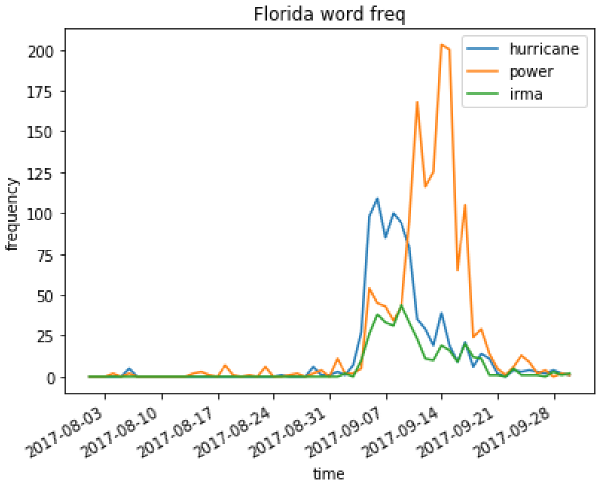
Miami: Hurricane Irma lasted from August 30th - Sept. 16. We don't see a spike in perplexity until about September 8th in the online community. This is also about the time that the words 'hurricane' and 'irma' spike in popularity. Perplexity tops out at about 160.



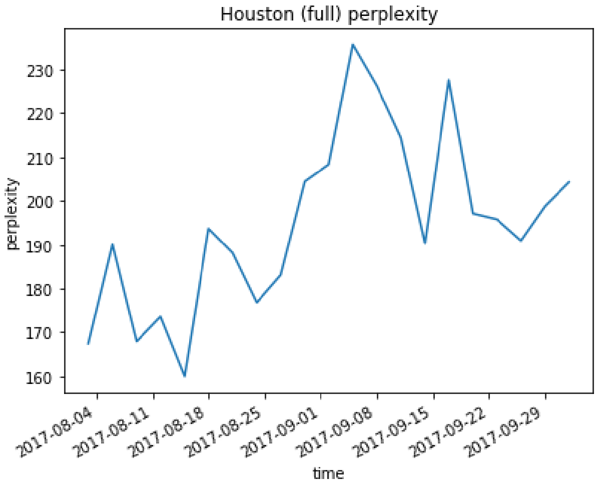


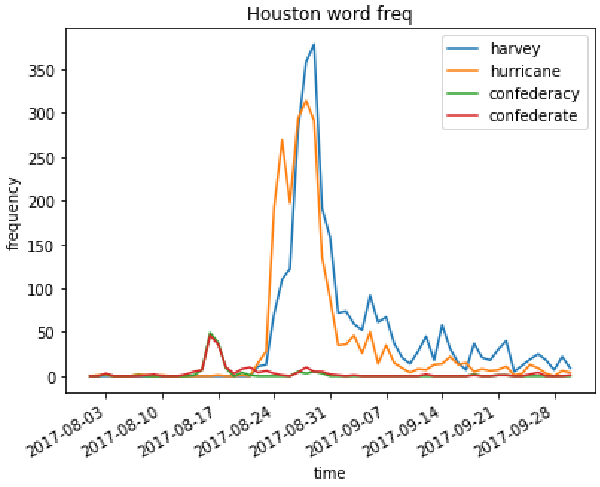
Florida: This subreddit spike happens about the same time as the Miami dataset. We see similar words show up and also ‘power’ is now being used significantly more. The perplexity reaches over 175 which is a little bit more than the Miami dataset of the same event.



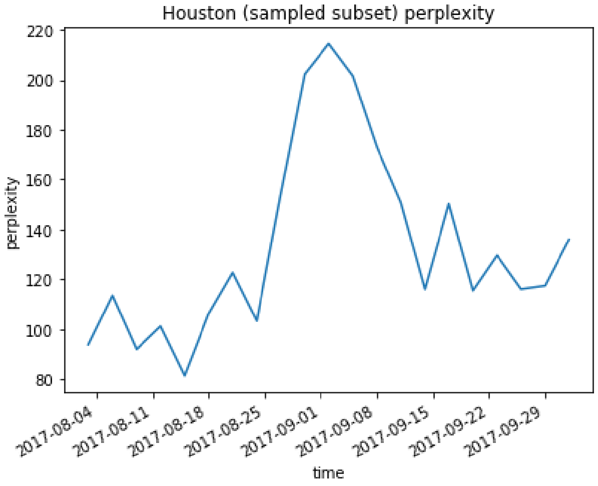


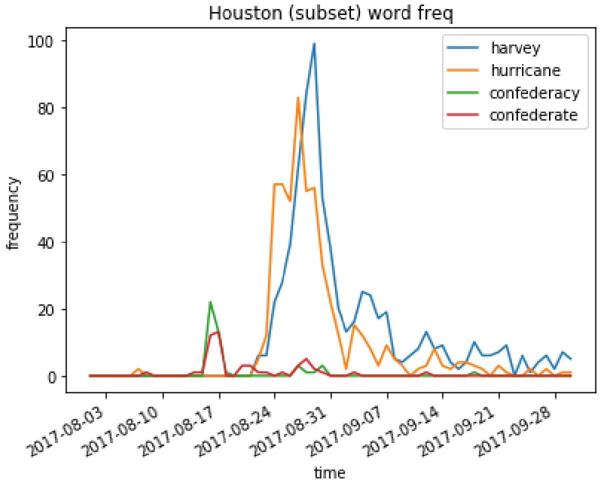
Houston (full dataset): Harvey hit houston around August 25. The full dataset shows a spike after the 25th and goes up to about 230 followed by another smaller spike. While this larger dataset shows a big spike it’s not as clear as the sampled dataset below. This dataset has significantly more comments than any of the other’s and is likely full of lots of noise.



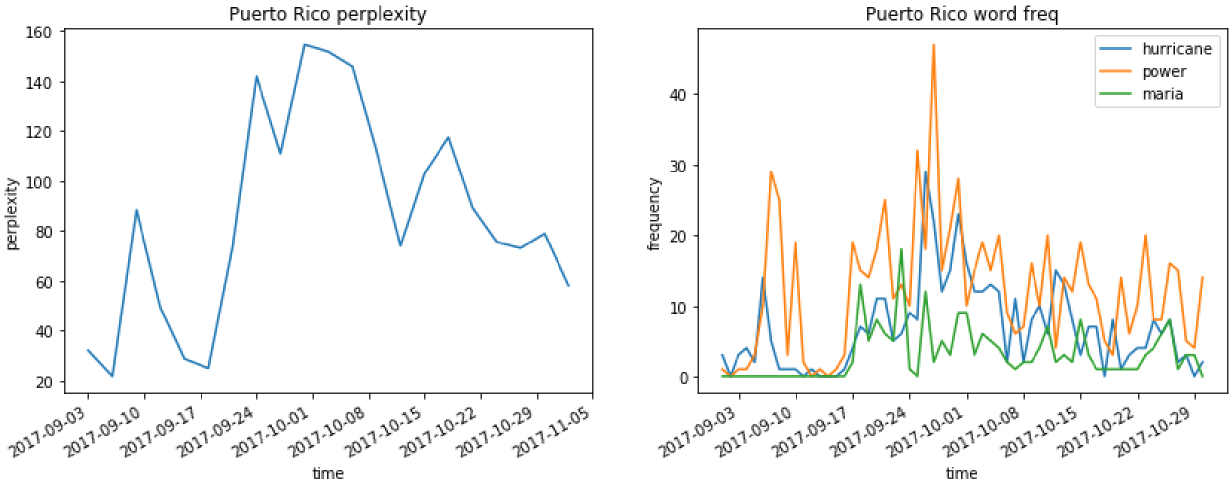


Houston (sampled dataset): Using a random sample of the full houston dataset that's about ¼ the size we can reduce the amount of noise and see a more evident spike in perplexity. We then see the perplexity return to a lower level following the event.

