What features of the data did you use to create your model?
Problem 1: Text as amazon_food_review.csv
Problem 2: Text from amazon_food_review.csv
Problem 3: Latitude and Longitude from yelp_academic_dataset_bussiness.json

How did you evaluate the performance of your model?
Problem 1: MulticlassMetrics to output the confusionMatrix and precision.
Problem 2: RegressionMetrics to output the explainedVariance and rootMeanSquaredError.
Problem 3: WSSSE to output the number of clusters.

What was the performance of your model? (i.e. using RegressionMetrics, MulticlassMetrics, or WSSSE)
Problem 1:
confusion matrix: [[1022. 1031.]
[648. 8591.]]
precision: 0.8518127419922562
evaluate result: 0.8573058506165904
Problem 2:
(Training)
explainedVariance: 0.668552281869132
rootMeanSquaredError: 0.817647361503431
(Test)
explainedVariance: 0.6562758618150969
rootMeanSquaredError: 0.8101056782929746

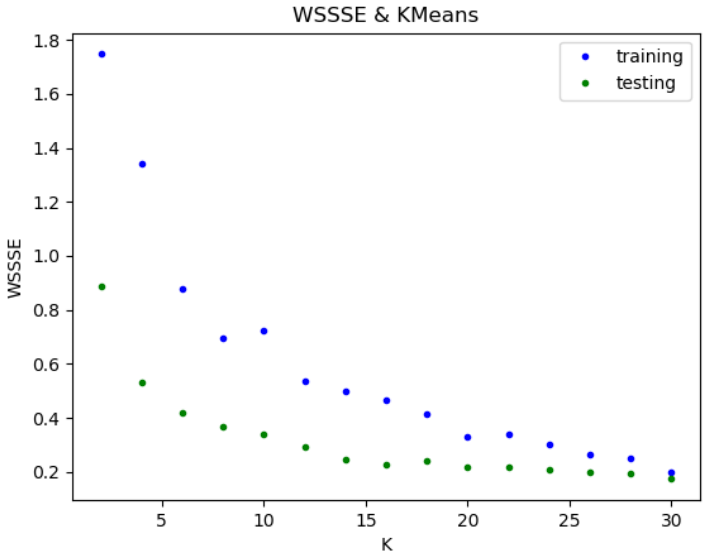
Problem 3:
(Training)
WSSSE: 5.99111680758
(Test)
WSSSE: 1.42005266289

What parameters did you change to try to improve the performance of your model? What was the effect of changing these parameters?

Problem 1: Previously, I divided the scoring into 5 different classes instead of 2. The output was around 0.6273503954277962. To optimize the prediction, I change is back to 2 classes increases sample size to 0.1, which provides 0.8574491213470694.

Problem 2: Using different minDocFreq parameter in IDF function, number of iterations and step size parameter in LinearRegressionWithSGD.train function doesn't provide significant changes of the result. Since we tend to ignore small number, I implement a threshold to filter out small numbers from HelpfulnessDenominator. It decreases the rootMeanSquaredError by 0.04.

Problem 3: I ran a simple loop to sample the model's performance under different K value by WSSSE (result shown below). As K increases, WSSSE decreases. After K > 15, the slope of the decreasing is converging to 0.



Left is model prediction on training data , right is the prediction on testing data. K for both figures is 3