Black Lives Matter: Clustering as a Way of Analyzing Online Discussions of Race

Omar Sow, CS 229 Fall 2017

Abstract:

The intention of this project is to investigate ways that clustering might be used as a tool in the otherwise largely qualitative field of critical race studies. This stems from interest in Beyond the Hashtag*, a report examining online discussion of the Black Lives Matter movement. This investigation found that clustering may provide a viable starting point for research into online discussions, by guiding focus towards contentious time periods (by comparing cluster sizes), and examining key features of different clusters. The clustering results were rarely clean, but demonstrated adequately that this should be a research step for qualitative investigators in digital spaces.*

Introduction:

This project explored how machine learning methods can be used to understand or further explore a very particular subset of Twitter. Specifically considering Black Twitter, the mediated cultural conversation of Black Americans on Twitter [1]. How can machine learning help further understand conversations taking place in that space. There have been various technical works investigating sentiment analysis in the context of Twitter, and there has been social science research into Black Twitter, but I fail to find the intersection of the two.

One example of this social science research was the report *Beyond the Hashtag*, analyzing 41 million tweets from June 2014 – May 2015 that included one or more of a set of hashtags circulated surrounding Black Lives Matter and the events of Fergusson [2]. The report has a sophisticated analysis but is hampered, I believe, by the limits of its tools and the size of its dataset, which is better suited to computer science than narrative-focused social science.

The intent is to expand on the work done in *Beyond the Hashtag* to explore the groups involved in discussions around the topic of Black Lives Matter. Authors of the piece were able to do some network analysis based on Twitter user connections, but did little to no analysis of tweet text. This project investigated whether programmatic text analysis through clustering provides new insights, or displays results matching background knowledge, about this topic.

Related Work:

In forming my approach to this project, I examined past attempts to use unsupervised learning in Twitter text analysis. This has provided an outline of various phases of development. Essentially, I want to produce a way to computationally model the expressed feelings/opinions/allegiances of those participating in conversations around Black Lives Matter during the events of Fergusson, as just one example of the ongoing national and international civil rights struggles.

First, I examined “Opinion Mining on Twitter Data using Unsupervised Learning Technique”, which compared spectral clustering as a sentiment classifier to supervised learning methods to a set of tweets. The results were positive within the constraints of this experiment, and its structure has been instrumental in organizing the steps of this paper’s own research and development [3].

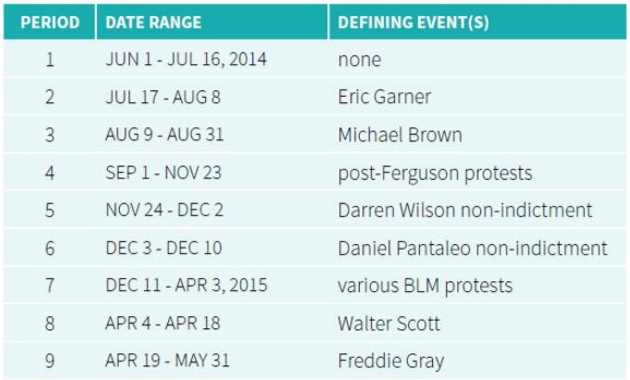
Similarly, there has been thorough research into the larger sentiment analysis problem in a Twitter context, including three past CS229 projects, which used proxies from the data (such as emojis) as labels to take a supervised learning approach. However, like this project, they were interested in modelling the conversations taking place. These papers included “Analyzing Twitter Sentiment of the 2016 Presidential Candidates”, “Twitter US Airline Recommendation

Prediction", and "Social Unrest: Classification and Modeling" [4-6]. The last example listed was particularly similar in approach to this paper. All found positive results, indicating that sentiment analysis on tweet content could be a rich field of inquiry. This motivated this project as a plausibly fruitful inquiry.

Finally, "Topical Clustering of Tweets" examined how accurately KMeans can cluster tweets, and provided a framework for my way of measuring and reporting results by taking top weight features of each cluster [7].