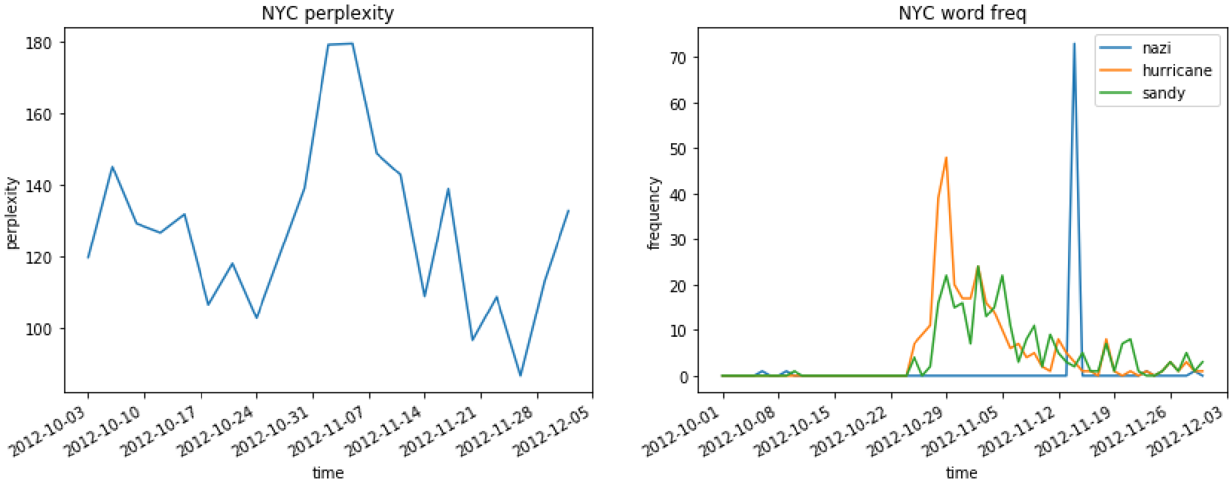




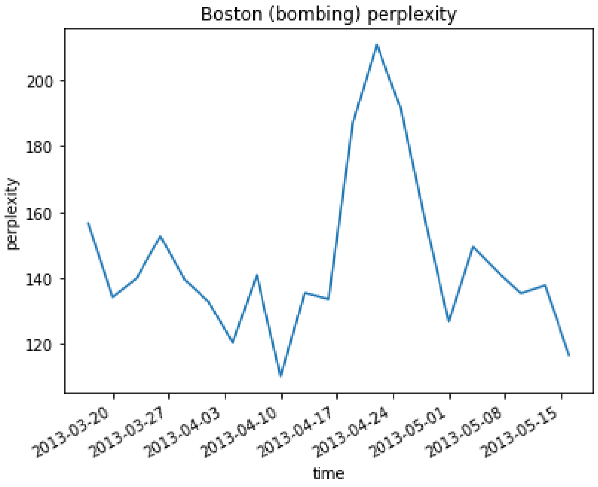
Puerto Rico: Hurricane Maria last from September 16, 2017 – October 3, 2017. We see a large increase in perplexity after the 17th, but it's not as clear as some of the other datasets. This data set has a mix of English and Spanish and it's not as clear that the perplexity is changing due to an event.

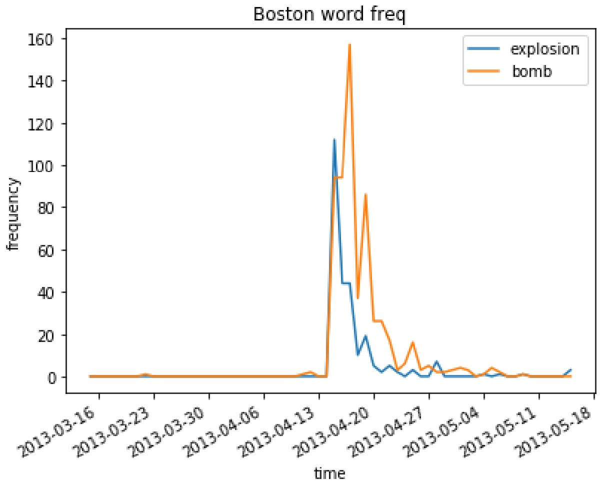


NYC: Hurricane Sandy hit NYC on Oct. 28. We see a spike around late October. And words like ‘hurricane’ and ‘sandy’ start to appear. There’s also a day when ‘nazi’ appears which may explain some of the jumps in the perplexity on the right hand size.

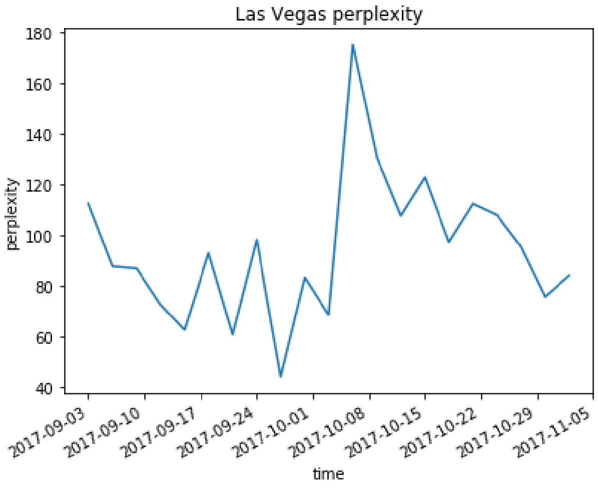


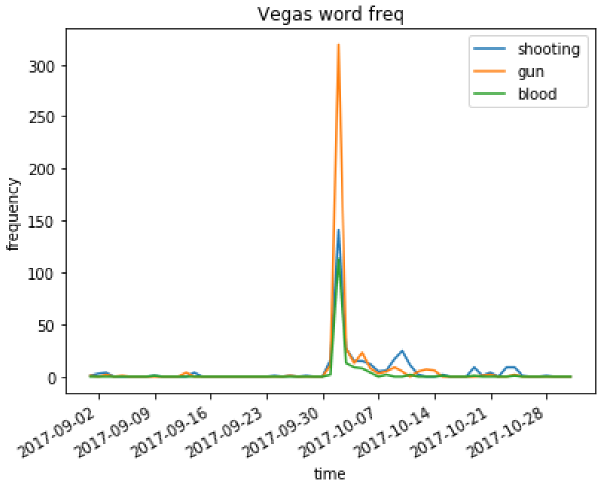
Boston: The bombing took place on April 15th. After which we see a large increase in perplexity of over 200. There’s also new words that enter the vocab at this time like ‘explosion’ and ‘bomb’.



