What can the people in New York City possibly have in common with those in Scottsdale Arizona?

Can spatiotemporal techniques be used to extract local knowledge from global injustice data?

by

Laura Hoyte

Ryerson University: Capstone Project

Background Information

Different cities – population, other

cities	total number of tweets	total population (000s)	median age	mean income per_capita (000s)	house_median_income (000s)	per_english_other (%)	per_white (%)
ATLANTA GA	**30,613**	448.9	33.4	37.2	47.5	10.2	40.0
COLUMBIA SC	2,317	132.0	28.1	24.7	41.3	7.8	51.7
SCOTTSDALE AZ	540	227.5	**46.3**	52.2	73.3	13.2	88.4
CINCINNATI OH	3,277	297.4	32.5	25.6	33.6	7.4	51.1
COLUMBIA MO	2,630	115.4	26.9	26.8	44.9	9.7	78.7
BURLINGTON VT	673	42.6	26.8	24.7	44.7	13.4	**86.2**
BEVERLY HILLS CA	**474**	**34.7**	42.3	**84.6**	**97.3**	**52.2**	82.2
MINNEAPOLIS MN	7,365	400.0	31.9	32.6	51.5	21.1	65.3
CHAPEL HILL NC	939	58.8	**25.7**	37.9	62.2	20.9	72.8
SANTA CRUZ CA	618	62.8	28.7	30.4	62.2	25.2	78.3
NEW YORK NY	**30,400**	**8,426.7**	35.8	33.1	53.4	49.1	43.3
COLUMBUS OH	7,320	824.7	32.0	25.0	45.7	14.4	61.5
DAYTON OH	1,670	141.4	33.4	**16.7**	**27.7**	6.7	54.1
MONTGOMERY AL	699	203.0	34.4	24.4	42.9	**5.7**	**36.1**
PROVIDENCE RI	1,496	178.7	29.3	22.3	37.5	48.1	51.1
ROCHESTER NY	2,565	210.7	31.0	19.2	30.9	19.8	45.05

The problem?

Extracting positive sentiment from geographically diverse populations in space and time– based on a polarizing negative event (injustice shooting of Michael Brown in Ferguson)

Topic modelling at the global level does not always

- *Give high coherent topics*
- *Reflect any consistent overall sentiment consensus*

Modelling at the “local” level presents a greater opportunity to extract anomalous sentiments

Previous research

[1]Tweets can be classified into interesting clusters regardless of content

[2]It relies on count data and cylindrical windowing techniques

Any unusual activity in the space and time domains will be reflected in the clusters created

Clustering reduces the dataset and simultaneously creates a clearer picture of sentiment (geographical)

[1] W. Tao Cheng*, "Event Detection using Twitter: A Spatio-Temporal Approach
[2] R. H. J. H. R. A. F. M. Martin Kulldorff, "A Space–Time Permutation Scan Statistic for Disease Outbreak Detection

Solution

Clustering reduces the dataset and simultaneously creates a clearer picture of sentiment (geographical)