Dataset and Features

The authors of Beyond the Hashtag have publicly released the entirety of their Twitter dataset (in the form of tweet IDs, which I 'rehydrated' through the Twitter API to get full tweet objects).



PERIOD	DATE RANGE	DEFINING EVENT(S)
1	JUN 1 - JUL 16, 2014	none
2	JUL 17 - AUG 8	Eric Garner
3	AUG 9 - AUG 31	Michael Brown
4	SEP 1 - NOV 23	post-Ferguson protests
5	NOV 24 - DEC 2	Darren Wilson non-indictment
6	DEC 3 - DEC 10	Daniel Pantaleo non-indictment
7	DEC 11 - APR 3, 2015	various BLM protests
8	APR 4 - APR 18	Walter Scott
9	APR 19 - MAY 31	Freddie Gray

Figure 1: Dates comprising each time period. Chart taken from Beyond the Hashtag report

Looking at Figure 1, taken directly from *Beyond the Hashtag*, I have split up the twitter data accordingly, into 9 sections, defined by events that shaped discourse around Black Lives Matter from 2014-2015.

Since clustering of all 41 million tweets in database would be computationally difficult. I took 1,000 tweets from each time period.

<u>Preprocessing</u>

I used the Tweet preprocessing script from Stanford GloVe library [8], modified to include the following preprocessing procedure:

- Removing stop words (from a short list in a text file)
- Ignore retweets
- Stemming words using the Porter stemming algorithm [9]
- Repeated words
- Replace hashtags with a tag and the content ("#fergusson" → "fergusson <hashtag>")

<u>Feature extraction</u>

Given tweet text, I need to tokenize it, and store relevant features in a dictionary object. Other papers seem to have used mixes of unigram and bigram feature extraction. I implemented the code for extraction of both these features from each tweet.

Methods

<u>KMeans Clustering</u>

KMeans functions by taking a dataset and some number of clusters to find (referred to as k). Given that, it initializes k clusters to be random points in the data. Then it runs a two-part iterative algorithm to minimize the following cost function.

$$J(c, \mu) = \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

[10]

This is accomplished by repeating, until convergence, the process of assigning all points to the nearest cluster, then recalculating each cluster to the average of all its assigned points.

I implemented a KMeans clustering algorithm, which is nearly identical to what was developed for the CS229 homework assignment. This allowed for more control over the output, which was important as I determined whether this was a worthwhile project. I also employed the Scikit-Learn library's implementation for efficiency's sake [11].

Spectral Clustering

In order to investigate an algorithm beyond the scope of the class, I also implemented spectral clustering. This method is an extension of KMeans that builds a similarity matrix for the data, capturing distances between points, then normalizes it and runs clustering on the rows of that matrix. The program takes the top eigenvectors of that similarity matrix, and uses dimensionality reduction to cluster. As an approach, it can outperform KMeans [12].

I compared the two clustering algorithms, looking purely at distortion, which can be visualized in Figure 2.

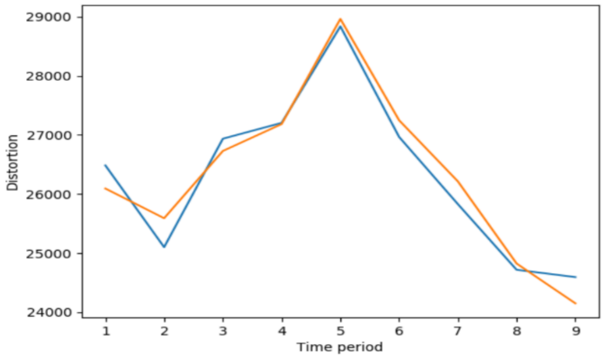


Figure 2: Distortion resulting from KMeans (blue) and Spectral (yellow) clustering across time periods.

On this dataset (given the size and nature of the data), both algorithms resulted in roughly equivalent distortion. Since the focus of the project is not on minimizing distortion, but investigating general potential uses of clustering, I moved forward with the simpler KMeans.

