Clustering as a Tool for Analyzing Online Discussions of Race

Omar Sow | osow@stanford.edu | December 12, 2017 | CS229 | Team-ID: 612

Question

How can clustering analysis serve to enrich research into conversations about race in the U.S. taking place in digital spaces, specifically Twitter?

Previous Qualitative Work

I aimed to expand on the sort of work done in the paper *Beyond the Hashtag,* exploring the use of hashtags associated with the Black Lives Matter movement [1,2]. They performed network analysis and qualitative analysis, but little text analysis on their dataset, which was publicly released.

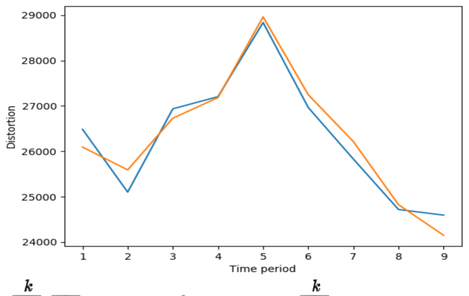
Goal of this project is to compare clustering results across the year, with the intent of identifying ways that unsupervised learning analysis could provide insights into these discussions.

Data

Given IDs for all tweets containing hashtags related to BLM, collected between June 2014 to May 2015. I split up data into 9 periods, based off previous qualitative work, and capturing defining events in the movements history (see Results). The complete dataset comprised over 41 million tweets, of which I took 5,000 from each time period.

Clustering Algorithms

Throughout experiments, compared simple KMeans clustering to Spectral Clustering [3,5]. Found qualitatively similar results and chart here shows **distortion comparison** of both.



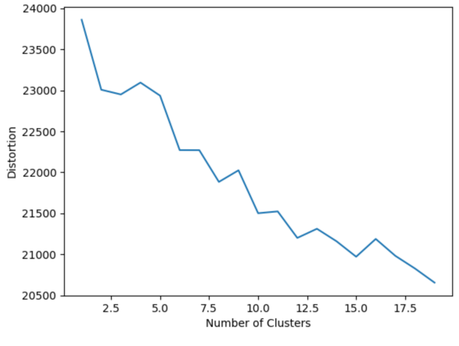
Goal:

Give set of tweet assignments S s.t.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

Picking K

Using the “elbow method”, selected the optimal k number of clusters. Based off preliminary analysis using a set of 10,000 tweets, looking at distortion resulting from different k-values. Result is inconclusive, used background knowledge **k = 4**



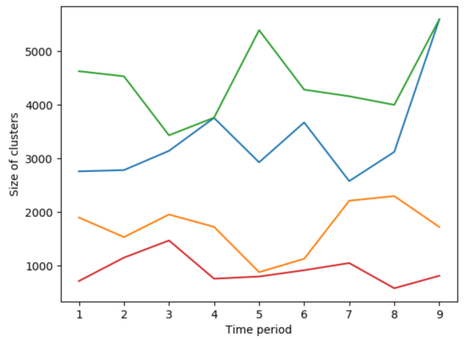
Process

Stage 1	Stage 2	Stage 3	Stage 4
Preprocessing of tweet strings to remove irrelevant information [4]	**Feature extraction** from tweets, using sparse vector representation of uni & bi-grams	**K-Means** clustering (both my own implementation and sklearn)	**Explore other clustering** options (spectral and DBSCAN)

Results

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

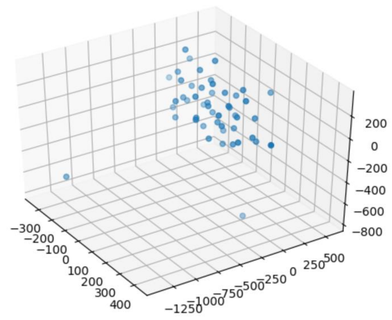
Discussion



Plot of size of clusters at each period. **Each line corresponds to the i-th largest cluster**.

No consistency in what comprises each cluster.

As a visualization, highlights time periods that were more controversial (3 and 4) or less controversial (5 and 7).



After dimensionality reduction, a display of 50 principal components of the data, demonstrating most data **clustered close together**, with some outliers.

- Divergence and convergence of clusters at controversial and less controversial periods, respectively
- Specific focuses by certain subgroups on otherwise ignored events (highlighted results in grey on Results table)
- Clustering may provide effective research starting point

Further Work

- Explore use of specifying certain tweets as initial clusters
- Expand portion of dataset to be used with more powerful computing

Works Cited

[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] Stanford Tweet preprocessing script: https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb

[5] 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/