Tom Halloin

Capstone Proposal Project #2: Analysis of Berkshire Hathaway Annual Letters Using Natural Language Processing (NLP)

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

Warren Buffett is one of the most successful investors of all time. This capstone will be an exploration of his annual shareholder letters using Natural Language Processing to determine the following:

1.) Prove or disprove the hypothesis that Warren Buffett has a clear, sustainable investment process that others can follow for similar results.

2.) Use Natural Language Processing to look for reasons Berkshire Hathaway's stock increased or decreased in value. Build a model that can explain growth in book value over time given certain language appearing in an annual report.

- **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 a typical investor that is interested in investing like Warren Buffett. Instead of relying on aphorisms such as "Be fearful when others are greedy and greedy when others are fearful", this project looks to provide a more systematic way to invest given the successes and failures of Berkshire over time.

- **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. See the Github for this page for progress on this step.

- **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. The statistical analysis and hypothesis testing will focus on the first question. 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.