

Sentiment Visualization

Data Analysis and Visualization using
Voxgov US Federal Government Media Releases



Presented by:
Mengying Li, Darrick Leow, Rongyao Huang, Yuqi Bi



Data

- 2012 US federal government daily media releases
(Volume: 200,539; Source: Voxgov)
- A subset of the release that has complete tags of political party, location & gender attributes
- 67,678 individual text files (JSON Format)



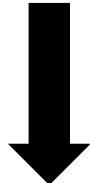
Research Outline

- ✧ Model latent topics based on the corpus
 - Assign each document to a topic
 - Calculate the frequency of each topic
- ✧ Reveal sentiment (negative/positive) for each document
- ✧ Explore the patterns of sentiment
 - Political party affiliation
 - Gender of respondents
 - Location source
- ✧ Deliver our results with efficient visuals



Process of Data Treatment and Analysis

Step 1



Data Cleaning

Data wrangling (clean, reformat) using R commands
e.g. stringR, Rjson, data.table

Step 2



STM

(Structural Topic Modeling)

- Creation of Corpus - removing of stop words, uppercase etc.
- Topic modeling: generate 10 topic categories
- Classification of individual text files into topics

Sentiment

Analysis

- Constructing a lexicon (combined from existing lexicons – Hu & Liu, McDonald, MPQA)
- Classifying each word as + or -

Step 3

Deliverables

Data analysis and conceptual prototyping
Visualizations generating using R, JavaScript, html, Adobe Photoshop



Process of Data Treatment and Analysis

Step 1



Data Cleaning

Data wrangling (clean, reformat) using R commands
e.g. stringR, Rjson, data.table

Step 2



Cutting Edge Topic Modeling Method

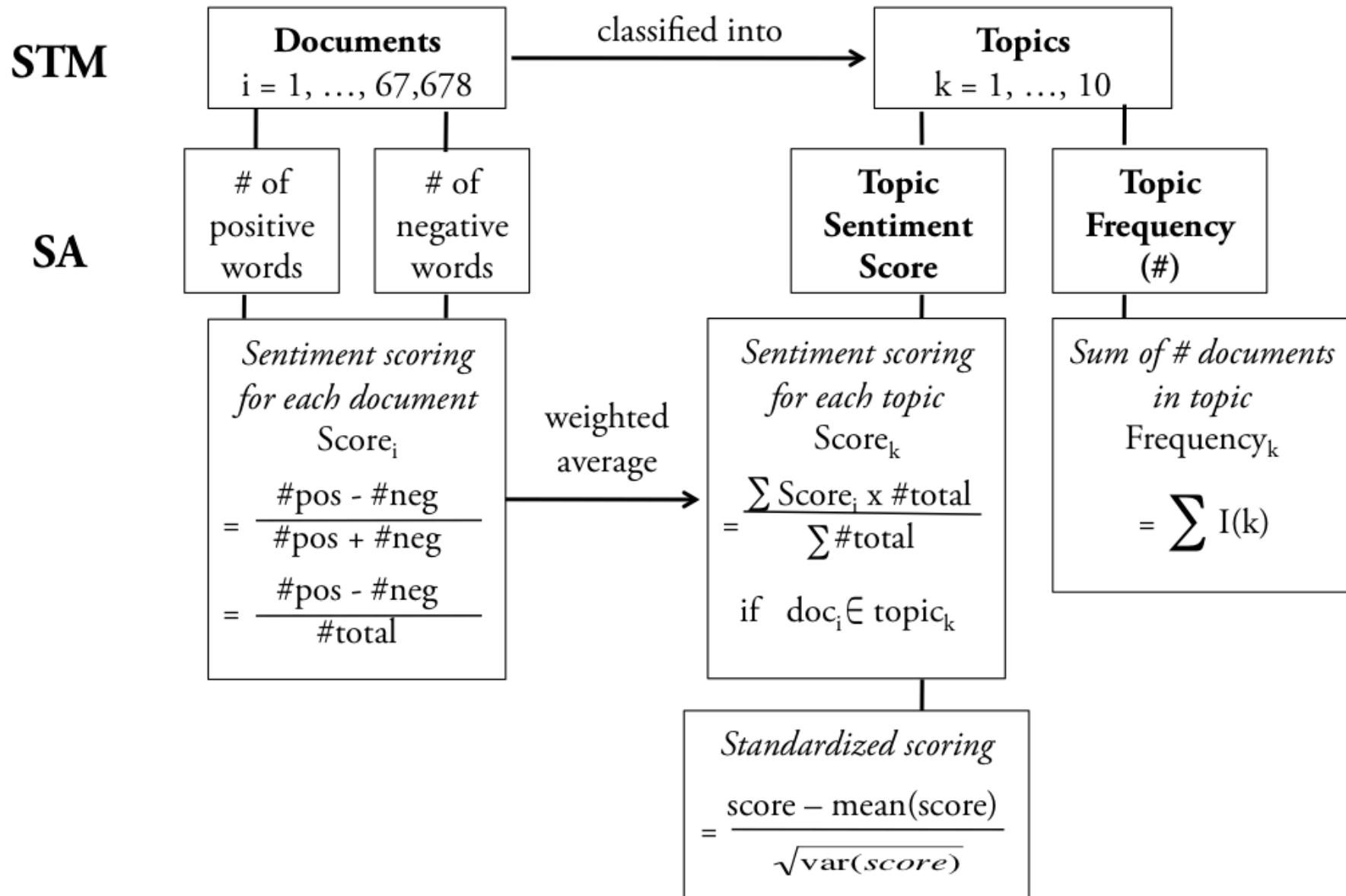
- Built upon the tradition of generative topic models
- Innovative in incorporating meta data (party, gender)

