Topic Modeling Tutorial for Communications Researchers

Contents

1 Overview

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.

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.

2 History of Topic Modeling

3 Latent Dirichlet Allocation

Blei, Ng, & Jordan (2003) introduced a generative statistical model called “latent dirichlet allocation” 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 model?

3.2 What are its assumptions?

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

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/