Big Data Final Project Proposal

Our team for this project is Logan King, Lori Chiu, and Sarah Torrence.

We chose this topic because it will allow us to understand how customer reviews of products are structured and how Natural Language Processing can be used to make predictions about customer reviews. Additionally, our group has a vested interest in this project because one of the members will be working for Amazon following graduation.

We are using a dataset found on Kaggle that contains product reviews from Amazon. The dataset is 54.41 GB in size. The reviews are separated by departments such as grocery, music, and apparel that we will combine into one final dataset. The entire dataset contains reviews that span between 1995 and 2015. The features included in the dataset are country, product category, customer id, upvotes, star rating, and review. We will focus on the text of the reviews to predict star rating.

In this project we would like to understand which customer reviews are more helpful. Specifically a few questions we would like to ask include:

- What is the average length of helpful reviews? Are more helpful reviews longer or shorter? This would entail understanding the lengths of the most helpful reviews (helpful votes).
- Can we predict what star rating a customer gave a product based on their review? This would entail a prediction model using the review text as an input and outputting the customer's review rating (star 1-5)
- What type of content is more or less helpful to other customers reading the reviews? This would include using the review text as inputs and utilizing nlp to understand what topics or content is more helpful to other customers (helpful votes)

There are several options for tools and algorithms for us to use in order to solve our questions of interest. Given the structure of data available, multiprocessing and or Spark can be used to perform operations on several partitions of our overall dataset at once. For NLP and predictive modeling, we can use operations available via MLlib and AWS.