Self-Constructed

Step 3

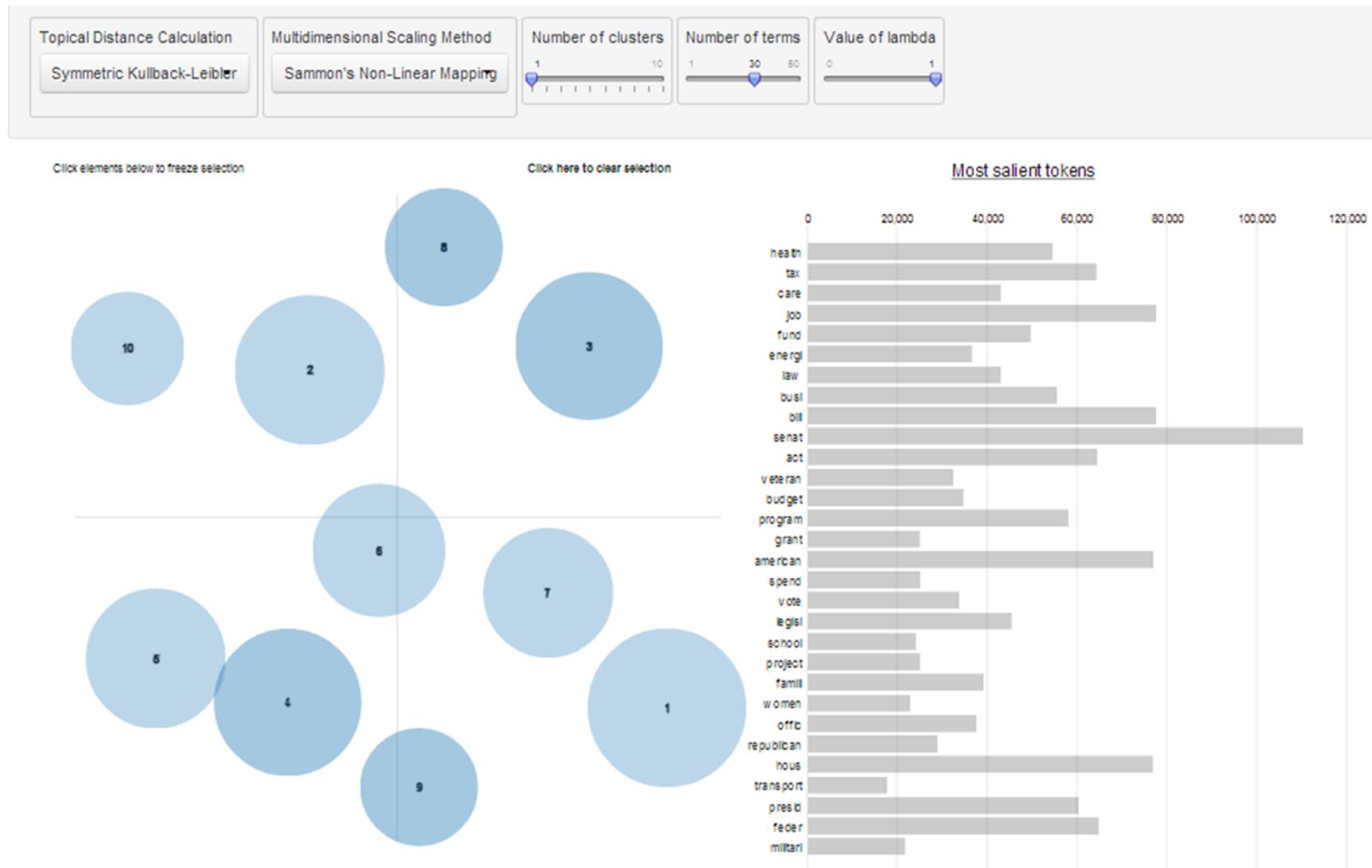
Deliverables

Data analysis and conceptual prototyping
Visualizations generating using R, JavaScript, html, Adobe Photoshop





Visualization #1a: Sentiment Topic Modeling





Topic Categories

