Homework 3 : Text Analytics
Megan Hazlett
Week of October 16, 2020

Github: https://github.com/MSIA/mlh9188_msia_text_analytics_2020

Download the yelp data set, unzip it, and put all files in a folder called yelp within the repo. You will be using the file yelp/yelp_academic_dataset_review.json

Question 1

The data set I am using is the yelp reviews. There are 8 million documents in this data set. The following is for the first 50000 entries: It's labels are 1-5 stars, with the most frequent label being 5 and the least frequent rating being 2. The average number of characters in a review is 571.43 with a range of 2 to 5000. The average number of words in the first 500 entries is 105.59.

Question 2

For building multiple logistic regression models to predict the number of stars for each yelp review, I modeled unigrams, bigrams, and a combination of unigrams. I also tried a few different hyperparameters including k, the number of important features to use and varying what method was used to select these important features. The following table contains my results.

Note: Was only able to train in 50,000 due to capacity issues.

Model	Features	k	Norm	Accuracy	F1 [1star, 2star, etc.]
1	Unigram	120	L2	61%	[0.71,0.13,0.26,0.28,0.76]
2	Bigram	120	L2	58%	[.68, 0.06, 0.25, 0.32, 0.73]
3	Unigram+bigram	120	L2	48%	[.37,.0.02,0.04,0.08,0.63]
4	Unigram	1200	L2	56%	[0.65, 0.02, 0.17, 0.29, 0.72]
5	Bigram	1200	L2	72%	[0.74,0.16,0.26,0.42,0.77]

6	Unigram+bigram	1200	L2	50%	[0.52,0.04,0.08,0.16,0.65]
7	Bigram	1200	L1	44%	[0,0,0,0,0.61]
8	Unigram+bigram	1200	L1	44%	[0,0,0,0,0.61]

Question 3

For building multiple support vector machine models to predict the number of stars for each yelp review, I modeled unigrams, bigrams, and a combination of unigrams. I also tried a few different hyperparameters including k, the number of important features to use and varying what method was used to select these important features. The following table contains my results.

Note: Was only able to train in 50,000 due to capacity issues.

Model	**Features**	**k**	**Max Iterations**	**Accuracy**	**F1 [1star, 2star, etc.]**
1	Unigram	120	50	53%	[0.61,0.04, 0.06, 0.08,0.69]
2	Bigram	120	50	50%	[0.49, 0.12,0.09, 0.07, 0.66]
3	Unigram+bigram	120	50	52%	[0.58,0.05, 0.04,0.10, 0.68]
4	Unigram	1200	100	55%	[0.62, 0.09, 0.13, 0.10, 0.70]
5	Bigram	1200	100	53%	[0.58,0.11,0.13,0.11,0.68]
6	Unigram+bigram	1200	100	53%	[0.60,0.07,0.14,0.08,0.69]
7	Bigram	120	100	50%	[0.47,0.12,0.08,0.10,0.66]
8	Unigram+bigram	1200	100	54%	[0.62,0.1,0.12,0.1,0.69]

Question 4

See code in github

Used mod 8 to make predictions.