HW1 Report - Menghan Xu
Python version: 3.8.8

Brief:

Loaded data from gz file directly and suppressed bad line warning. Before sampling 200,000 reviews, Idropped rows with empty reviews so they wouldn't get selected.

When sampling, I used random seed for dataset consistency. BeautifulSoup is easy to use with html parser along with regex substitution, I was able to remove the html tags and any URL from the review body.I was not able to install contraction or pycontractions using pip, so I had to create a json based on the listed contractions in English from Wikipedia as a dictionary for contraction, then try to replace any match in each review. lowercase is straightforward with str.lower(), and I used regex again for remove non alphabetical and extra whitespaces. The steps in cleaning were done in such an order that the contractions don't get altered, such as "i'm" becomes "im".

For pre-processing, I used nltk stop word library to remove any stopwords and join the rest of the words from review with a single whitespace. For lemmatization, I first tried to look up the POS tag for each word, then lemmatize it with the tags, but it took about 10 mins each time for the 200,000 reviews, and after comparing the model results and feature counts, I turned to just passing the word to lemmatize(), which is much faster and results are nearly the same.

I did not split the training and testing dataset until the pre-processing was done, and I only used 160,000 train data with TfidfVector, because in real world problems, we cannot train with the data(test data in this case) we haven't collected. The total features turned out to be just above 50,000, also only single gram.

The model training steps are similar by fitting the vector matrix and classification to the model and predict on both train and test data set. I used learning rate 0.1 in perceptron and 1000 iterations, the result was fast so I did not experiment more on it. The SVM model fit was extremely slow with default options, so I turned to LinearSvm with tolerance 1e-5 for faster completion. In logistic regression, I just used 500 iterations. For MNB, I just kept it default.

Outputs:

Statistics of three classes, (positive, star=3, negative) :
3856492,349547,668848
Raw data classification:
Positive review: 3856492
Negative review: 668848

Average length of reviews before and after data cleaning:
324.359335,310.915845

Average length of reviews before and after data preprocessing:
310.915845, 191.15878

Three sample reviews in raw format:

1. Very Nice!! Exactly as described!!!!!
2. Easy to use. Ice cream in 20 minutes, couldn't be better.
3. Even though this product works great, the edge can break or crack easily. The edge is where the glass doubles back to create the space between the two walls. When you have any liquids in the cup, all that weight is being cantilevered from the outer wall to the inner wall, i.e. suspended. So, if you happen to bump or shake the cup while it is full, that puts a lot of stress on that edge. I've had 1 cup chip. I also bought the water jug but that one just broke when I bumped it against the sink after filling it. Yea, the worst possible scenario. So, if you are willing to pay for the great looks and functionality buy this but do expect that they won't last as other cups that cost less.

The double wall really does insulate the contents, hot or cold. So, your contents last longer and your hand does not get hot or cold. Great for ice cream and tea. Their tea pot is also very good. I haven't broken that one.....yet.

Three sample reviews after cleaning & preprocessing:
1. nice exactly described
2. easy use ice cream minute could better
3. even though product work great edge break crack easily edge glass double back create space two wall liquid cup weight cantilevered outer wall inner wall e suspended happen bump shake cup full put lot stress edge cup chip also bought water jug one broke bumped sink filling yea worst possible scenario willing pay great look functionality buy expect last cup cost le double wall really insulate content hot cold content last longer hand get hot cold great ice cream tea tea pot also good broken one yet

Model report:

Accuracy, Precision, Recall, and f1-score for training and testing split for Perceptron:
0.89514375,0.8751737658171449,0.9215438508695108,0.8977604436454494,0.8549,0.8358918817103033,0.8844544095665172,0.8594877257541278

Accuracy, Precision, Recall, and f1-score for training and testing split for SVM:
0.9324,0.9334688040942274,0.9310396597022395,0.93225264951269,0.8939,0.894899690587883,0.8934728450423518,0.894185698613743

Accuracy, Precision, Recall, and f1-score for training and testing split for Logistic Regression:
0.91286875,0.9161820910960027,0.9087201301138497,0.9124358547569547,0.896475,0.8996437352601736,0.8933233682112606,0.8964724118102952

Accuracy, Precision, Recall, and f1-score for training and testing split for Naive Bayes:
0.88603125,0.8907566218664099,0.8797572876266734,0.8852227880130671,0.8674,0.8739363857374393,0.8597409068261086,0.8667805294619984