Springboard Capstone Project Proposal #2: Natural Language Processing

- **What is the problem you want to solve?**

This capstone will be an exploration of Berkshire Hathaway's annual shareholder letters using Natural Language Processing to create a topic model. This model will try to extract meaning from these letters by identifying recurring themes or topics and their change from letter to letter. In addition, I want to track named entities such as corporations, people, and industries over time.

- **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?**

The "client" in this case is an investor that is interested in investing like Warren Buffett. Warren Buffett is one of the most successful investors of all time. If there is a link to topics in the letters and Berkshire Hathaway's share price, investors can use these topics to create their own investment ideas that to try to imitate Buffett's performance.

What data are you using? How will you acquire the data?

The data comes from Berkshire Hathaway's shareholder letters available at the link https://www.berkshirehathaway.com/letters/letters.html. The letters come in both HTML and PDF format, so part of the challenge will be scraping data from both formats into a feasible data set. I will update progress on this step via GitHub.

- **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.**

The first part of this project includes downloading and extracting the data. Once I have acquired all of the annual reports, I will create a natural language processing pipeline to clean the reports. Louie's presentation provides a good roadmap for creating this pipeline and ideas for EDA. Machine learning will start with linear regression, then progress to more complex methods as appropriate.

- **What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

Deliveries will include a series of Jupyter notebooks, a 10-12 page paper documenting the process and explaining the results, and a slide deck to explain results to technical and non-technical audiences.