1. Military
2. Jobs & Businesses
3. Education
4. Budget
5. Energy
6. Government Accountability
7. Disadvantaged Populations
8. Emergency Relief
9. Law & Judiciary
10. Health



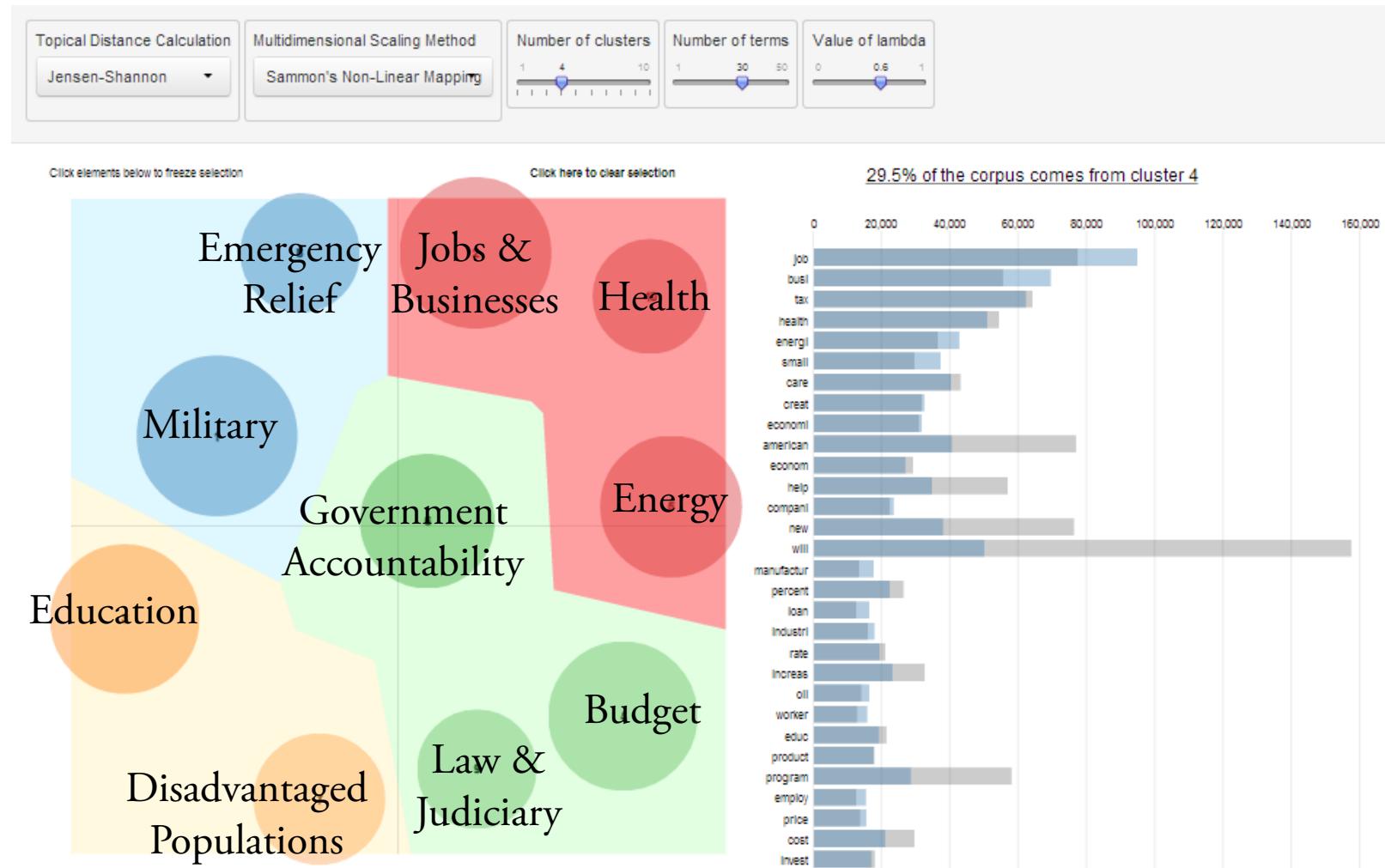
Tracing Back to ‘Most Relevant Text’

