Clayton Murray

Final Capstone proposal

Problem. The Yelp platform offers a trove of information to both business and consumer. We can mine lots of specific insight from the large dataset provided by Yelp, but can we deduce the basics? Specifically: *is there a method to detect the value of a consumer's review in the eyes of other consumers?* In this project, I plan to answer the question of whether the Yelp platform can confidently capture review value to offer sufficient, statistical advantages to businesses.

Value. This offers value in two distinct ways:

1. ***Benefit Yelp*** *by increasing the amount of accurate reviews*. If we can analyze what comprises a publicly valuable review, we can detect when authors create these reviews and incentivize them to create more. For example, when Jane Doe submits reviews flagged valuable by the algorithm, the platform can incentivize Jane Doe *directly* to contribute valuable reviews via coupons, badges, or in other ways, thus increasing the amount of strong reviews on the platform. Attempting to do this through the business under review will undoubtedly create positive bias.
2. ***Benefit companies*** *by recognizing factors that are more significant than others.* In the event a review or reviewer is flagged as valuable, businesses can focus on the content therein. This allows them to capitalize upon recognized successes or address negative concerns. For example, if a set of reviews are flagged as low value and gripe about issues x, y, and z, but Jane Doe and other reviewers who are recognized as valuable note that restaurant *Zoey's* service and food were a bit weak on Cinco de Mayo, the establishment can confidently address the peak event concern as the more important issue.

Source. The engineers at Yelp have collected a vast and interconnected dataset featuring images and data about reviews, consumers, and businesses. This set is fairly large, but it has a centralized download source via the challenge section of the website, located here: https://www.yelp.com/dataset/challenge.

Techniques. Although I have not yet begun, there are a few primary technologies and methods I hope to apply:

- Standard stack (Python3, pandas, matplotlib, seaborn, Jupyter, etc.)
- genism and word2vec, spacy library possibly using GPU implementation, and other NLP tools
- feature reduction techniques, likely LSA
- Multi-class statistics, such as ANOVA
- KNN for unsupervised; logreg, boosting, and trees for classification or determining probabilities

Possible technologies:

- Tensorflow
- NLP summary and extraction
- Dask, batch processing, or PySpark in the event data is too vast for in-memory calculations
- Others as needed during project

Challenges. Since we have spent lots of time working on technical things like coding, statistics, and deep diving various methods, I feel like the most challenging aspect of the project will be *interpretation*. There will be results produced by the analysis; are the results telling us enough to act? Are we overlooking

anything? Are we trying to capture too much? The interpretation and adjustments to do so will be the trickiest part.