# Event detection with social media data

...

Fred Boehm (fred.boehm@wisc.edu), Robert Turner and Bret Hanlon
Department of Statistics
University of Wisconsin-Madison
Madison, Wisconsin, USA

# Overview

- Motivation
- Strategy
- Twitter
- streamR and parseTweetFiles R packages
- Latent Dirichlet Allocation (LDA)
- Super Bowl 2015 study & findings
- Next steps
- Resources

# Motivation

- Topic models, such as latent dirichlet allocation (LDA), enable analysts to identify themes, or “topics”, in a collection of texts
- How can we apply LDA to time series data (such as streaming tweets from Twitter)?
- Can we identify events that correspond to changes in social media discourse?
- Do such events lead to short-lived or persistent changes in discussions on social media?

# Strategy

1. Choose a short-lived social or political event that generated a lot of discussion on Twitter
2. Fit topic models of tweets before, during, and after the event
3. Compare topics for three time periods:
    a. Do any new topics appear during the event?
    b. Do new topics persist after the event?

# Twitter

- Users post messages of 140 characters or fewer in length
- Streaming API enables users to download a ~1% sample of all tweets during a chosen time period.

# streamR & parseTweetFiles packages

- Pablo Barbera (of NYU) developed the streamR package for interacting with Twitter's streaming API
- Robert Turner developed the parseTweetFiles package for processing tweets

# Latent Dirichlet Allocation

- A generative Bayesian model for identifying themes, or “topics”, in a collection of texts
- Developed by Blei, Ng, and Jordan (2003)
- Nearly 15,000 citations (Google Scholar)

# Latent Dirichlet Allocation

- Models a collection of documents
- Each document is a mixture of topics
- Topics are, technically, distributions over the vocabulary

# Super Bowl 2015 study

- We collected tweets from Twitter's streaming API before, during, and after the 2015 Super Bowl
- We fitted LDA models of tweets at three time periods:
  - 12h Pre-Super Bowl (5:30am to 6:30am on February 1)
  - During Super Bowl (5:30pm to 6:30pm on February 1)
  - 12h Post-Super Bowl (5:30am to 6:30am on February 2)
- We then examined topics with LDAvis R package and via topic-specific word clouds