Choosing K:

Background knowledge, based off the *Beyond the Hashtag* report, posits that there are four large groups participating in this conversation: supporters of Black Lives Matter, people opposed to Black Lives Matter, mainstream news media, and unaligned parties.

However, in order to investigate a potentially more rigorous way to determine the k value, I attempted to employ the 'elbow method'. This entailed running clustering, and comparing distortion, for a range of cluster sizes.

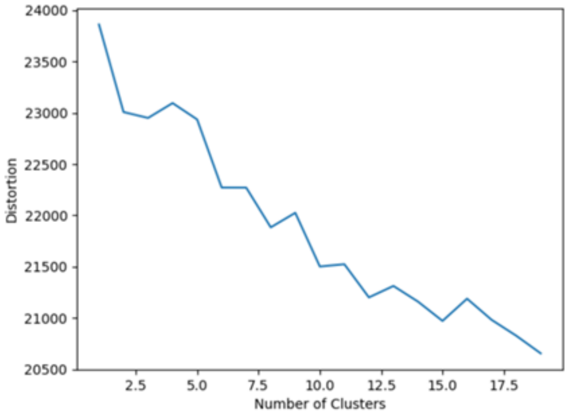


Figure 3: Distortion when running KMeans on 5,000 tweets and different numbers of clusters

The idea behind the elbow method is that, although increasing number of clusters will continue to lead to lowered distortion (and I attribute the jagged quality of the curve in Figure 3 simply to the random initialization effects). However, the parameters of this project are that I want to find ways that clustering can provide meaningful qualitative results upon convergence, and within a range of cluster sizes that might prove meaningful in any way (likely no more than 5), the results seemed inconclusive. Therefore, I decided to perform clustering with k = 4.

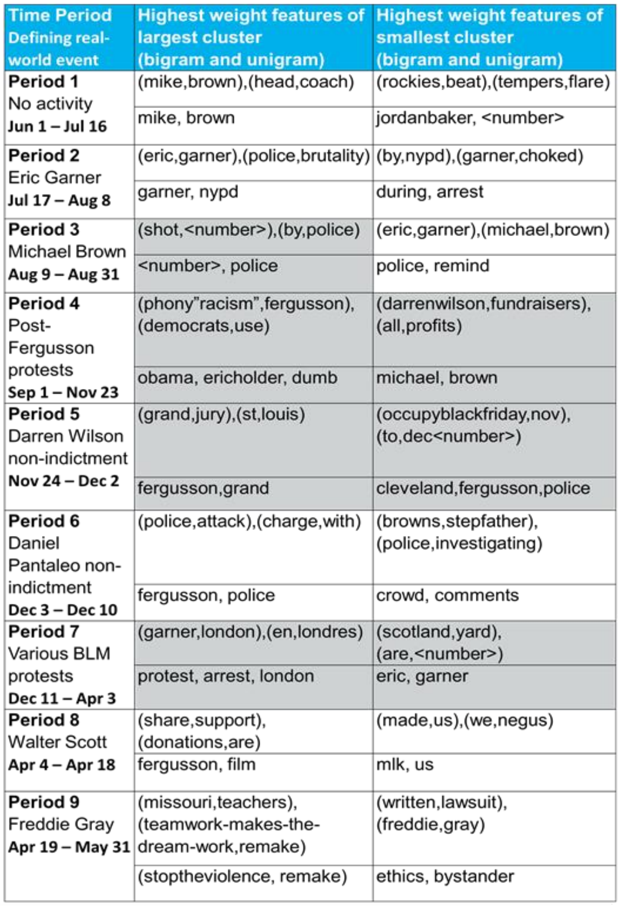
Results

The Most Important Features of the Clusters

Following the development of the clustering program, a way for results to be displayed, and syncing up my own implementations with Scikit-Learn where appropriate, results in Figure 4 were obtained.