i. Extract sentiment at the global level via topic modeling – LDA, SVD

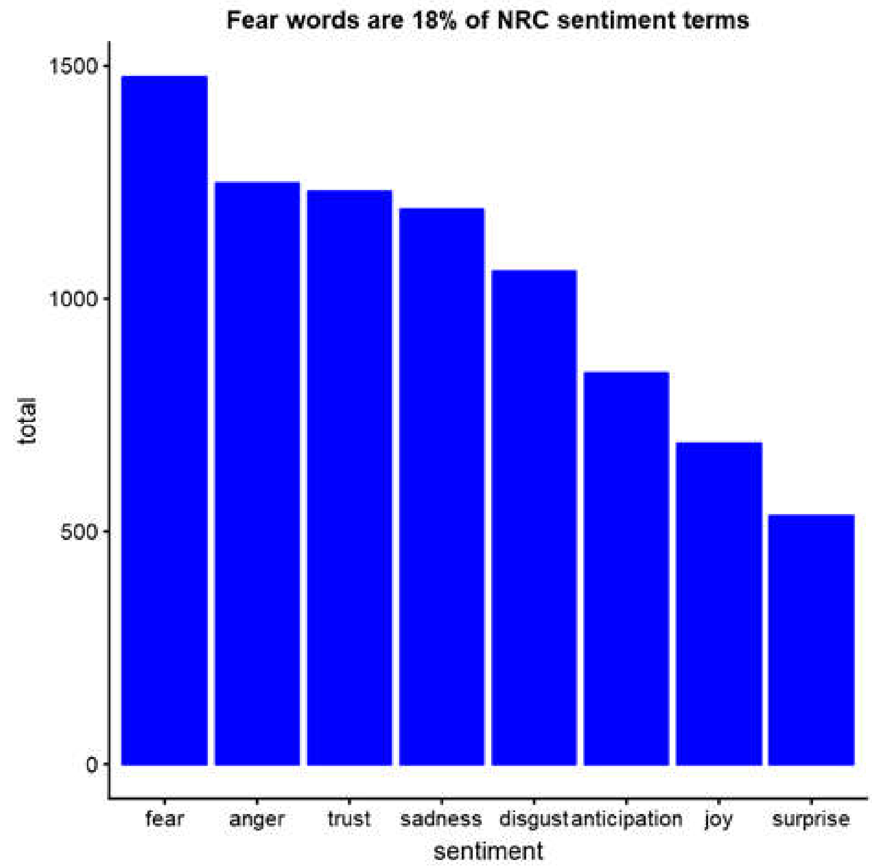
ii. Create clusters using STSS

iii. Extract sentiment at the spatiotemporal level using via topic modelling - SVD and Nonnegative Matrix Factorization (NMF)

Label tweets custom directory

•NRC emotion Lexicon

- Tweets classified into positive and negative sentiments
- Eight sub categories
- Same term can be classified into multiple categories (context-based). Count the number of categories

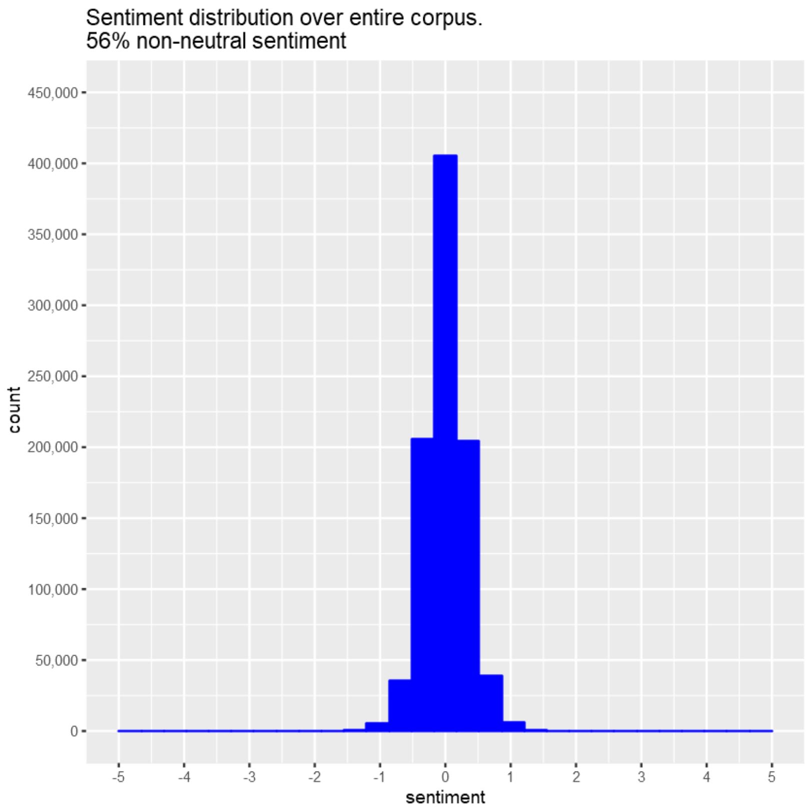


Customize NRC sentiments

- Rescale dictionary terms into
  positive (1), negative (-1), neutral (0)