Health – an example

- Affordable care act benefits seniors on medicare advantage;
- Cardin, Cummings join Lieutenant Governor Brown to announce launch of first MD oral health campaign “Healthy Teeth, Healthy Kids” aims to help children access dental care.



Visualization #1b: Sentiment Topic Modeling (Clustered)

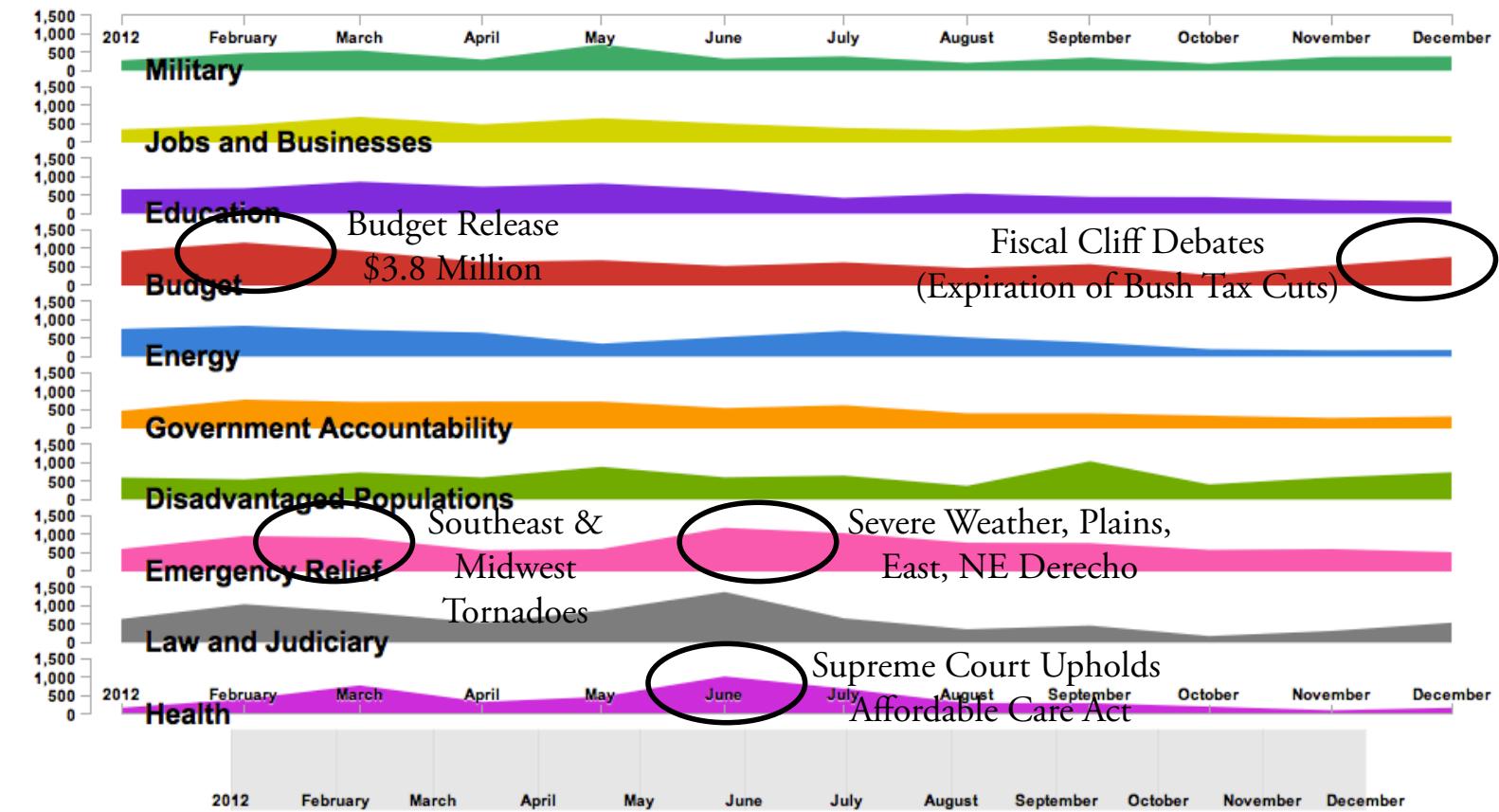




Visualization #2: Topic Distribution by Frequency over Time

2012 VOXGOV FEDERAL GOVERNMENT MEDIA RELEASE

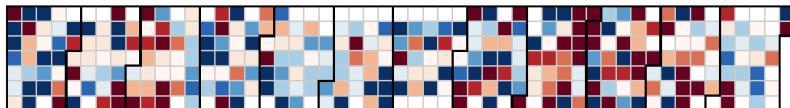
Topic Distribution of Media Release By Frequency