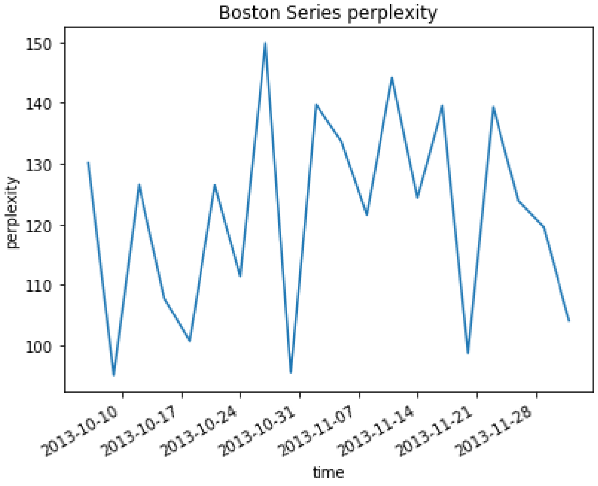


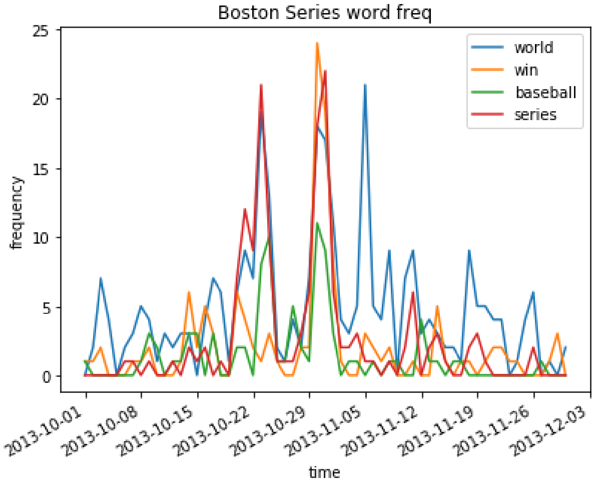
Las Vegas: A shooting took place on October 1st. Theres also a large spike in perplexity after this time. We see new words introduced like ‘shooting’, ‘gun’, and ‘blood’.





Boston: The Red Sox won the World Series in late October of 2013. There’s not much evidence of any large even taking place. This event was not unexpected in the online community and the perplexity has no clear spike. There’s lots of noise and no real evidence of something significant happening.

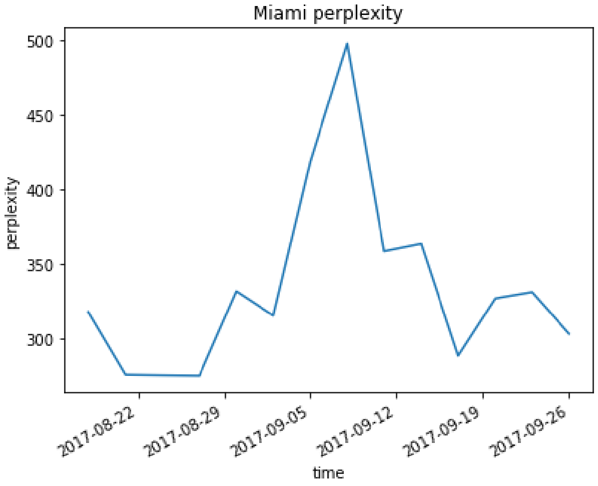


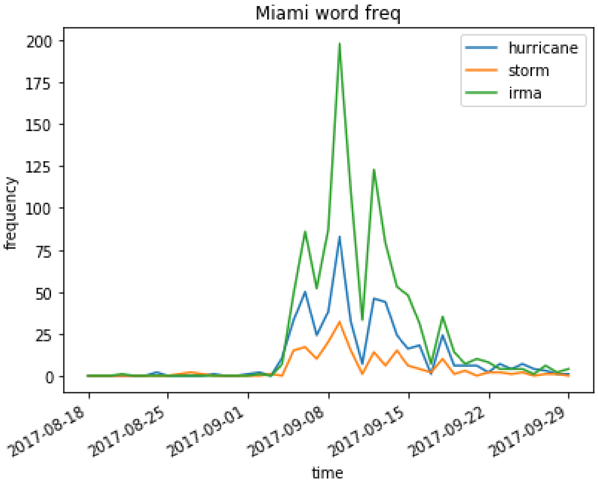


Twitter Perplexity Results

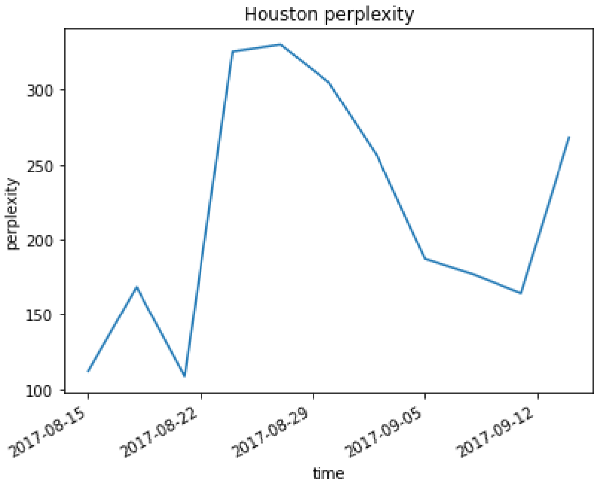
A lot of the Twitter results align very well with the results seen in the Reddit dataset.

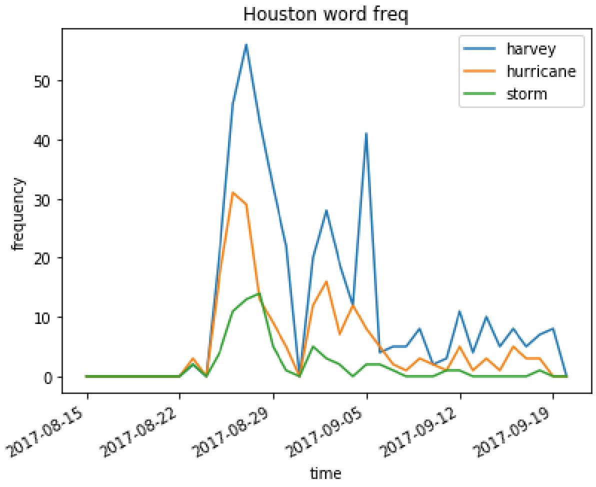
Miami saw a similar spike in hurricane-related vocabulary and perplexity at the same time frame in which the hurricane occured.



