NATURAL LANGUAGE PROCESSING

- Before making predictions, book reviews must be converted into a corpus.
- Normalize corpus with lowercase letters; eliminate stop words and special characters.
- Use CountVectorizer and TfidfVectorizer to convert individual reviews into a sparse matrix.
- Iterate over n-grams. One word, is the default. Also try 2 word combinations and 3 word combinations.
- Best results consistently came from CountVectorizer(ngram_range=(1,2)) using, 1 and 2 word combinations.

PREDICTIONS

- Question: Is a particular review helpful?
- Y is the helpful rating.
- Instead of predicting an exact rating, the data is split into helpful and unhelpful scores.
- Reviews with a helpful rating of over 85% are helpful.
- Reviews with a helpful rating of under 50% are not helpful.
- Leaving out the middle is justified because these reviews could go either way. User results may not be as accurate due to bias.