This table shows the highest weight features (both bigram and unigram) of the largest and smallest clusters formed after running KMeans for the data subset from each period. Shaded in grey are the most potentially useful results. We notice that in all non-shaded cells, the topic of discussion of clusters was similar to the description (based on background knowledge), of the real-world event that defines the time period. In the grey cells, we see particularly focused discussions.

In Period 3, the largest cluster was focused on discussion of the number of times that police shot Michael Brown. In Period 7, both clusters were concerned with a particular Black Lives Matter protest in London. In period 4, we see a split between critics of the Black Lives Matter protests (who might do things such as call former President Obama and his Attorney General, Eric Holder, dumb; or run fundraisers for Darren Wilson. I can deduce this connection from the features based off domain knowledge, which is exactly what a qualitative researcher in the field would be able to do.



Time Period Defining real-world event	Highest weight features of largest cluster (bigram and unigram)	Highest weight features of smallest cluster (bigram and unigram)
Period 1 No activity **Jun 1 – Jul 16**	(mike,brown),(head,coach)	(rockies,beat),(tempers,flare)
	mike, brown	jordanbaker, <number>
Period 2 Eric Garner **Jul 17 – Aug 8**	(eric,garner),(police,brutality)	(by,nypd),(garner,choked)
	garner, nypd	during, arrest
Period 3 Michael Brown **Aug 9 – Aug 31**	(shot,<number>),(by,police)	(eric,garner),(michael,brown)
	<number>, police	police, remind
Period 4 Post-Fergusson protests **Sep 1 – Nov 23**	(phony"racism",fergusson), (democrats,use)	(darrenwilson,fundraisers), (all,profits)
	obama, ericholder, dumb	michael, brown
Period 5 Darren Wilson non-indictment **Nov 24 – Dec 2**	(grand,jury),(st,louis)	(occupyblackfriday,nov), (to,dec<number>)
	fergusson,grand	cleveland,fergusson,police
Period 6 Daniel Pantaleo non-indictment **Dec 3 – Dec 10**	(police,attack),(charge,with)	(browns,stepfather), (police,investigating)
	fergusson, police	crowd, comments
Period 7 Various BLM protests **Dec 11 – Apr 3**	(garner,london),(en,londres)	(scotland,yard), (are,<number>)
	protest, arrest, london	eric, garner
Period 8 Walter Scott **Apr 4 – Apr 18**	(share,support), (donations,are)	(made,us),(we,negus)
	fergusson, film	mlk, us
Period 9 Freddie Gray **Apr 19 – May 31**	(missouri,teachers), (teamwork-makes-the-dream-work,remake)	(written,lawsuit), (freddie,gray)
	(stoptheviolence, remake)	ethics, bystander

Figure 4: Table of highest weight features of the largest and smallest cluster from each time period.

<u>Cluster Size</u>

In Figure 5, we see the size of the ith-largest clusters for *i* from 1 (in green) to 4 (in red). The main point of interest here is that we can infer that when a topic is very controversial, the participants will split into more distinct group, resulting in larger clusters (i.e. a smaller difference between the largest and second largest cluster).

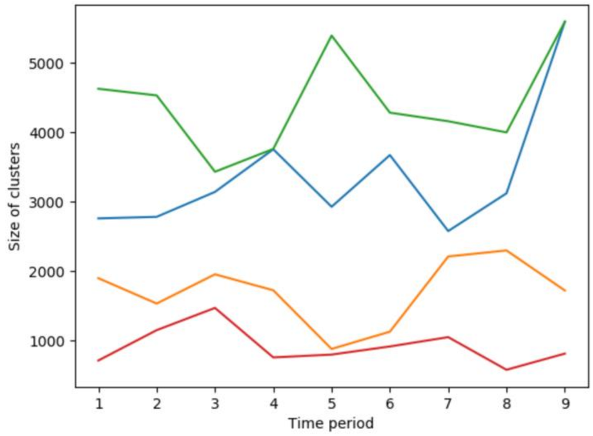


Figure 5: Size of i-th largest cluster size across time periods. Green line = largest cluster, Blue = second largest, yellow = third largest, red = smallest

We see this at time period 4, for instance. With background knowledge, this makes sense since post-Fergusson protests led to extremely controversial debates, both in digital space and mass media. In contrast, time periods where the debate was less controversial, and there would have been larger presence of unanimous groups, might include time period 5, which marked the non-indictment of Darren Wilson, and led to widespread outrage by the online Black community.