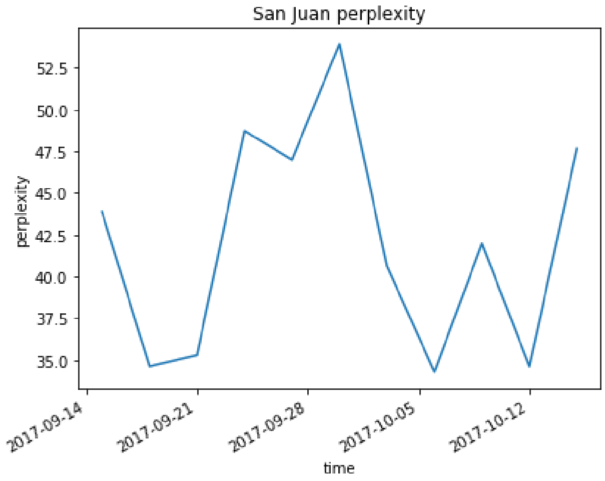


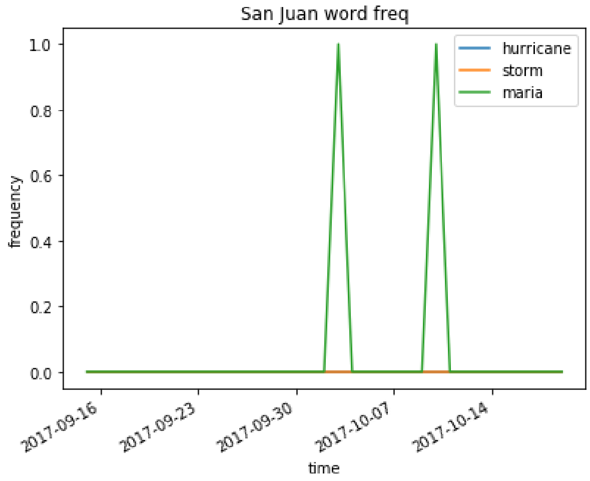
Houston experienced a similar spike during Hurricane Harvey; however its effect lasted a much longer time. In fact, Harvey is still being discussed well after it made landfall. This isn't surprising when one considers the amount of devastation that storm caused.



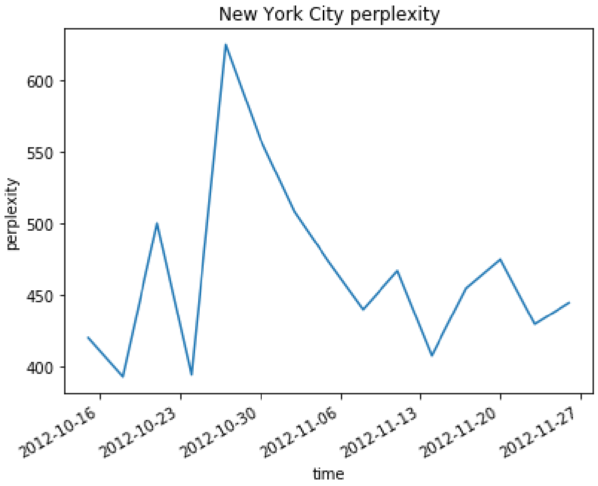


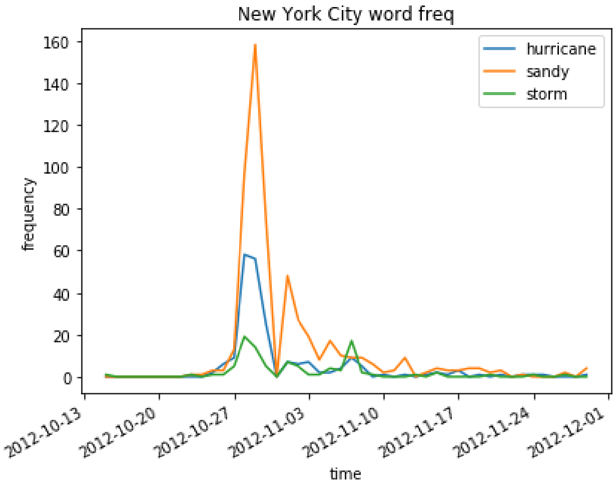
The San Juan dataset demonstrates how the model performs without a significant amount of data. A spike does seem to line up with when Maria entered the vocabulary; however other days also seem to trigger spikes randomly.



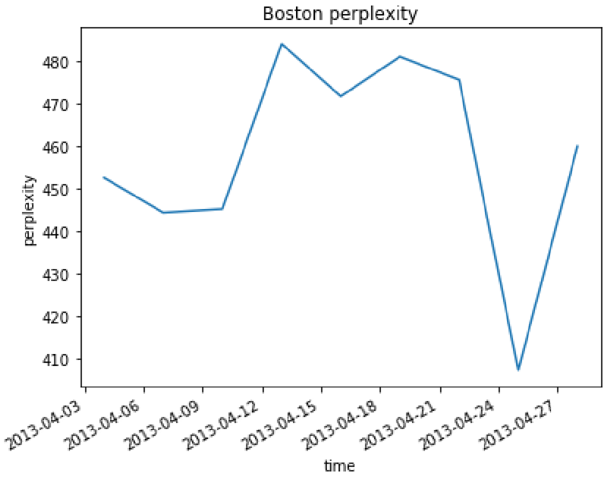


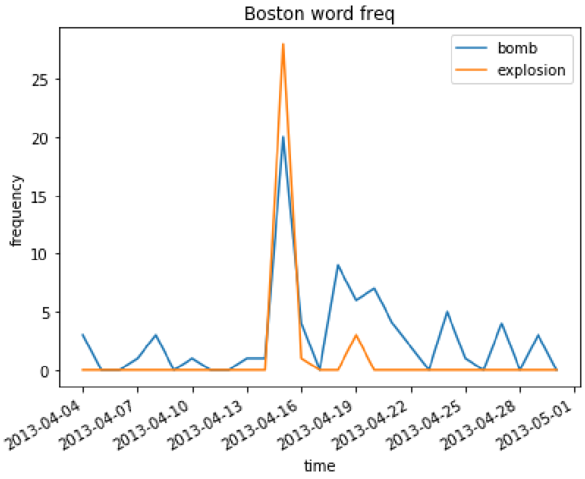
The New York City dataset demonstrates a very clear spike when Hurricane Sandy made landfall. This dataset doesn't show the same type of build-up in perplexity that other hurricanes displayed.



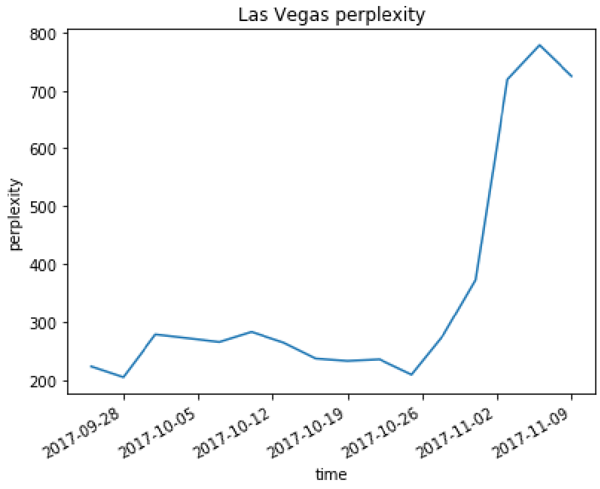


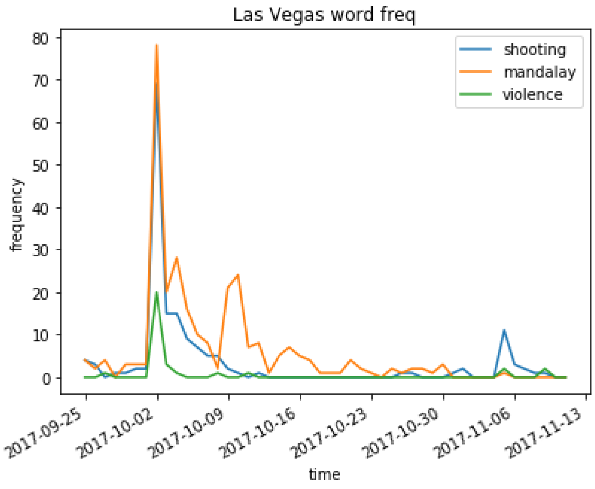
The Boston bombing incident did trigger a large spike in perplexity at the time of the event. The perplexity took quite some time to come down; however that is likely due to the lingering effects of a terror attack.





