Initial Ideas for Second Capstone Project
Andy Harless
Springboard
2019-04-27

1. Use Yelp restaurant reviews to predict changes in food service labor demand. (The expectation is that references to service inadequacy—e.g., when it's hard to get hold of a waiter to take your order—would predict increases in demand. But also the model might produce some surprises.) I can use JOLTS data on job openings, which are available monthly broken down separately by industry and region, and assume the two breakdowns are independent in order to estimate regional changes in demand within the industry. (This estimate, perhaps leaded by one or two months, would then become the "target" variable.) Right now my thought is to calculate embeddings for the words (or bigrams/trigrams) in a review, cluster (across the whole training set) on those embeddings, and use relative cluster density (within each review) as a feature vector. (Alternatively, maybe it would be better to aggregate by restaurant or even by locality before measuring density.) Then, for supervised learning, I could take each review as a separate case and predict the month-by-region target. (Statistically this is ugly because there is so much dependency among cases in the same month-by-region cell, but for machine learning it might be OK.) Would be cool if I could then use this model to anticipate which specific restaurants will want to hire. Anyhow these are just thoughts, and there's a lot more thinking to do. Also before the final proposal I might change my mind completely about what I want to predict. (And assuming I do want to predict labor demand, maybe the Conference Board HWOL data are better than JOLTS for this project, something I need to look into.

2. Use textual analysis of Federal Reserve minutes to predict policy decisions (possibly with a lag?). This paper (warning: slow PDF load) is something of an inspiration for this idea. In their case, the textual analysis results (from a simple bag-of-words sentiment model) are not part of a supervised learning paradigm but just used as an indicator of the Fed's sentiment, which the authors then correlate with economic data to make inferences about the Fed's preferences (e.g. sentiment becomes most positive when the inflation rate is near 1.5 percent). My thought is to use a more sophisticated textual analysis method (but what exactly?) as part of a supervised learning procedure in which observed policy actions (or inactions) would be the target.

3. Use textual analysis of Twitter data to predict bond market action. Related to #2, but the Fed's decisions occur in a context that is mediated and anticipated by markets, so it would be interesting to see if the crowd knows something before even the markets.

And then I can come up with dozens of other half-baked ideas (for example, use an LSTM network to help write poetry) but the hard parts are making sure that the data exist, that there is a methodology that would work, and that it would meet the requirements of the assignment.