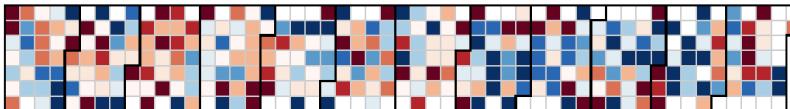


Visualization #3a: Sentiment Measure by Topics over Time

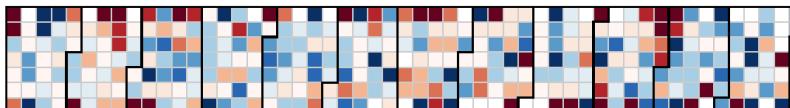
Military



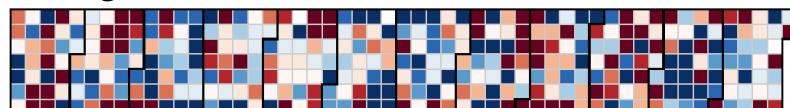
Jobs & Businesses



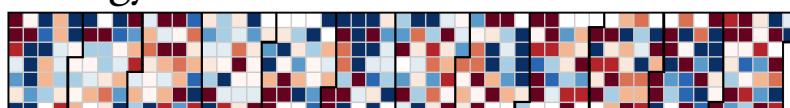
Education



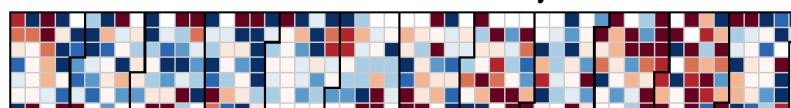
Budget



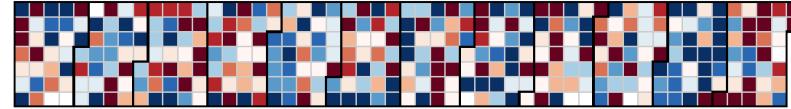
Energy



Government Accountability



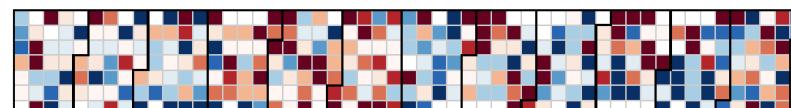
Disadvantaged Populations



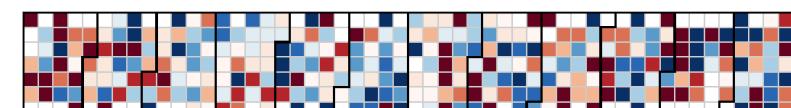
Emergency Relief



Law & Judiciary