The Las Vegas dataset shows an initial spike at the time of the Mandalay Bay shooting; however it is eclipsed by the sudden spike at the end of the test window. This much larger spike is due to the Adobe MAX conference that took place in Las Vegas at that time.





NMF Topics for Twitter Data

Before Hurricane Harvey, people are concerned with traffic, but not after. There were several topics detected all about Hurricane Harvey. This may speak to the widespread damage of effect on most of Houston.

Vegas Before

Topic #0: vegas las strip north bellagio venetian cosmopolitan downtown wynn night baby center rock convention weekend welcome fabulous em paris club
Topic #1: bit ly vegastraffic xxmewb accident http blvd rd clark nb ave sb reported approaching ramp right beltway eb dr sahara
Topic #2: beautiful life lifeisbeautiful festival weekend amazing lib friends lifeisbeautifulfestival art

people gorillaz best live experience truly like fun really ready
Topic #3: com twitter pic http nfl raidernation mp raiders tour home circlepix week https oakland dlvr listing washington retweet listed jets
Topic #4: lasvegas lasvegasstrip lifeisbeautiful like usa bellagio hiring vegas strip home vivalasvegas travel jobs flamingo bar world morning place ready great
Topic #5: casino hotel resort paris mirage aria flamingo hollywood planet orleans luxor excalibur bay mandalay york rock spa hard island bellagio
Topic #6: just posted photo video like stratosphere got planet fitness bar resort cause live bay mandalay club home park won hollywood
Topic #7: time great fun thank got night amazing having good lifeisbeautiful weekend year took favorite ago ve friends saw work lunch
Topic #8: nevada north las henderson valley spring enterprise paradise good sunrise way summerlin southern lol em south look manor lets trails
Topic #9: beer untp drinking http photo ipa tenayacreek bangerbrewing ale beerhaus light hop little parisvegas double area brewing house golden company
Topic #10: day great office favorite pool week lifeisbeautiful good start like coffee national everyday days sunday swim work lets having blessed

Vegas After
Topic #0: vegas las strip bellagio venetian cosmopolitan north wynn en downtown grand good palazzo mgm vegasbaby fabulous sign welcome view weekend
Topic #1: bit ly vegastraffic xxmewb accident http blvd rd clark ave nb sb reported approaching right eb dr charleston wb west
Topic #2: swarmapp nv las henderson vegas bar center lasairport cafe pic twitter grill lunch home restaurant station burger starbucks buffet picking
Topic #3: beer untp drinking http photo ipa khourysfinewine ale hopnutsbrewing zombies bottle share tenayacreek lager stout light bangerbrewing hop bar eagle
Topic #4: lasvegas bellagio lasvegasstrip sincity en usa lv tour sema home hiring travel depechemode fremont mandalaybay vegas wynn fun circlepix dtlv
Topic #5: casino hotel resort rock luxor hard mirage aria mandalay bay spa paris planet hollywood red york excalibur rio island suite
Topic #6: just posted photo video got park like want listed palace nightclub caesars retweet center red restaurant henderson world store don
Topic #7: nevada north las paradise henderson vegas spring valley enterprise summerlin en south que morning dog home work got like sunrise
Topic #8: tonight party nightclub come lets inside hakkasanlv lo hi sunny git forecast today ready join octth clear marqueelv vip live
Topic #9: twitter pic nfl mp raidernation http raiders tour home circlepix week oakland ravens vs retweet listed dlvr listing virtual newest
Topic #10: vegasstrong welcome prayforvegas fabulous mandalay bay sign city mandalaybay vegasgoldenknights proud resort lasvegasstrip prayforlasvegas strong goldenknights game memorial vegas church

Houston (Hurricane Harvey)

Before:
Topic #0: traffic delay stop mins fwy cleared accident outbound lp inbound hwy sw katy rd ly bit stall northside nb gulf
Topic #1: swarmapp tx houston club pasadena intercontinental airport george pearland bush city brunch checkin sam sports starbucks ubercheckin automatically uber grill
Topic #2: bubly http beds baths tx st dr pearland ln pasadena bath rd houston porte ct deer la way lake bellaire
Topic #3: beer untp drinking http photo ipa flying ale hops hop saintarnold conservatoryhtx premiumdraught summer grill heights meet west casa better

Topic #4: twitter pic dlvr rt status http astros shit did way white lol people road circlepix fun george cool local try

Topic #5: houston texas southeast westside htown downtown htx north museum night en arts work sunday sundayfunday live mi la good fine

After

Topic #0: houston texas westside pray share southeast downtown byb help city htown en day like god byt prayforhouston time eastside flood

Topic #1: water bit ly high lanes traffic main affecting http fwy sb hwy wb nb lp downtown baytown eb right harris

Topic #2: bubly http beds baths beer tx st untp dr closed drinking astros shelter flooding pasadena pearland new ln porte relief

Topic #3: harvey hurricane houston got rain needs relief prayers center look today flooded friday come good george stuck open thing home

Topic #4: hurricaneharvey safe prayforhouston stay houston need park downtown htx flooded rain ready houstonstrong dry buffalo going water home getting prayers

Topic #5: just posted photo video stadium hard downtown got houston house southeast getting la didn meyerland beer untp old don say

Puerto Rico (Hurricane Maria)

Maria Before

Maria twitter data proved hard to obtain. Because of the small sample size and the Spanish language, we ran into problems and didn't end up detecting coherent topics, despite trying different models and parameters.

Topic #0: juan san argentina marquesado capital todo que hoy foto te es gracias feliz estadio dia el la del club boca

Topic #1: la hoy san es sabado siempre del juan gracias foto feliz estadio yo el lo dia club capital boca bicentenario

Topic #2: que lo es pero todo siempre yo por hoy quiero mas capital dia el juan del club boca estadio feliz

Topic #3: el hoy quiero boca yo foto es que capital feliz la juan gracias estadio lo dia del club bicentenario las

Topic #4: del bicentenario siempre estadio juan sabado san que club hoy gracias foto feliz yo es el las dia capital boca

Topic #5: te siempre quiero estadio sabado pero hoy que gracias juan dia la del el es club capital feliz foto boca

Maria After

Topic #0: juan san argentina lucianopereyra marquesado dia vida foto lo domingo quiero club como del la el es feliz capital al

Topic #1: que lo mute la para el juan foto feliz es vida domingo del como club capital argentina al dia los

Topic #2: la lucianopereyra que una mute vida san es foto feliz el dia domingo del como club capital argentina al juan

Topic #3: el domingo para dia la juan foto feliz es vida lo del como club capital argentina al las los una

Topic #4: mi vida lucianopereyra para marquesado san es feliz foto el domingo la dia del como club capital argentina al juan

Topic #5: te quiero mas vida como que feliz mute san argentina capital club juan del dia al domingo el

es foto
Topic #6: los quiero san mute es domingo juan foto feliz el del dia las como club capital argentina al la vida

Miami (Irma)

Miami produced a relatively highly coherent topic, ranked #5 among topics. People care less about going to the beach(topic #5) before after the hurricane hit.

