PRESENTATION OUTLINE

- I wanted to do a project with NLP and text generation
- Wanted to incorporate gathering my own data and web scraping
- Since there are many tutorials, and almost infinite data, i chose amazon reviews
- Overview of products I chose and why

SCRAPING THE DATA

- Followed two tutorials, combined the two
- Found a chrome extension that orders amazon products by most reviewd
- Used Chrome Developer tools and look for the correct things to grab
- Tagged as a robot and had to change my user agent
- Talk about using the URLs to get good and bad reviews

PREPROCESSING THE DATA

- Need to make the data useful
    - Had some ideas going in about what might separate reviews
- Talk about python libraries I used to get the data
    - JSON reader, CSV reader and dictwriter
- Using NLTK
- Errors in text - weird characters, unicode errors, etc
- Bag of Words and Bigram - what it is
    - Talk about sending wrong data to CSV (first 200 instead of best 200)
    - WEKA difficulties - too many attributes!
- Smart data
    - What the different attributes are

CLASSIFYING THE DATA

- Table of results
- Talk about difference between full dataset and Selected dataset
- Bigram - what are the useful words?
- Smart - how are the different attributes distributed - adjective for instance

GENERATING NEW DATA

- LSTM - long short-term memory, more complex version of a recurrent neural network, one that has some persistent memory. Allows a cell to ‘remember’ or ‘forget’ some or all of the persisting information.
- experience training the model - long hours and computer trouble
- Talk about Markov Chain
  - Suppose that you start with $10, and you wager $1 on an unending, fair, coin toss indefinitely, or until you lose all of your money. If I know that you have $12 now, then it would be expected that with even odds, you will either have $11 or $13 after the next toss. This guess is not improved by the added knowledge that you started with $10, then went up to $11, down to $10, up to $11, and then to $12
- Knowing more about LSTMs from this, if i could use whole words instead of individual letters, or double letters or triple letters, it might work out better. Just the same with the Bigram data, using 3 or 4 words, or bringing it down to a few letters would certainly change the output, and might make the output more nonsensical or more legible. Training next step is training on jerry seinfeld episodes.