- Create average sentiment using

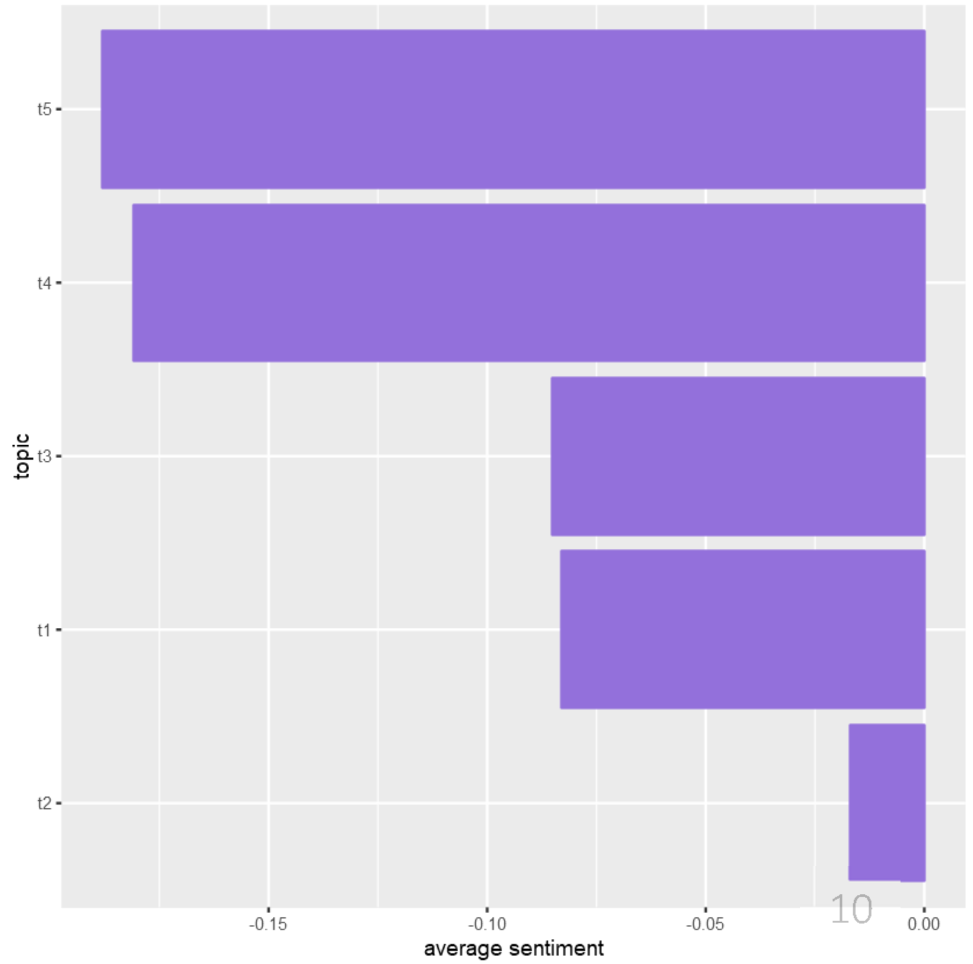
$$\frac{total\ sentiment}{count}$$



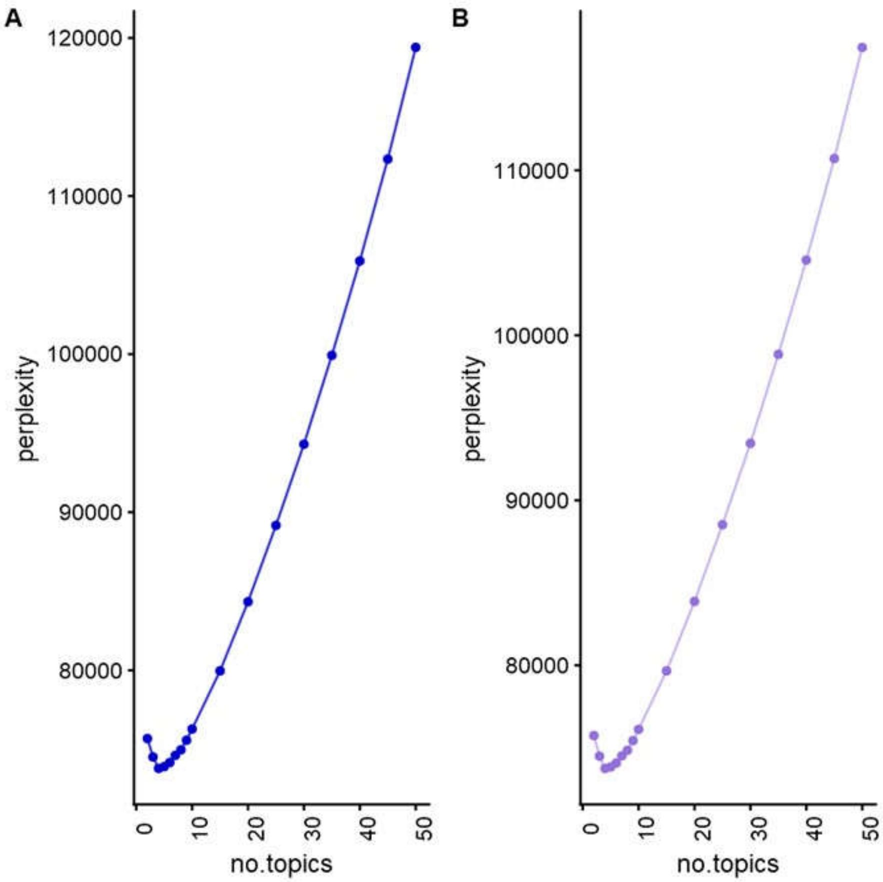
Global topic models

Latent Dirichlet Allocation (LDA)