Miami Irma before
Topic #0: miami international airport mia miamibeach southbeach downtown brickell city vacation beach home north vibes iflymia tbt like good fontainebleau miamilife
Topic #1: chance tonight storm pm hi forecast lo augth fl cloudy partly showers tue wednesday sunny storms mon sat thu tuesday
Topic #2: bit ly http sfltraffic blocked lane sb st accident nb disabled vehicle sr left ramp right express tpke cleared nwth
Topic #3: en una los mi acaba publicar las foto doral hoy este del ya por park noche tu kendall esta nuestro
Topic #4: com twitter pic http tour https home realestate keyes listing net dlvr utm_source utm_medium tecnohoy status circlepix virtual looking shop
Topic #5: beach south usa hotel miami southbeach miamibeach sunny ocean isles fontainebleau summer hollywood em hallandale north drive sea life riu
Miami Irma After
Topic #0: miami airport downtown international miamibeach mia brickell city em north tbt southbeach gardens iflymia good dade hurricaneirma night today live
Topic #1: sfltraffic bit ly http sr blocked accident expy nb lane st nwth sb rd palmetto disabled cleared vehicle ave right
Topic #2: chance storm tonight hi forecast lo pm septh fl showers sat mon storms sunday tue wednesday heights tuesday thu friday
Topic #3: en mi una los doral hoy brickell acaba publicar del por te foto este ya esta mexico para las dios
Topic #4: com twitter pic http tour home realestate https dlvr net keyes utm_medium utm_source tecnohoy listing circlepix looking virtual buyer status
Topic #5: irma hurricane hurricaneirma ready safe post huracan like miamibeach got stay relief week storm aftermath brickell southflorida thank help power

NYC (Sandy)

Sandy proved to be harder to detect in NYC twitter data.

Sandy Before
Topic #0: http instagr park street newyork hall ly bar photo dinner bit posted city square art music day hotel house theater
Topic #1: new york ny bar city square street hotel hall lincoln times cafe west shop theatre club kitchen st world grill
Topic #2: pic twitter com path event http great tonight party right dinner awesome theatre birthday better food oh work el cool
Topic #3: like feel people look really nice shit looks make girl try fucking stop away open im niggas lmaoo getting ll
Topic #4: just foursquare mayor posted photo ousted think day saw came haha did fuck face wanna oh come noticed baby literally

Topic #5: love baby izod thanks guys lmfao friend omg haha fuck true nigga great team ya damn face tho missing foreve

Sandy after
Topic #0: http instagr park center ly com sandy bit photo square night rockefeller posted bar st house dinner store city central
Topic #1: pic com twitter path http sandy true look bad line ready got finally night fun yes tree happy storm street
Topic #2: new york ny square city park jersey club times bar manhattan cafe st garden madison terminal restaurant bit hall ly
Topic #3: lol ok oh suck did right cool ass gonna hate man phone big fun omg rt got better ur happy
Topic #4: just foursquare mayor posted photo ousted want saw got life took did im ve live week day finished asked half
Topic #5: like look looks feel bitch girls shit say bitches sure eating im sound house niggas watching sounds need wtf better

Boston (Bombing)

Boston detection of the bombing had medium coherence.

Boston bombing before
Topic #0: http instagram com dinner photo house td beer day garden st east center dlvr happy little thanks untp bruins square
Topic #1: twitter pic com paxeast great lot today best did pax lmfao new hey bruins tonight wish lmao yes game restaurant
Topic #2: just got think lmao damn said did haha way home going cuz watch want don today new girl time long
Topic #3: like feel people shit ass looks really week rt come seriously ll girls bad live cause makes better look gets
Topic #4: boston ma time ly bit house http center east excited school best cambridge way restaurant paxeast college night report bad
Topic #5: know don let day wanna shit want doesn thanks mt dont time say yeah tomorrow trying probably ve won start
Topic #6: good time going day tonight look im gonna looking yeah bad think looks lmao ve today far ill right pretty
Topic #7: lol got right bro shit okay funny didn bitch ya lot come worst friends actually oh high house new team
Topic #8: love people better watching night hey best dont hate cause ve big new ll make thanks going time music pic
Topic #9: need new fuck damn make life oh stop fucking wait house yeah don cause birthday bad seriously start friends haha

Boston Bombing After
Topic #0: http instagram com bostonstrong beautiful redsox park yq fenway new prayforboston photo food day bar sunset square police dinner boston
Topic #1: pic twitter com bostonstrong little today game im happy rt gonna prayforboston home bruins night yes tonight school best seen
Topic #2: boston ma marathon house bostonstrong city manhunt way vs watertown bar great grill la police sports http bruins market center

Topic #3: like feel looks really make guy shit dont nigga stop wait yes wanna prayforboston haha lmao old cause school away

Topic #4: fenway park sox red redsox mlb astros vs night houston game lets pic instagram sunset new watching home bostonstrong best

Topic #5: just news saw did said life didn friend posted photo shit walk want gonna watching need heard work getting yes

Topic #6: love friends life miss man baby really girl news beautiful city great weird ya thank alive wish god let oh

Topic #7: don know shit want fuck wanna thingsthatirritateme let food haha tell rt alive trying new miss say http watching hate

Topic #8: good day today thanks way great oh bad ll happy big better ready know place twitter tomorrow god long dinner

Topic #9: time shit having dinner today favorite fucking tell mom friends away night lmao city year want girl make watching way

Topic #10: people think life tell pretty weird need friends open know thing new kid making way day thingsthatirritateme make said hours

Topic #11: got did phone guys say wait work report home little police new way text literally ill feel mom nigga market

Topic #12: going tonight watch trying make open rt lets birthday nice shots year today phone favorite tv cambridge city person great

Topic #13: lol really think girl mad dont fuck need thing text little fun bitch club party walk watching wish shit did

Topic #14: right work man said person literally heart amazing oh doing bad scanner wrong police ill ya starbucks fuck restaurant having