Health

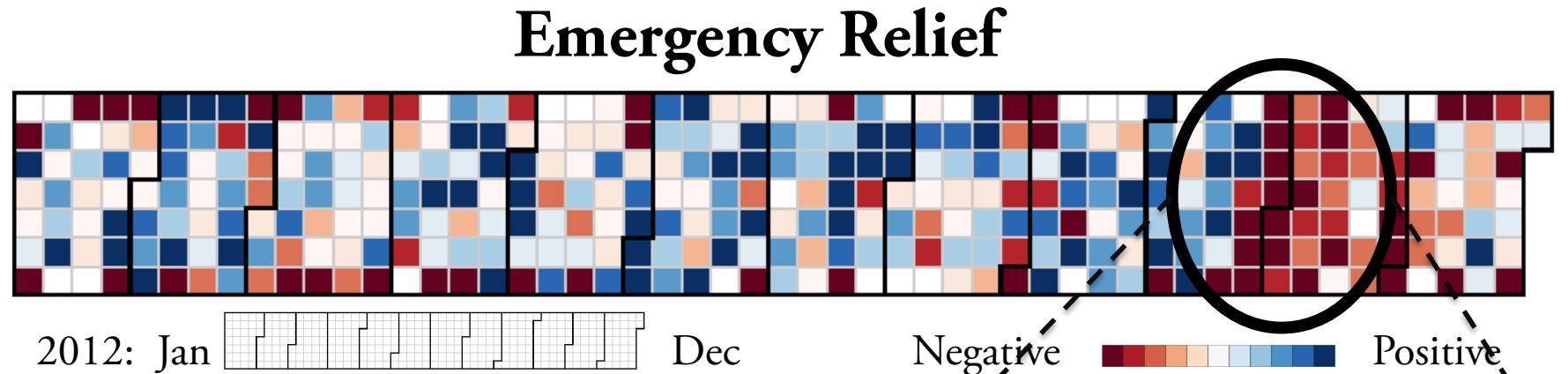


2012: Jan Dec

Negative Positive



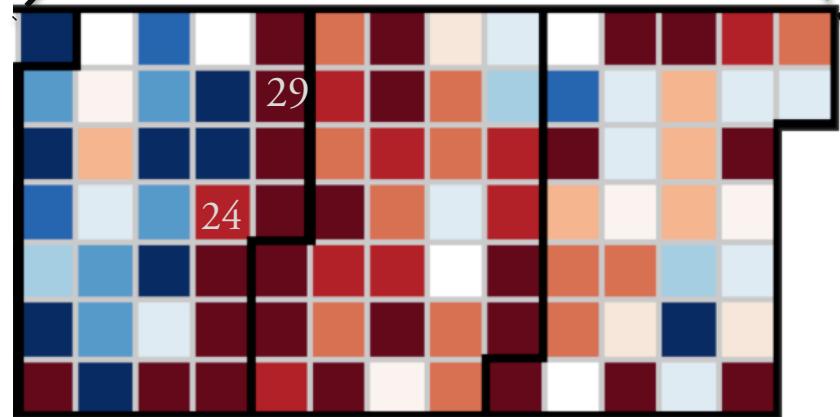
Visualization #3b: Sentiment Measure by Topics over Time



Hurricane Sandy

22nd Oct '12: Tropical Wave
 24th Oct '12: Classified as a Hurricane
 29th Oct '12: New York/New Jersey

24 States in the US affected
 >\$68 billion USD





Visualization #4: Sentiment Benchmarking by Political Party





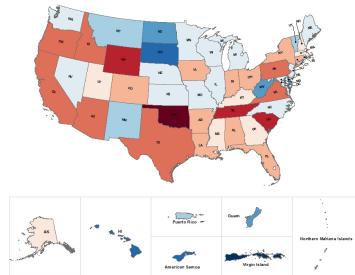
Visualization #5: Sentiment Benchmarking by Gender of Respondent



