95-851 Making Products Count: Data Science for Product Managers
Homework 3: Natural Language Processing
Fall 2020

Due 11:59 PM October 7, 2020

Overview

In this assignment, you will practice using natural language processing techniques to understand customer feedback about the Kindle, Amazon's e-reader product. You will analyze a set of reviews extracted from Amazon's reviews for attributes of the Kindle that customers viewed positively and negatively, and build model to predict 5-star ratings based on the text in the review, and identify attributes of the reviews themselves that lead to other customers finding that review helpful. The submission will be a Jupyter Notebook with the name *DSPM_HW3_<your Andrew id>.ipynb*. The deadline for the homework is **October 7, 2020 11:59 PM EST.**

Tasks

The data file for you to analyze is in the assignment in Canvas as kindle_reviews.csv.

1) You'll be creating models to predict the reviews that lead to recommendations for the product, 5-star ratings, and reviews being considered helpful, which are found in the field reviews.doRecommend, reviews.rating, and reviews.numHelpful respectively. Create summary statistics and histograms for each of these fields. Do you see any issues in using these fields as outcome (target) variables? (3 points)
2) Prepare the text of the reviews in reviews.text field for analysis by eliminating stopwords. What are the top 10 most frequent words? What are the top 10 nouns? What are the top ten adjectives? (3 points)
3) What are the top ten most frequent words in reviews that do not recommend purchase of the Kindle? (3 points)
4) Create a model that will predict when a customer will give the Kindle a 5-star rating based on the text of the review. Evaluate the accuracy of your model. (3 points)
5) Create a model that will predict when at least two customer will find a review helpful. Evaluate the accuracy of your model (3 points)

HW3 Resources

- Learning to classify text can be found at http://www.nltk.org/book/ch06.html
- This is a subset of the full dataset available about the products, the schema for which can be found in https://developer.datafiniti.co/docs/product-data-schema .