LDA Wikipedia Topics for Reddit Data

NYC (Sandy):

```
2012-10-25
   know, says, told, night, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   com, http, www, org, html, net, https, gov, php, edu
   borough, metropolitan, brooklyn, manhattan, christie, nyc, harlem, stony, bronx, quinnipiac
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   allow, required, need, needed, possible, able, provide, available, allows, allowing
   bus, express, transport, metro, transit, transportation, buses, routes, tram, route
   storm, tropical, weather, cyclone, tornado, rain, winds, depression, homes, damage
2012-10-28
   know, says, told, night, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   com, http, www, org, html, net, https, gov, php, edu
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   borough, metropolitan, brooklyn, manhattan, christie, nyc, harlem, stony, bronx, quinnipiac
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   storm, tropical, weather, cyclone, tornado, rain, winds, depression, homes, damage
   online, internet, content, facebook, media, twitter, available, social, websites, copyright
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   bus, express, transport, metro, transit, transportation, buses, routes, tram, route
   street, avenue, neighborhood, streets, downtown, ave, boulevard, corner, neighborhoods, plaza
2012-10-31
   know, says, told, night, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   borough, metropolitan, brooklyn, manhattan, christie, nyc, harlem, stony, bronx, quinnipiac
   com, http, www, org, html, net, https, gov, php, edu
   bus, express, transport, metro, transit, transportation, buses, routes, tram, route
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   street, avenue, neighborhood, streets, downtown, ave, boulevard, corner, neighborhoods, plaza
   allow, required, need, needed, possible, able, provide, available, allows, allowing
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   bridge, tunnel, traffic, bridges, crossing, lanes, arch, span, river, toll
   car, cars, driver, chevrolet, motor, speedway, toyota, drivers, nascar, auto
   online, internet, content, facebook, media, twitter, available, social, websites, copyright
   run, running, ran, robinson, walk, runs, aaa, longest, rbi, rbis
   power, lc, lag, dissipated, anantapur, dissipating, volts, impedance, capacitive, kilowatt
```

Miam (Irma):

```
2017-09-03
   know, says, told, nigh#1t, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   com, http, www, org, html, net, https, gov, php, edu
   storm, tropical, weather, cyclone, tornado, rain, winds, depression, homes, damage
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   allow, required, need, needed, possible, able, provide, available, allows, allowing
   florida, miami, cuba, orlando, tampa, cuban, havana, castro, batista, celia
   pm, bbc, sunday, saturday, friday, morning, programme, monday, thursday, tuesday
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   water, supply, dry, wet, pump, fresh, float, pumping, pumps, floating
   tells, goes, begins, takes, finds, gets, tries, help, comes, returns
   bass, custom, ups, vintage, watt, upright, dunlop, pickup, walnut, combo
   hurricane, caribbean, jamaica, status, trinidad, kingston, bermuda, indies, tobago, barbados
   tax, pay, money, price, paid, income, budget, revenue, cost, costs
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
2017-09-06
   know, says, told, night, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   com, http, www, org, html, net, https, gov, php, edu
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   storm, tropical, weather, cyclone, tornado, rain, winds, depression, homes, damage
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   florida, miami, cuba, orlando, tampa, cuban, havana, castro, batista, celia
   allow, required, need, needed, possible, able, provide, available, allows, allowing
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   tax, pay, money, price, paid, income, budget, revenue, cost, costs
   water, supply, dry, wet, pump, fresh, float, pumping, pumps, floating
   pm, bbc, sunday, saturday, friday, morning, programme, monday, thursday, tuesday
   hurricane, caribbean, jamaica, status, trinidad, kingston, bermuda, indies, tobago, barbados
```

Florida (Irma):

```
2017-09-03
   know, says, told, night, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   com, http, www, org, html, net, https, gov, php, edu
   storm, tropical, weather, cyclone, tornado, rain, winds, depression, homes, damage
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   florida, miami, cuba, orlando, tampa, cuban, havana, castro, batista, celia
   bass, custom, ups, vintage, watt, upright, dunlop, pickup, walnut, combo
   pm, bbc, sunday, saturday, friday, morning, programme, monday, thursday, tuesday
   water, supply, dry, wet, pump, fresh, float, pumping, pumps, floating
   tells, goes, begins, takes, finds, gets, tries, help, comes, returns
   registry, bayonne, curran, wallis, montserrat, nic, horden, camilo, isla, guiana
   allow, required, need, needed, possible, able, provide, available, allows, allowing
   hurricane, caribbean, jamaica, status, trinidad, kingston, bermuda, indies, tobago, barbados
2017-09-06
   know, says, told, night, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   com, http, www, org, html, net, https, gov, php, edu
   storm, tropical, weather, cyclone, tornado, rain, winds, depression, homes, damage
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   allow, required, need, needed, possible, able, provide, available, allows, allowing
   florida, miami, cuba, orlando, tampa, cuban, havana, castro, batista, celia
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   tells, goes, begins, takes, finds, gets, tries, help, comes, returns
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
```

Houston (Harvey):

```
2017-08-19
   know, says, told, night, said, come, going, away, things, room
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   announced, got, wanted, decided, revealed, going, week, good, start, months
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   com, http, www, org, html, net, https, gov, php, edu
   texas, houston, austin, dallas, hamilton, antonio, tyler, tx, worth, paso
   allow, required, need, needed, possible, able, provide, available, allows, allowing
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   racing, gt, championship, formula, prix, motorsport, driver, ferrari, porsche, wrc
   dish, bread, cheese, sauce, cake, beef, potato, dishes, pizza, michelin
2017-08-22
   know, says, told, night, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   storm, tropical, weather, cyclone, tornado, rain, winds, depression, homes, damage
   allow, required, need, needed, possible, able, provide, available, allows, allowing
   com, http, www, org, html, net, https, gov, php, edu
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   texas, houston, austin, dallas, hamilton, antonio, tyler, tx, worth, paso
   water, supply, dry, wet, pump, fresh, float, pumping, pumps, floating
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   dam, reservoir, flood, irrigation, water, dams, floods, flooding, capacity, drainage
2017-08-25
   know, says, told, night, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   com, http, www, org, html, net, https, gov, php, edu
   storm, tropical, weather, cyclone, tornado, rain, winds, depression, homes, damage
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   dam, reservoir, flood, irrigation, water, dams, floods, flooding, capacity, drainage
   texas, houston, austin, dallas, hamilton, antonio, tyler, tx, worth, paso
   water, supply, dry, wet, pump, fresh, float, pumping, pumps, floating
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   allow, required, need, needed, possible, able, provide, available, allows, allowing
```

Puerto Rico (Maria):

```
2017-09-19
   know, says, told, night, said, come, going, away, things, room
   com, http, www, org, html, net, https, gov, php, edu
   announced, got, wanted, decided, revealed, going, week, good, start, months
   se, pa, si, ne, ba, ka, ni, cohen, ra, ko
   josé, carlos, juan, luis, cruz, miguel, maría, antonio, fernando, garcía
   el, del, nacional, las, los, universidad, plaza, colegio, casa, monte
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   user, users, app, android, support, features, allows, available, software, application
   online, internet, content, facebook, media, twitter, available, social, websites, copyright
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   en, iceland, er, viking, det, med, icelandic, karin, den, smiley
   la, rochelle, ogier, amis, musique, remi, stanislas, ninh, après, étoile
   puerto, rico, juan, reyes, rivera, torres, rican, ppd, hernández, rl
   storm, tropical, weather, cyclone, tornado, rain, winds, depression, homes, damage
2017-09-22
   know, says, told, night, said, come, going, away, things, room
   com, http, www, org, html, net, https, gov, php, edu
   online, internet, content, facebook, media, twitter, available, social, websites, copyright
   announced, got, wanted, decided, revealed, going, week, good, start, months
   el, del, nacional, las, los, universidad, plaza, colegio, casa, monte
   josé, carlos, juan, luis, cruz, miguel, maría, antonio, fernando, garcía
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
   puerto, rico, juan, reyes, rivera, torres, rican, ppd, hernández, rl
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   se, pa, si, ne, ba, ka, ni, cohen, ra, ko
   user, users, app, android, support, features, allows, available, software, application
   network, mobile, communications, networks, communication, wireless, nokia, internet, services
   bass, custom, ups, vintage, watt, upright, dunlop, pickup, walnut, combo
   hurricane, caribbean, jamaica, status, trinidad, kingston, bermuda, indies, tobago, barbados
   en, iceland, er, viking, det, med, icelandic, karin, den, smiley
   la, rochelle, ogier, amis, musique, remi, stanislas, ninh, après, étoile
```