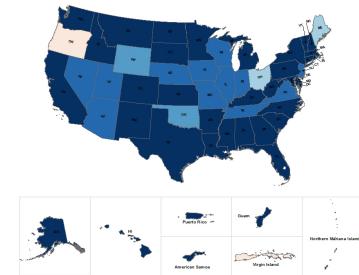


Visualization #6a: Sentiment by Location

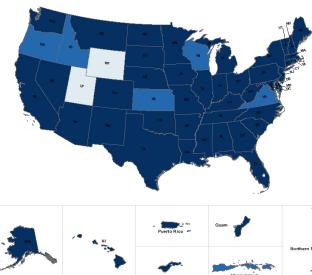
Military



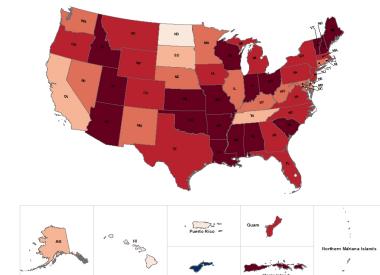
Jobs & Businesses



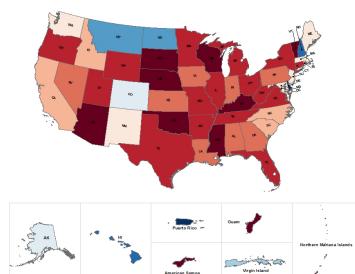
Education



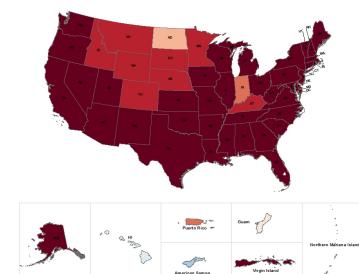
Budget



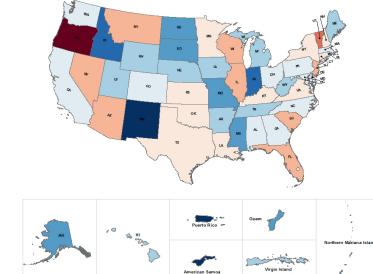
Energy



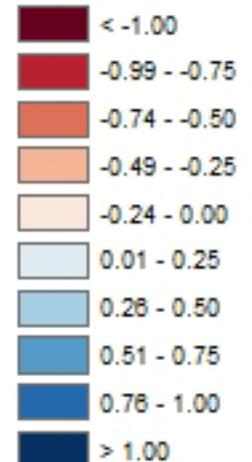
Govt. Accountability



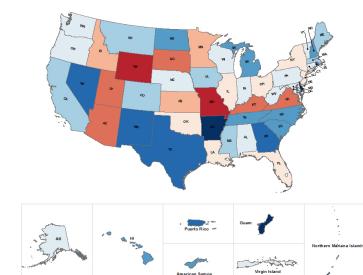
Disadvantaged Populations



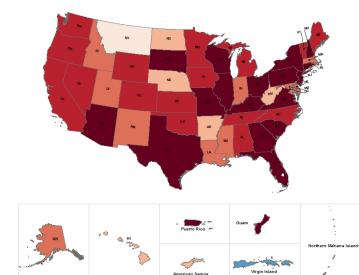
Score



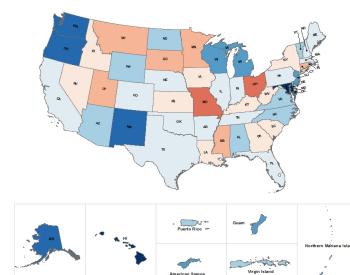
Emergency Relief



Law & Judiciary



Health

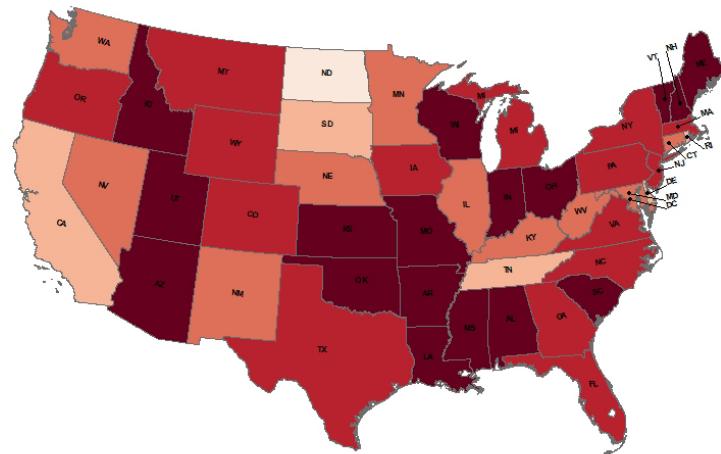
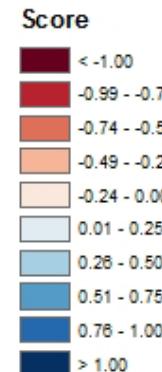
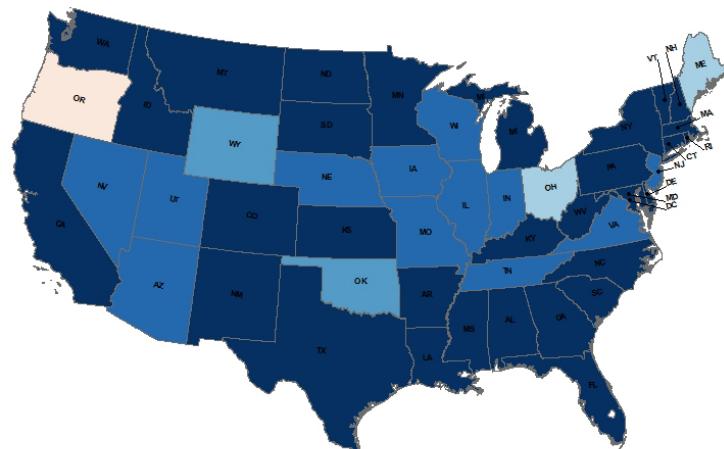




