Topic Modeling for Communications Researchers

Contents

1 Motivation

We are in the big data era. Social media inundates us with status updates and tweets, while bloggers share their views on current events. Website creation has become easier than ever. These new media interact with more established media, such as print news media, television, and radio. What do they have in common? They all can be viewed as data sources in which words - whether written or spoken, tweeted or blogged - are the data.

Big data presents big opportunities for communications researchers. We present a concise summary of a collection of machine learning techniques that, together, are called “topic modeling”. We discuss how these methods may be used in communications research, and we apply topic models to illustrative examples to demonstrate their value to communications research questions.

2 Overview

With the tremendous rise in computing speed and memory capacity over the last quarter century, researchers working at the interface of quantitative methods and social sciences gained the capacity to treat written texts as data. While document analysis is still in its infancy, scientists nevertheless have made great progress towards computational dissection and interpretation of texts. Among the most foundational contributions is the development of probabilistic topic models. We detail below, with limited use of statistical terminology, how these methods work and why they may be useful in communications research. We also provide computer code (in the R programming language) that implements these methods.

3 Latent Dirichlet Allocation

Blei, Ng, & Jordan (2003) introduced a generative statistical model called “latent dirichlet allocation” (LDA) in 2003. Although others had described similar statistical models (Pritchard, Stephens, & Donnelly, 2000), Blei et al. (2003) first applied the statistical model to text analysis.

3.1 What is the intuition behind the model?

As Blei (2012) writes, the key to understanding LDA is to recognize that a given document - be it a research article, a novel, or a blog post - exhibits multiple topics. Each topic, in turn, is, in a technical sense, a distribution over words. For example, a topic related to evolution may heavily weight the words “evolution”, “evolutionary”, “biology”, “phylogenetic”, and “species”.

An example may help to illustrate our point. Blei & Lafferty (2007) fitted a 100-topic model to 17,000 research articles from the journal Science. They found that a 1996 article “Seeking Life’s Bare (Genetic) Necessities” (Pennisi, 1996) exhibited topics related to “evolution”, “genetics”, “disease”, and “computers”. **Would it be more sensible to use RT’s Super Bowl results here?**

3.2 What is the model?

LDA models have a hiearchical structure in which words make up documents, and a collection of documents is a corpus. The corpus is assumed to have (unobserved) topics, or themes. The purpose of LDA, then, is to discover the unobserved topics from the texts.

3.3 What are its assumptions?

3.4 What are its limitations?

4 Computational implementation of LDA with R

We present below instructions and code for using LDA in the R statistical environment (R Core Team, 2015).

5 Visualizing topics

- LDAvis
- wordcloud

6 Results from Tweets Analysis

7 Discussion

8 Future directions

9 Online resources

References

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of*

Applied Statistics, 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Pennisi, E. (1996). Seeking life's bare (genetic) necessities. *Science*, *272*(5265), 1098–1099.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/