Vegas (shooting):

```
2017-10-01
   know, says, told, night, said, come, going, away, things, room
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   announced, got, wanted, decided, revealed, going, week, good, start, months
   killed, attack, killing, attacks, injured, wounded, massacre, security, forces, dead
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   rifle, pistol, automatic, weapon, bullet, weapons, carry, barrel, firearms, bullets
   com, http, www, org, html, net, https, gov, php, edu
   allow, required, need, needed, possible, able, provide, available, allows, allowing
   hotel, las, vegas, nevada, inn, resort, casino, hotels, spa, paradise
   gun, guns, battery, inch, fired, firing, artillery, armor, mk, turret
   violence, victims, abuse, rape, crimes, sexual, violent, victim, punishment, forced
   shot, shooting, shoot, broken, shots, arrow, bow, hoffman, arrows, shotgun
   police, crime, officer, officers, investigation, inspector, detective, criminal, sheriff, enforcement
   said, stated, reported, claimed, saying, statement, described, claims, announced, told
2017-10-04
   know, says, told, night, said, come, going, away, things, room
   announced, got, wanted, decided, revealed, going, week, good, start, months
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   com, http, www, org, html, net, https, gov, php, edu
   hotel, las, vegas, nevada, inn, resort, casino, hotels, spa, paradise
   shot, shooting, shoot, broken, shots, arrow, bow, hoffman, arrows, shotgun
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   reception, reviews, review, wrote, gave, said, critics, positive, critical, described
   killed, attack, killing, attacks, injured, wounded, massacre, security, forces, dead
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
```

Boston (bombing):

```
2013-04-14
   know, says, told, night, said, come, going, away, things, room
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   com, http, www, org, html, net, https, gov, php, edu
   announced, got, wanted, decided, revealed, going, week, good, start, months
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   killed, attack, killing, attacks, injured, wounded, massacre, security, forces, dead
   boston, massachusetts, lincoln, rhode, springfield, providence, salem, worcester, booth, randolph
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   online, internet, content, facebook, media, twitter, available, social, websites, copyright
   good, hope, desire, sense, happiness, wish, passion, respect, pleasure, spirit
   friend, friends, love, mother, relationship, young, woman, father, girl, lives
   damage, damaged, accident, caused, destroyed, struck, lost, occurred, hit, recovered
2013-04-17
   know, says, told, night, said, come, going, away, things, room
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   announced, got, wanted, decided, revealed, going, week, good, start, months
   com, http, www, org, html, net, https, gov, php, edu
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   killed, attack, killing, attacks, injured, wounded, massacre, security, forces, dead
   police, crime, officer, officers, investigation, inspector, detective, criminal, sheriff, enforcement
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   online, internet, content, facebook, media, twitter, available, social, websites, copyright
   boston, massachusetts, lincoln, rhode, springfield, providence, salem, worcester, booth, randolph
   news, media, report, reporter, coverage, journalist, journalists, journalism, graham, correspondent
```

Boston (World Series):

```
2013-10-28
   know, says, told, night, said, come, going, away, things, room
   guy, fun, happy, mad, big, pretty, dad, funny, parody, dirty
   love, night, girl, let, baby, good, live, kiss, beautiful, come
   announced, got, wanted, decided, revealed, going, week, good, start, months
   boston, massachusetts, lincoln, rhode, springfield, providence, salem, worcester, booth, randolph
   scene, scenes, featured, filmed, audience, production, filming, shown, footage, interview
   com, http, www, org, html, net, https, gov, php, edu
   dish, bread, cheese, sauce, cake, beef, potato, dishes, pizza, michelin
   fact, case, certain, question, necessary, matter, claim, possible, considered, principle
   baseball, league, giants, pitcher, sox, games, hit, mlb, professional, runs
```

```
allow, required, need, needed, possible, able, provide, available, allows, allowing
tax, pay, money, price, paid, income, budget, revenue, cost, costs
online, internet, content, facebook, media, twitter, available, social, websites, copyright
bus, express, transport, metro, transit, transportation, buses, routes, tram, route
students, student, campus, academic, alumni, activities, universities, volleyball, colleges, arts
```

Snapshot Language Modeling Process

Process

1. Get data for each day for each disaster
2. Preprocess the data
3. Break the data into training and test sets using windowing* methodology specified below. For each 3 day window, use our baseline trigram model with add K smoothing or Kesser Ney. Measure perplexity against the held out test data. **
4. Observe how perplexity changes over time prior to, during, and after a disaster.
5. We will look at increases in uses of certain words over time (e.g. Irma, Harvey, Maria, etc)

*We used several methodologies for windowing that we are considering and have started on implementing.

1. Implementation 1:
    a. Step 1: Train on a month of data about 2 weeks before an event. Test on 3 days after training data. Get result.
    b. Step 2: Train on additional 3 days of data. Test on 3 days after training data. Get result.
    c. Repeat step 2, until 30 days after event.
2. Implementation 2:
    a. Step 1: Train on a month of data about 2 weeks before an event. Test on 3 days after training data. Get result.
    b. Step 2: Keep the same training data and test on 3 days after previous test data. Get result.
    c. Repeat step 2, until 30 weeks after event.
3. Implementation 3:
    a. Train on 3 days of data, measure perplexity on following 3 days of data
    b. Train on 3 days following previous train data, test on following 3 days of data
    c. Repeat step 2, until 30 days after event

**for reddit, we tried using the full post and the first 15 words of each posts, since length of post is related to perplexity.

Baseline: we are using a trigram language model with add K smoothing as our baseline.
Improvements: try both bigram and trigram models with KN

Topic Modeling Process

Process 1 (Baseline)

1. For each disaster, create an topic LDA
2. Choice of hyperparameters by testing different values of K
3. Compare topics produced by different topic models for different disasters

Baseline: We are using the approach from (Hoffman et al, 2010) for our baseline as coded in the gensim package.
Improvements: author topic modeling(Rosen et al, 2004), or hierarchical LDA from gensim

Process 2

1. Train a nmf model from scikitlearn with generalized Kullback Leiber divergence

2. Using tfidf to give higher weighting to more important words
3. Get rid of duplicates
4. Generate topics before and after a disaster
5. Compare the topics before and after and see if modeling picked out coherent topics.

Process 3

1. Train an LDA on Wikipedia topics using gensim package
2. Generate topic vectors over 3 day windows of each Reddit data set
3. Evaluate:
    a. Look at the top topics and 10 highest probability words for each topic to see if selected topics are relevant
    b. Look at Jensen-Shannon distance to see if the distance between the 3 day windows of topics changes enough to be helpful