Visualization #6b: Sentiment by Location

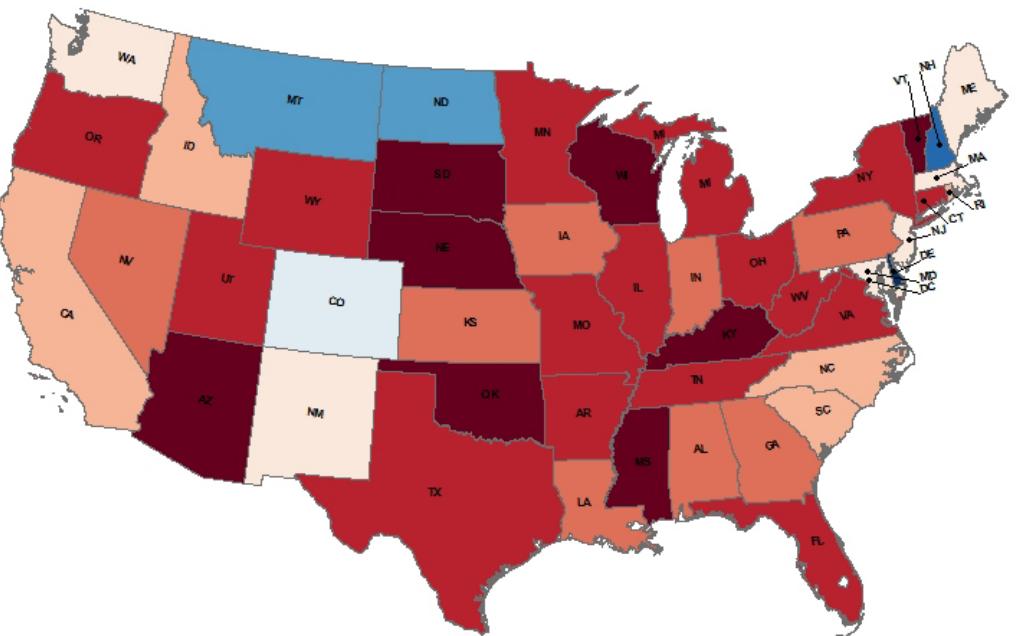
Jobs &
Businesses

Budget





Visualization #6c: Sentiment by Location Energy



Score





Limitations & Next Steps

- Non-perfect topic modeling categorization
(Objectively use ‘best guess’ to identify topic categories)
- Stanford Sentiment Analysis ([live demo](#))
- Other possible visualization (cubism)



Acknowledgements

- Data Source



- Software Packages (in R)

STM (Structural Topic Model), LDAvis, Shiny, ggplot2

- Lexicons

MPQA Subjectivity Lexicon, Hu & Liu, McDonald

- Visualization Sources

Multiple Area Charts, Calendar View, ArcMap

- Advisors

Prof Sharon Hsiao, Prof Gregory Eirich,

LDAvis developers Carson & Kenny,

Parallel Coordinates D3 developer Kai. S. Chang

Thank You

Q&A

