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

Off Campus Courses: Introduction to Statistical Inference, Data Technologies (both at University of Auckland)

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 conducted text analysis research with Chris Wild at University of Auckland (New Zealand) and I am doing a text analytics Special Topics course with Nick Horton. In addition to other work with Chris Wild, I created corpora and document term matrices to make topic models based on an elementary introduction I had gotten in the Topic Models chapter in the Text Mining with R textbook (https://www.tidytextmining.com). I extended my text analysis project in my Special Topics course to create an “iNZight light” interface in the form of a reactive Shiny app. This interface is built for introductory students who want to get an easy and understandable glimpse at text analytics.

Based on this work, I have had a rudimentary introduction to LDA. For my comprehensive project, I would like to do an expository review of LDA that would extend my text analysis work with Nick and Chris.

Data Access

I will be using the full set of Emily Dickinson poems for the section in which I will apply LDA to a dataset. I already have this data from working with it for Special Topics and 495, but it is also accessible via the Gutenberg Library (https://www.gutenberg.org/files/12242/12242-h/12242-h.htm). This data is clean and has already been parsed by Nick, so I shouldn’t run into any issues when trying to work with it.

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.

My first proposed task is to learn about LDA based on the original LDA paper. This paper can be accessed here: [http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf](http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf). I will analyze the paper in enough depth to be able to describe what LDA is and the methodology behind it in a logical and straightforward manner.

Secondly, I would like to show my understanding of the paper and other supplemental resources by explaining it to others. This part will take the form of a final report in which I will include an explanation of my understanding of LDA based on the primary sources. Ideally, this would make it clear that I have grasped LDA and am competent enough to teach it to others.

The third and final task is to apply what I’ve learned to the corpus of Emily Dickinson poems as an example analysis. I will do a thorough analysis of what topics were created and what words make up the topics. An example similar to what I will do is done here: [https://www.r-bloggers.com/a-rather-nosy-topic-model-analysis-of-the-enron-email-corpus/](https://www.r-bloggers.com/a-rather-nosy-topic-model-analysis-of-the-enron-email-corpus/). The final deliverable for the example analysis portion would most likely take the form of a Shiny application.