Group 6

Research question- What are the most common topics in e-book reviews on Amazon, and how do these topics differ among positive, neutral and negative reviews?

Hypothesis- ?

Two parts? Maybe just focus on part 1 if time gets difficult?

Part I

* Text cleaning, custom stopword removal (to preserve sentiment like ‘did not like’ etc.), parsing into unigrams/bigrams/trigrams
* Coherence score graph to determine best number of topics
* Topic LDA modeling on review text to understand and compare recurrent topics in positive reviews (4-5 stars), neutral reviews (3 stars), and negative reviews (1-2 stars) in e-book reviews
  + May have to be mindful of reviews of delivery/service vs. actual book content within topics

Part II

* Use the probability distribution from the topic modeling + (potentially) count vectorized/tf-idf/embeddings to create a predictive multi-class classification model to predict the review score of a (new) review
  + Or maybe the independent variables can be the topic probabilities and some of the features from the dataset for each review like helpful votes, vine, verified review, word count, etc.
* This will relate the topics found in Part I, and the tendency for certain topics to appear within positive/neutral/negative reviews
* Are there any Amazon product review embeddings?

Vader analysis logic:

<https://python.plainenglish.io/twitter-sentiment-analysis-using-vader-tweepy-b2a62fba151e>

Create stop word list:

<https://ithaka.github.io/tdm-notebooks/creating-stopwords-list.html>

Explanatory data

<https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a>