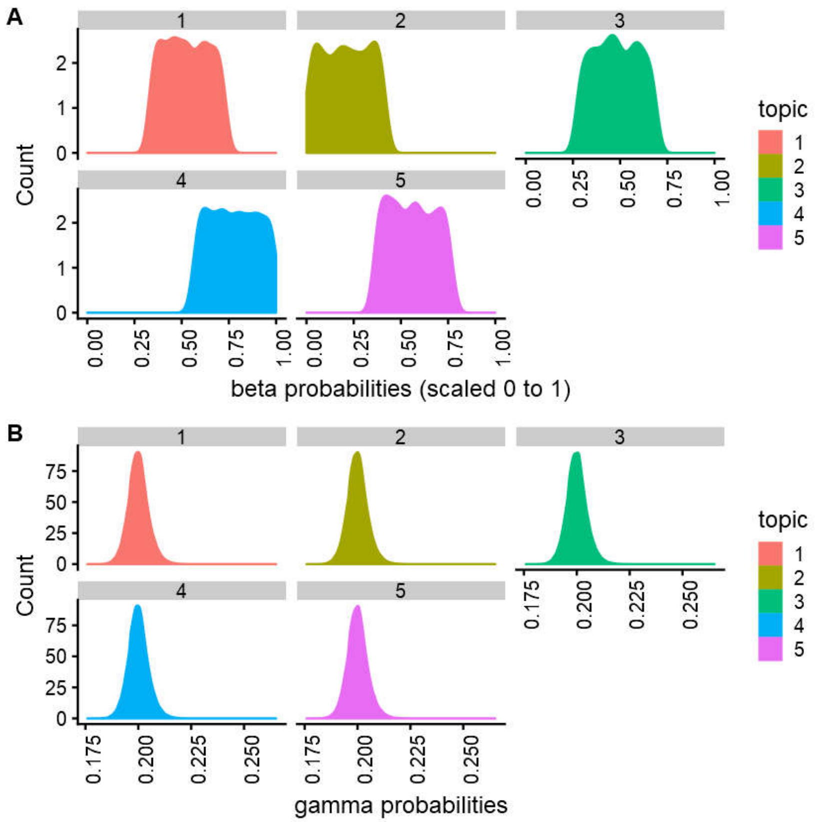
- Best k = 5 (at minimum perplexity)
- Bigrams and unigrams
- Positive only , negative only and all sentiment models
- Results mirrored regardless of ngrams and sentiment
- Beta, gamma distributions and k
- Top 20 terms
- All topics have an average negative sentiment
- **Advantage – finding best k for short text**



LDA Results



Best k = 5 at minimum perplexity. (A) train set. (B) test set



(A) Term-topic distributions top 1000 terms. (B) Topic-document distributions

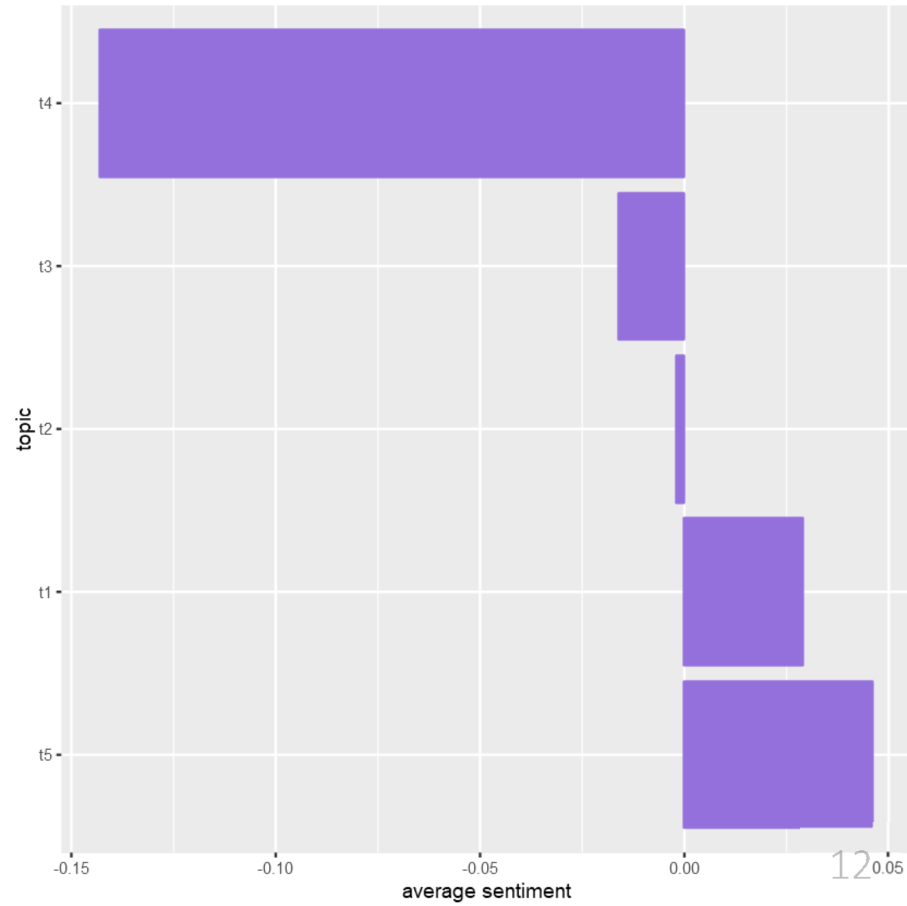
Singular Value Decomposition (SVD)

