Project 4 Summary

For this project I set out wanting to create an automatic text summarizer. The cosine similarity pair problem inspired me to see if I could automate a way to generate descriptive titles for an Amazon reviews. There was plenty of data to learn and extrapolate from, however I quickly shifted gears into seeing if I could combine what I learned from classification with unsupervised learning. My project morphed into one where I would create a model that could to distinguish between reviews that are more helpful to potential buyers and ones that are less so.

Using Amazon food review data, I anticipated there existed many different categories within the reviews. At the top most level, there are distinct food types (vegetables, sweeteners, etc.) being reviewed. However, I was unable to pull those out using more traditional methods. By utilizing Latent Dirichlet Allocation I was able to discover a lot of underlying classes of groups that existed in my data. This was perhaps the most exciting part of the project for me as it was absolutely incredible to see these categories “float” to the surface.

Although I did not ultimately end up successfully distinguishing reviews based on their helpfulness, I have a few ideas as to how I can get closer to my original objective. One of the biggest areas for improvements is tuning the original classified I build. Due to the distribution of the helpfulness of the reviews, it is difficult to generate balanced classes based on the ratio I picked. I believe the best approach would be to actually go in and classify reviews by hand -- while time intensive and not particularly fun, it may actually be the fastest way to create a data set better models can be build upon. Following this, I would also like to employ more classic classification methods such as SVM to see how they perform compared to unsupervised learning.