Discussion

Overall, this project was about thinking of ways to apply this class' concepts to a completely different field. I was less concerned with specifics of the clustering, such as distortion, choosing instead to focus on ways that these results could align with domain knowledge as to what was happening at the time. This is sort of a reverse engineering approach, which led to results that were consistent with real-world events.

As such, moving forward, this method could be applied early on in the research process when examining digital conversations about race, and serve as a guiding tool for research. A researcher could know to focus on points where there might be more disagreement, and therefore richer analyses, by looking at cluster size. They might also be able to discover specific sub-conversations by examining the cluster results.

These two options could provide a set of quantifiable analyses for qualitative researchers to employ when examining digital space, where the sheer amount of data is often incompatible with traditional research methods.

Further Work

There is great potential for future work on this project.

First, it could be applied, using more powerful hardware than I have, to cluster on larger subsets of the data, and explore the results of more fully considering the dataset in its entirety.

Second, with more time it could be useful to attempt initializing clusters with tweets that we determined, based off background knowledge, to belong to different particular discussions. For example, initialize one cluster with a tweet by a Black Lives Matter founder, and another with a tweet by Fox News, then compare results of the clustering.

References

[1] A. Brock, "From the Blackhand Side: Twitter as a Cultural Conversation", Taylor & Francis, 2012. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08838151.2012.732147?journalCode=hbem20.

[2] M. Clark, D. Freelon and C. McIlwain, "Beyond the Hashtag", Cmsimpact.org, 2016. [Online]. Available: http://cmsimpact.org/wp-content/uploads/2016/03/beyond_the_hashtags_2016.pdf

[3] M. Unnisa and A. Amin, "Opinion Mining on Twitter Data Using Unsupervised Learning", Semantics Scholar, 2016. [Online]. Available: https://pdfs.semanticscholar.org/3e89/06938726a44698c0711b67798cf0ce9ca30c.pdf

[4] T. Patanam, D. Saadati and F. Uraizee, "Social Unrest: Classification and Modeling, 229", Cs229.stanford.edu, 2016. [Online]. Available: http://cs229.stanford.edu/proj2016/report/social-unrest-classification.pdf.

[5] D. Chinn, A. Zappone and J. Zhao, "Analyzing Twitter Sentiment of the 2016 Presidential Candidates", cs229.stanford.edu, 2017. [Online]. Available: https://web.stanford.edu/~jesszhao/files/twitterSentiment.pdf.

[6] X. Duang and T. Ji, "Twitter US Airline Recommendation Prediction", cs229.stanford.edu, 2016. [Online]. Available: http://cs229.stanford.edu/proj2016spr/report/042.pdf. [Accessed: 15- Dec- 2017].

[7] K. Dela Rosa, R. Shah, B. Lin, A. Gershman and R. Fredericking, "Topical Clustering of Tweets", Cs.cmu.edu, 2017. [Online]. Available: http://www.cs.cmu.edu/~encore/sigir_swsm2011.pdf.

[8] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

[9] "Porter Stemming Algorithm", Tartarus.org, 2017. [Online]. Available: https://tartarus.org/martin/PorterStemmer/. [Accessed: 15- Dec- 2017].

[10] A. Ng, "CS229 Lecture notes 7", Cs229.stanford.edu, 2017. [Online]. Available: http://cs229.stanford.edu/notes/cs229-notes7a.pdf. [Accessed: 15- Dec- 2017].

[11] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[12] A. Ng, M. Jordan and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm", Andrew Ng, 2017. [Online]. Available: http://www.andrewng.org/portfolio/on-spectral-clustering-analysis-and-an-algorithm/