- Best k = 5 (from LDA)

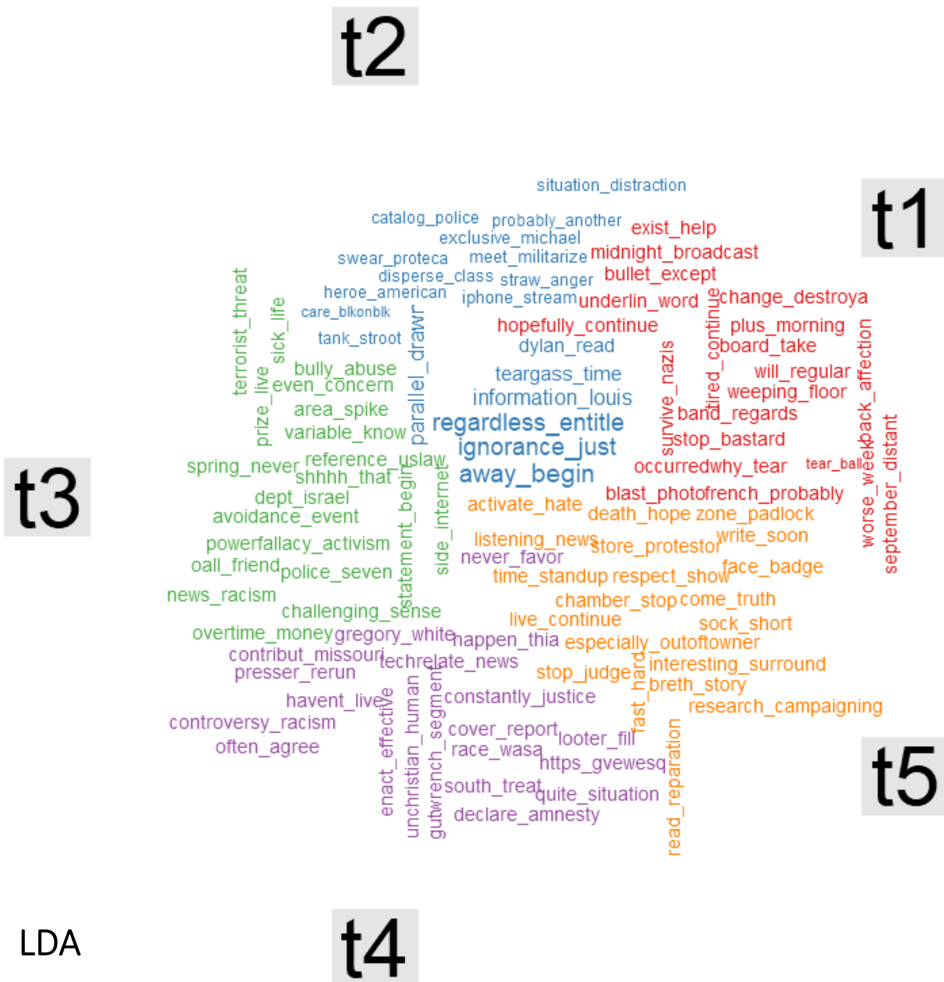
- Bigrams and unigrams

- Top 20 terms:-
- Topic 1, 5 have an average positive sentiment
- Topic 2, 3, 4 have an average negative sentiment

