Capstone Proposal:

1) What is the problem you want to solve? As Yelp grows in popularity, an increasing number of users rely on its reviews to identify new restaurants. This project would identify potential methods to improve recommendations for new restaurants based on user similarity.

2) Who is your client and why do they care about this problem? What will they DO or DECIDE based on your analysis that they wouldn't have otherwise? The client is Yelp. A more accurate recommendation process could increase the number of users on Yelp and encourage them to spend more time on the platform. Results may inform similar strategies with other recommendation platforms.

3) What data are you going to use for this? How will you acquire the data? The Yelp academic dataset found at: https://www.yelp.com/dataset This dataset contains over 5200000 reviews of over 174000 businesses.

4) In brief, outline your approach to solving this problem

a) Data cleaning and preprocessing: subset reviews based on prevalence of reviews and type of business, punctuation removal and lemmatization, phrase modelling, stopword removal
b) Topic modelling, likely with latent dirichlet allocation (LDA)
c) Prediction modelling of reviews based on a reviewers previous reviews (baseline model)
d) Cluster of reviewers based on characteristics and topics identified in LDA
e) New prediction model incorporating

5) What are your deliverables?

I will deliver documented code describing my analyses and a slide deck including visualizations and major results. Included in the visualizations will be networks describing the clusters of users and reviews.