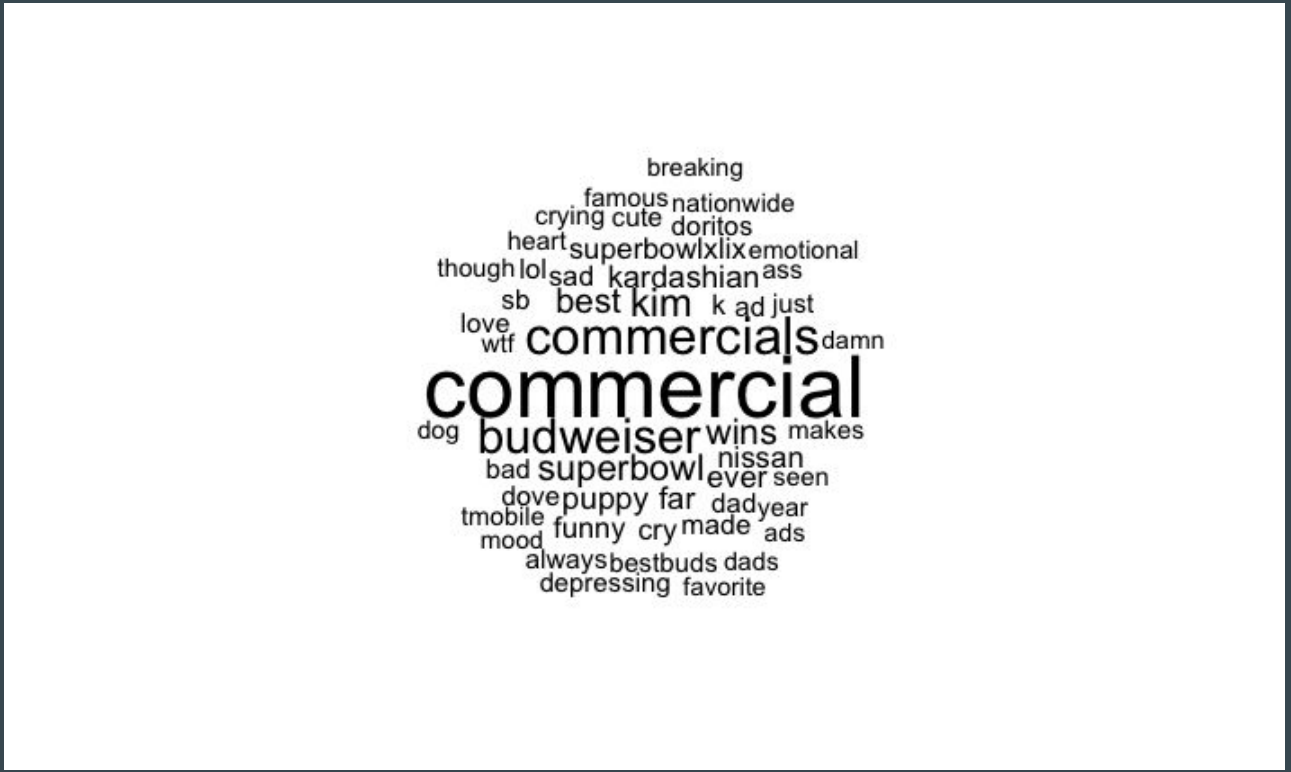
# Super Bowl 2015 Facts

- Kick-off at 5:30pm CST on February 1, 2015
- New England Patriots v. Seattle Seahawks
  - Final Score: New England 28, Seattle 24
- Star players included Russell Wilson, Tom Brady, Richard Sherman
- Halftime show featured Katy Perry, Lenny Kravitz, and Missy Elliott

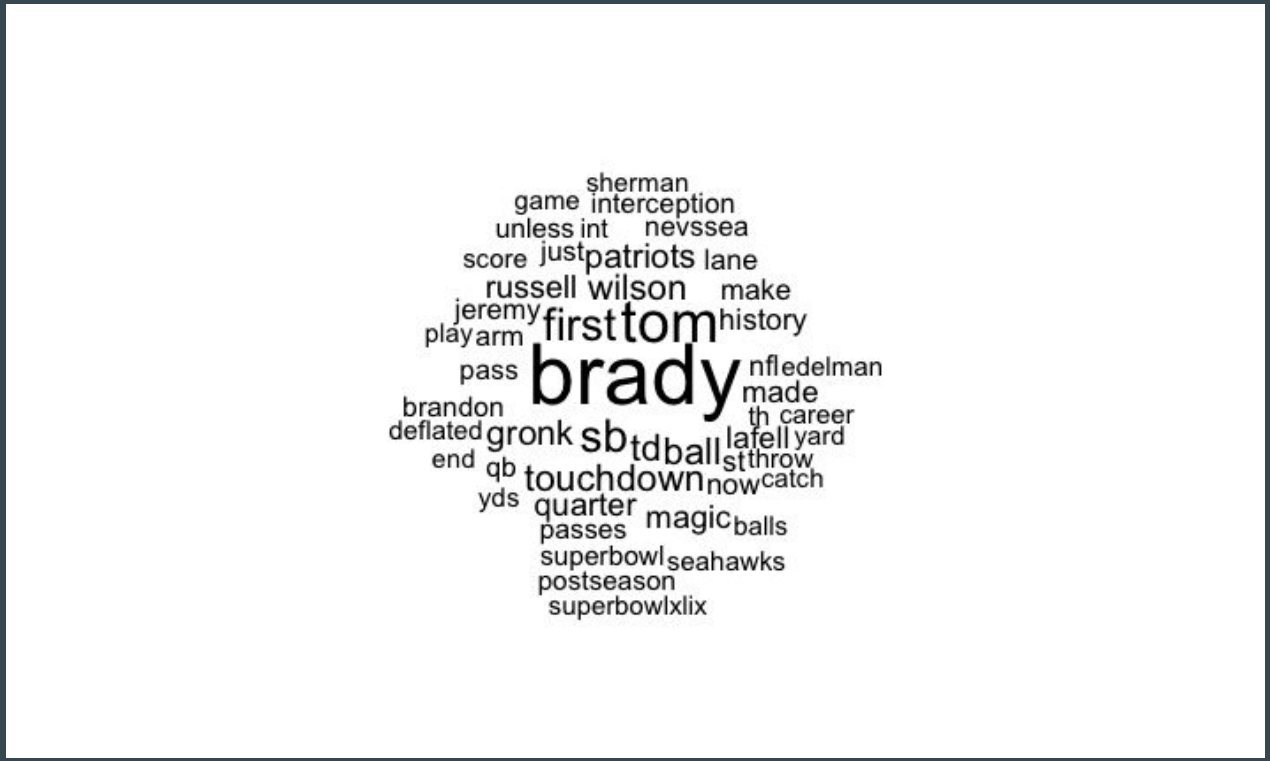
# Results from Super Bowl 2015 study

- Fitted LDA models of three distinct collections of tweets
- 20 topics per collection
- 12h pre-Super Bowl: No topics related to Super Bowl
- During Super Bowl: Topics related to star players and half-time performers
- 12h post-Super Bowl: Minor mentions of Super Bowl, but no Super Bowl-only topics

# Super Bowl Topic



# Super Bowl Topic



# Next Steps

- Develop topic models for streaming data
- Develop evaluation methods for topic models
  - Posterior predictive checks
  - Topic coherence metrics
- Use tweets over time to identify events that persist in social media discussions

# Thank you!

*Contact information:*

Fred Boehm: fred.boehm@wisc.edu

*Resources:*

parseTweetFiles R package: https://github.com/rturn/parseTweetFiles