Data Analysis with Topic Models for Communications Researchers

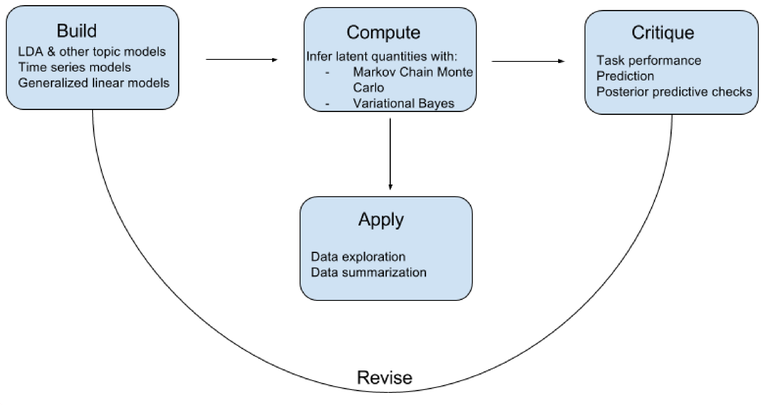
Frederick Boehm (fred.boehm@wisc.edu), Department of Statistics, University of Wisconsin-Madison

Abstract

We introduce topic modeling as a tool when analyzing textual data. We illustrate our methods with analyses of New York Times transcripts and tweets from three days in March 2016. We argue that such analyses will be useful in mass communications and journalism research. These methods are especially useful for identifying topics, or themes, in large collections of texts, when reading each piece individually is impractical.

Introduction

Social media users flood us with tweets, status updates, and blog posts. Data analysis with topic models enable researchers to identify themes, or topics, in a collection of texts. We present results from separate analyses of 1) New York Times articles from March 19, 20, and 21, 2016 and 2) a collection of tweets (from Twitter) from the same three days. Our manuscript contains computing code to reproduce our findings.



Methods

1. Fit latent Dirichlet allocation (LDA) models, separately, to 1) our collection of tweets and 2) our collection of print media articles from the New York Times
2. Visualized the topic modeling results with word clouds
3. Assessed resulting topics for coherence

Results

Figure 1: Word cloud for one topic (from a 20-topic model) of tweets from March 19, 20, and 21, 2016.

Figure 2: Word cloud for one topic (from a 20-topic model) of tweets from March 19, 20, and 21, 2016.

Figure 3: Word cloud for one topic (from a 20-topic model) of New York Times articles from March 19, 20, and 21, 2016.

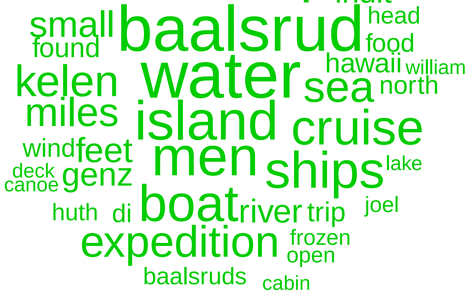


Figure 4: Word cloud for one topic (from a 20-topic model) of New York Times articles from March 19, 20, and 21, 2016.

Discussion

From the word clouds above, and those not shown, we find differences between the topics of discussion on Twitter and the topics in the New York Times. For instance, we find in the tweets an entire topic that deals with sports (with an emphasis on soccer) and another topic that involves music and entertainment. The New York Times analysis yields topics that match many of the newspaper sections. The two topics above might be called "Music" and "Travel & Adventure".

References

[1] David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.