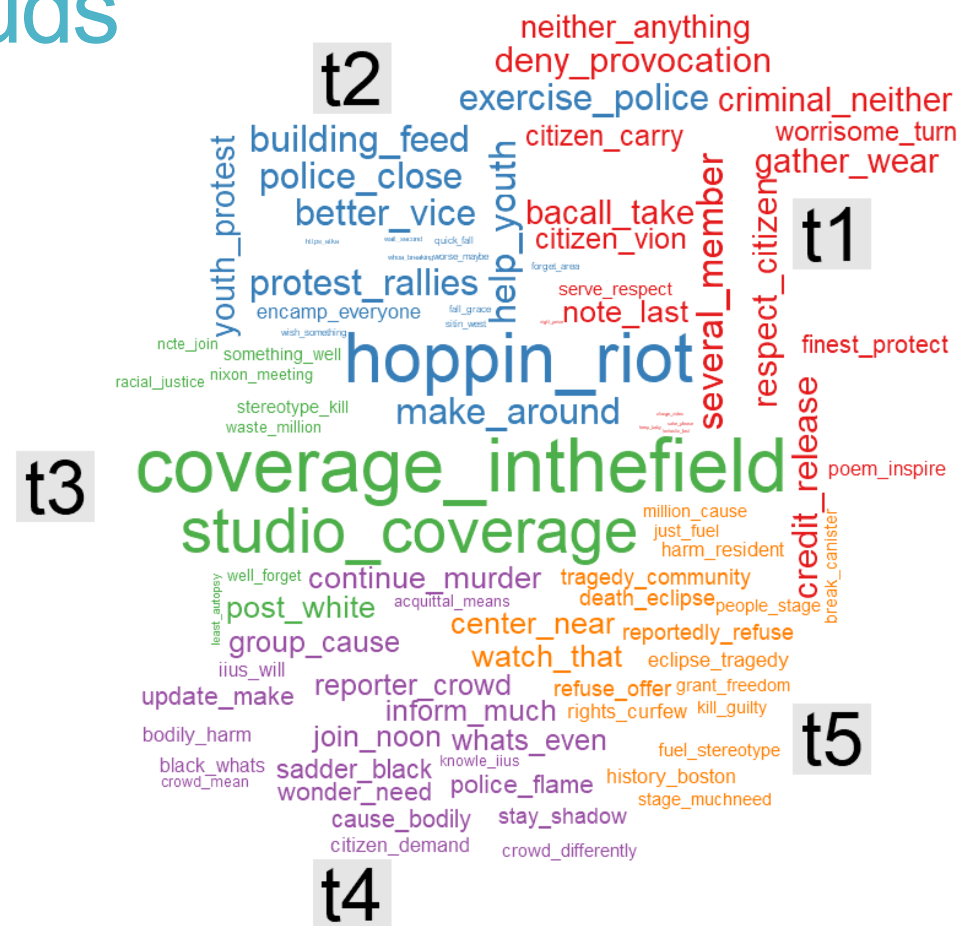
- **Problems - Sentiment difficult to interpret visually and bigrams redundant**



LDA and SVD wordclouds



LDA



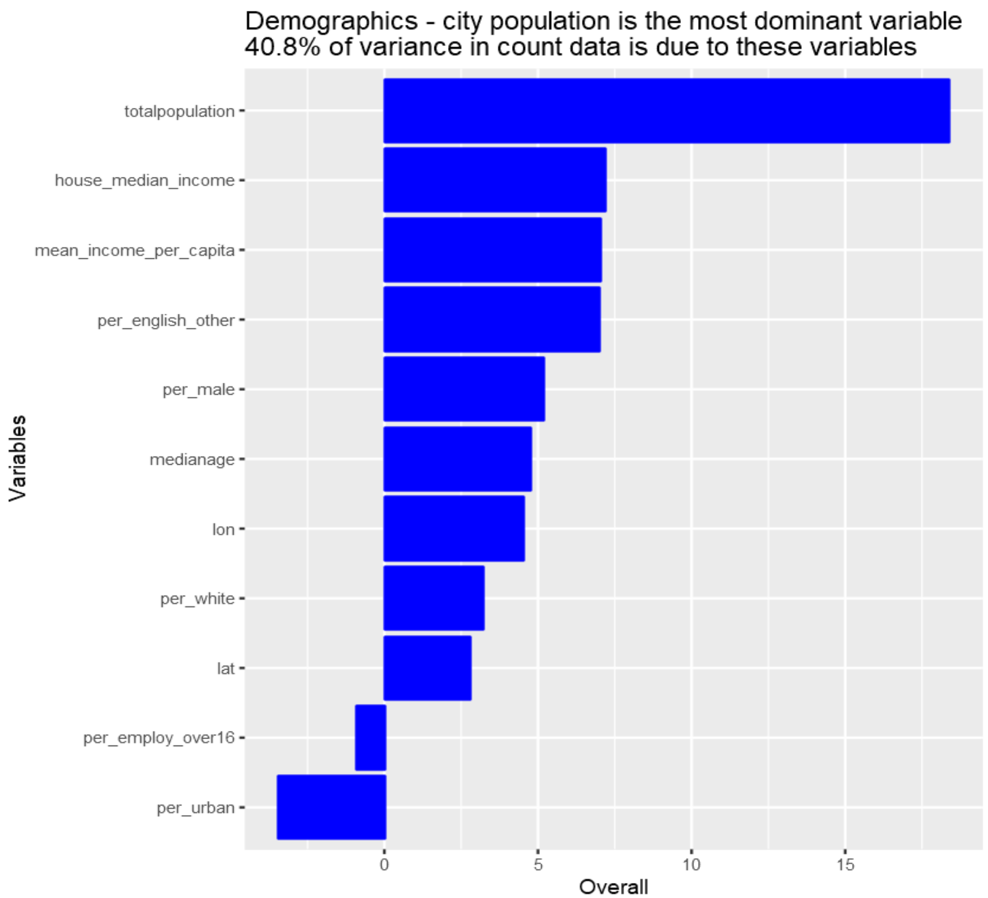
SVD

Spatiotemporal clustering

Using Spatiotemporal Scanstatistics (STSS)

