Walker Page
Capstone Project 2 Proposal

I. The Problem

- **What is the problem you want to solve?**
- I want to find an efficient and accurate way to automatically generate tags/labels of business reviews that can indicate the content of the reviews and make it easier to group sets of reviews together in useful ways.

II. Stakeholders and Significance

- **Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?**
- Being able to easily understand and summarize customer reviews can enable businesses to better understand their customers and the key factors that are shaping customer experience.

III. The Data

- **What data are you using? How will you acquire the data?**
- The dataset, which contains over 6.5 million yelp reviews, is available here: https://www.yelp.com/dataset.

IV. My Approach

- **Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**
- I will approach this project in two stages. The first stage will be approached as an unsupervised topic modeling problem. The second stage will be approached as a supervised classification problem. The goal is to identify a relatively small number of topics that are intelligible and useful for interpreting the content of the reviews, and then use these topics to classify unobserved reviews. As a baseline for generating the topics, I will use Latent Dirichlet Allocation (LDA) to generate the topic model, and then I will subsequently consider constructing other models to improve upon the baseline. Initially, I will divide the dataset into two random subsets. I will use the first subset to generate the topics. The second subset will be used to train and test classification models that use the topics as classifiers. The second subset will thus need to be subdivided further into a training and test set. One challenge will be to find some way to label the training data. This will likely need to be done manually.

V. Deliverables

- **What are your deliverables? Typically, this includes code, a paper, or a slide deck.**
- My deliverables, as required, will include all jupyter notebooks I develop, a final report, and a presentation slide deck.