Cassidy Maher

Preliminary Comprehensive Evaluation Proposal

November 3, 2018

**Electives**: Data Science (STAT231), Statistical Communication (STAT320)

**Off Campus Courses:** N/A

**Project Summary:** *In a few paragraphs, provide a brief synopsis of what you’ve already done, including a description of the data you have access to, and what questions/issues you have already tackled.*

***Brief Synopsis of Previous Work***

I decided to work with a topic that I encountered in my research at UoA with Chris Wild and in my Text Analysis Special Topics with Nick *instead of* extending my 495 project. I had a brief introduction to Latent Dirichlet Allocation (LDA) when I was spending time learning the basics of text analysis before beginning my interface for the basics of text analysis. I created corpora and document term matrices to make topic models with guidance provided by the Topic Models chapter in the Text Mining with R textbook.

I understand the basic setting of LDA in which every document within a corpus is a mixture of topics and every topic is a mixture of words. The only work I did with LDA was to tell R how many topics I wanted to be made and then it would cycle through my documents and determine topics and words that belonged to each topic (with the understanding that words can fall into multiple topics). I had no real understanding of the underlying processes and how this may extend my previous coursework.

***Data Access***

I shouldn’t have any issues with data access because I need text data. A good resource for finding clean text data is from the Gutenberg library, so I will probably turn to that for most of my data. There are also ways you can get newspaper articles and various other sources that are relatively easy to wrangle. For example, I already have a few hundred articles from BBC that are split into sections such as technology and entertainment.

**Proposed Tasks:** *Describe 2-4 additional tasks that will allow you to demonstrate your ability to communicate statistical ideas, perform analysis, and (ideally) use a new statistical technique to address a question of interest to you. Your proposal should clearly identify how your additional tasks build upon (and thus are different from) the capstone project.*

A lot of LDA is based on probability distributions and Bayesian processes in the context of unsupervised machine learning. I have seen these topics (some more than others) in my coursework and outside jobs but have not seen all three converge before. Since this relates to the text analysis work I have been doing over the past year, I am eager to understand more of the theoretical processes that go into making such a useful tool.

The primary task I will undertake is to learn the statistical theory behind LDA. Although I have considered the basic principles of LDA and have run a few examples of it in R, I have not looked at the Bayesian processes behind creating an unsupervised LDA model. Each document has a probability distribution over topics and each topic has a probability distribution of words. I’d like to better understand why we use the distributions that we do, how the distributions are related, and how these aid in creating Bayesian generative models.

This work will allow me to explore Bayesian analyses more in depth than I have before. Although I didn’t take a class entirely on Bayesian methods, we did brush the surface in 495. The fact that LDA is Bayesian allows it to continue to get better while new data is added. More specifically, I will learn how generative models work (something I’m not sure I’ve seen in the curriculum here). The generative model aims to find the a distribution as close as possible to the true distribution of the training data. This is how the latent variables (which is every variable except the observed words in the documents) are created in LDA.

If I have time, I think that running simulations could be a useful supplement to learning the theory behind LDA. I would compare the results I get from my simulations to the results I get from simply running the LDA command in R.

My final deliverable will likely be a report focusing on the theory behind LDA. I will touch on probability distributions, Bayesian modeling, and unsupervised machine learning. I would also like a small section on the real world applications of LDA.