Count data and other variables

- Based on count data and underlying Census information
- Requires count data proportional local population (univariate) or
- Additional data for multivariate
- Other possible variables (see figure)
- *Find variable importance using a random forest*
- Correlation positive counts and population = 0.40
- Select totalpopulation as only variable



STSS Clusters

Baseline model

- Count data zero-inflated and overdispered
- Baseline model using negative binomial GLM with positive counts >= 130
- Baselines useful for

Examining zero-inflated data

- *Proving overdispersion exists in the data*
- *Which scanstatistics method(s) will be useful in clustering*
- *Providing a guide for the STSS segmentation*

- Use STSS zero inflated Poisson

Model	AIC Results	Theta (dispersion parameter)	Mean Average Error (MAE)
Negative binomial (negative counts >= 130)	20,339	0.812	147.6[1]

STSS Clustering - results

Data distribution:	zero-inflated Poisson
Type of scan statistic:	expectation-based
Setting:	univariate
Number of locations considered:	194
Maximum duration considered:	12
Number of spatial zones:	715
Number of Monte Carlo replicates:	100
Monte Carlo P-value:	0.01
Gumbel P-value:	0
Most likely event duration:	11
ID of locations in MLC:	173

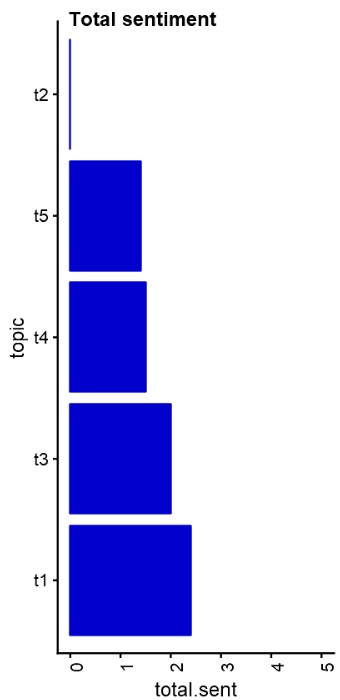
STSS Topic modelling

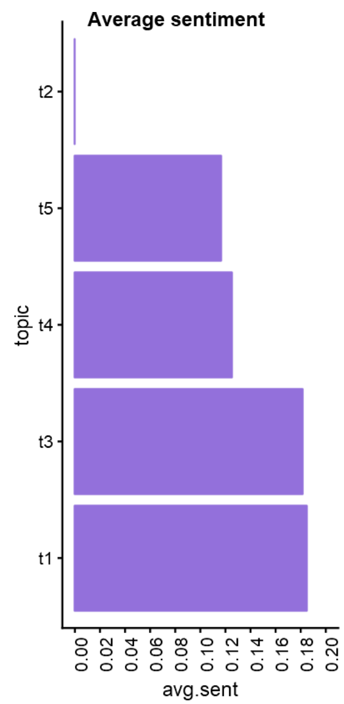
SVD top 20 terms - all cities

- All 5 topics are positive sentiment on average

- Visually only topic 1 might have a positive sentiment context

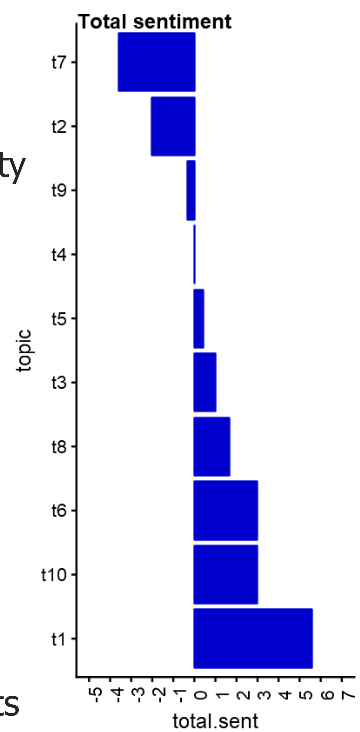
- **Disadvantage: Hard**
- **to interpret positive sentiment visually**

