**Project Proposal:**

* We will analyze Amazon review text for Kindle books to try to predict a user’s rating of a book, based on the text of their review. (Sentiment Analysis)
* Once we can accurately predict this (time permitting), we will try to develop an algorithm that picks out features of negative reviews and generates a response to address the user’s concern.
  + Book was boring => Here’s something more interesting
  + Book was dense => Here’s something easier

**Will’s Comments:**

* Length of the review impacts polarity
* Spacey library (pre-processing to create feature index)
* Split the most positive and most negative reviews
  + Topic modelling on subsets
  + Take all negative / positive words, find antonyms, get books associated with opposite

**Research Questions:**

* Can we predict a reviewer’s star rating from their text summary?
* Can we cluster negative responses based on sentiment (e.g. boring, confused, etc)
* Can we cluster reviewers based on their general sentiment (always positive, always negative, etc)?
* Can we cluster products based on their review description?

**Dataset Summary**

* Amazon Kindle Book Reviews since the launch of kindle
* Review text, review summary, star rating, helpful votes, etc.
* Product metadata

**Methodology**

* TextBlob to analyze text data
* Alternatively, PyCaret NLP libraries
* Alternatively, Spark Example
* K-means or other clustering methodology to cluster users, products, etc based on their description (e.g. “boring books”)

**Final Presentation Flow**

* Data set summary
* S3 Storage / Git LFS / Hosted ML
* Data cleaning for machine learning
* Trained model
* Test Model
* Response Engine
* Further Clustering Analysis
* (Optional) Test on Adjacent Products (e.g. Fashion)