Matt Sisco
From data to solutions (EECSE6898)
2/19/2016
Weekly report 3

Prof. Netzer began his talk by giving us an overview of the state of data science in business. He suggested that although businesses have begun attempting to utilize the abundant data available to them on the web, they are still not very sophisticated nor effective in their use of these data. Next, he told us about a project of his own which used blog data to examine market structures of automobiles. Essentially he created a matrix of all of the co-occurrences of different car models occurring in the same blog post. He then weighted these co-occurrences by the likelihood of seeing both car models together according to the multiplication of their individual likelihoods of appearing anywhere. Then he used network analysis to cluster them.

I was impressed by how accurately his method was able to group automobile brands into market segments that almost perfectly reflected the market structures revealed by survey methods and automobile trade-in records. Natural language processing can be an extremely complicated endeavor depending on what you're trying to reveal in the input text. However, the success of Prof. Netzer's project points out to me that sometimes very simple methods of parsing textual data can also be very fruitful. Prof. Netzer did not lengthily discuss his work using this method to investigate the side effects of pharmaceutical drugs, but he touched upon it and mentioned that it also worked quite well.

More specifically, Prof. Netzer's method has inspired me to try a similar approach in modeling the news data that I have now received from the News Rover project of Prof. Chang. I was able to extract about 80,000 online news articles that discussed climate change, and 4,000 transcripts of televised programs discussing this topic. In a separate project I have been working on comparing the effects of experiences with different weather events on attention to climate change using Twitter data. To supplement this work I'm now going to analyze the news data based on weighted co-occurrences of the names of different weather events with the mention of climate change. I believe this will provide another perspective on which types of extreme weather events are linked to climate change in the general public's mind.

Prof. Netzer also showed how with the same method he was able to track the movement of public perceptions of Cadillac from the American car category to the luxury car category. This was impressive because it lined up well with a marketing campaign that Cadillac initiated in 2004 to accomplish just this. Prof. Netzer showed the trends for Buick as a comparison, and they did not show a similar pattern which means the Cadillac trends are specific to Cadillac. When I originally read the paper

describing this project, I was somewhat skeptical about this case study. I wondered how many different automobile marketing campaigns he looked at before he found one that aligned with his data. I also wondered, after he decided to use the Cadillac case study, how many comparison vehicles he looked at before he found and decided to use Buick which nicely shows no change over the same time period. After listening to Prof. Netzer describe the work, I no longer felt concerned about these two questions. He explained why Cadillac was an easy choice for a case study, and did not seem to have looked at several other options in his data first. He also provided good reasoning for why Buick was an ideal comparison group for Cadillac, and he explained that most other brands also showed no movement over this time period.

The second project he presented was using essay text to predict whether or not a loan applicant would eventually default on the loan if he/she received it. This was an interesting project to see the results of because it was fascinating to see what the most predictive components of the essays were. For example he found that speaking about God and/or speaking about one's family both predicted that the person would more not pay the money back. Prof. Netzer compared the predictive value of the text-based models to predictions based on the standard financial variables such as credit score and debt-to-income ratio. He reported that his text-based models did a significantly better job in predicting future defaults than the financial variables.

While I believe that the textual data definitely has predictive value to add, I also wonder how much his team tried to make the financial models that they were comparing against as predictive as possible. To build the text-based model he tried several different machine learning algorithms and iteratively refined the model to maximize its predictive strength. I wondered if a similar amount of energy put into refining the financial-indicators model which he compares against. This question wasn't addressed in his presentation, and thus I am not wholly convinced that purely text-based predictions can categorically outperform financial-indicator models. I will look forward to this paper being published to look more into this.