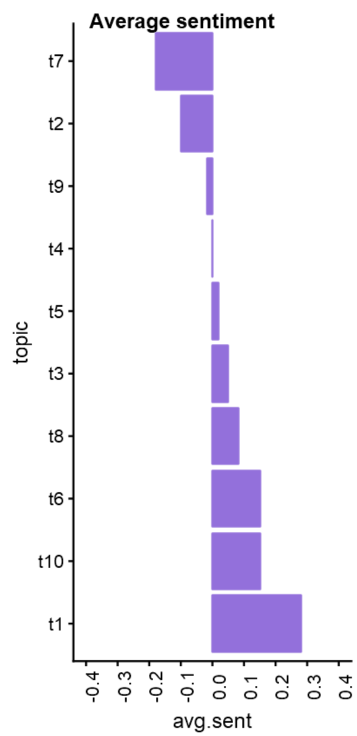


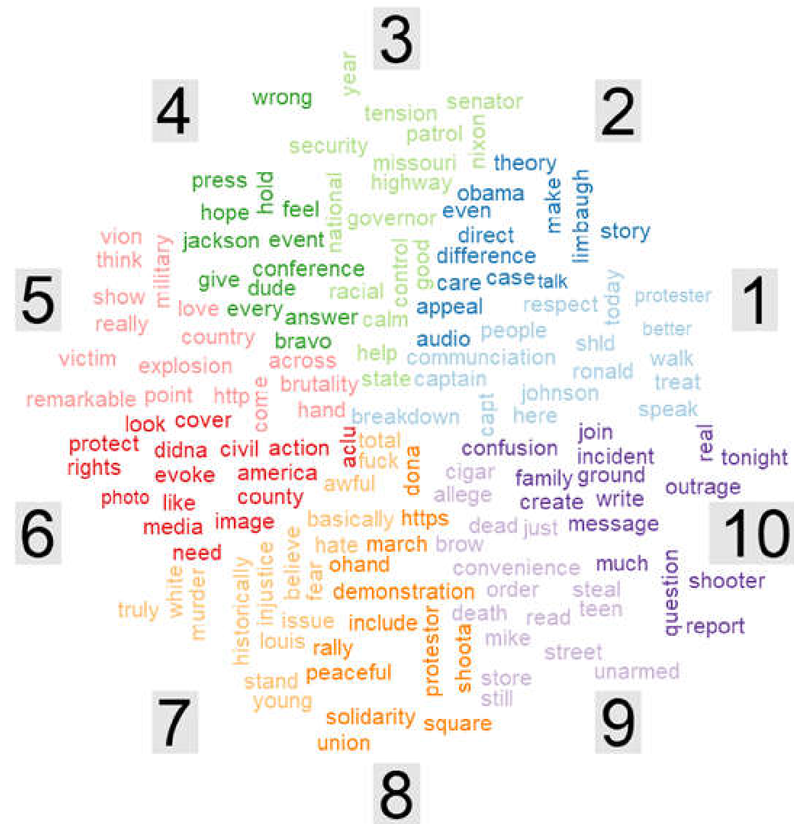


NMF wordclouds - NYC

- Possible themes
  - 4 - violence/military
  - 8 - peace/rally/solidarity
  - 6 - media,
  - 9 - initial incident
  - 1 - police
  - 2- political reaction
  - 3 - political/security
  - 10 - organizing protests /rallies/questions
  - 7 - injustice

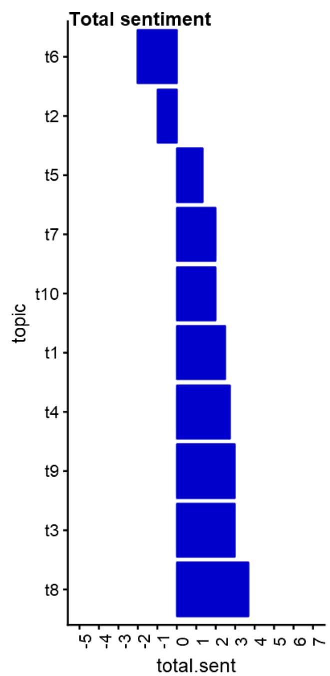


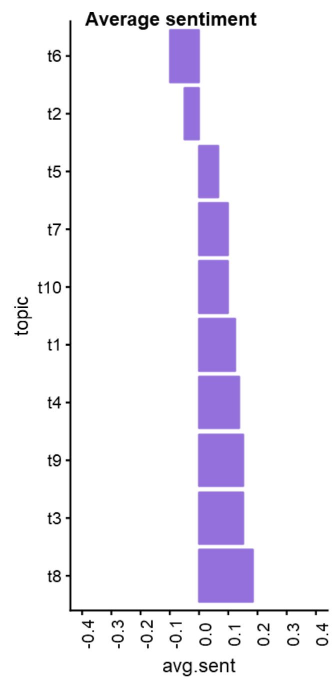




NMF top 20 terms – all cities

- Possible themes
  - 2 – injustice
  - 3 - rally/concern/solidar
  - 8 - legal/political
  - 9 - facts/truth
  - 6 - initial incident
  - 5 - organizing protests/r





Conclusion

- Ferguson is a negative polarizing event - based on injustice and/or race
- Pockets of positive sentiment exists in cities ranging from New York City to Scottsdale, Arizona, and Columbus, South Carolina to Beverley Hills, California
- Positive themes exists such as
  - Peace/rallies/solidarity
  - Legal investigation
  - Facts/truth
  - Political leadership
  - Injustice
- STSS (and topic modelling with NMF) does provide the opportunity to extract spatiotemporal anomalous positive sentiment in geographically diverse citiesa