HomeWork 8

Lemmatising and stemming is the process of grouping all the words which typically mean the same and and have different verb form tense or singular/plural. Lemmatising is converting all the influenced words into its root word and Stemming focusses on converting all the influenced words into one word (May or may not be the root word) and after analysing both the accuracies of stemming and Lemmatising

Unigram Lemmatised Reviews :

```
“original_clean_reviews=review_cleaner(train['review'],lemmatize=True,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews,y=train[“sentiment"],ngram=1,max_features=1000)"
```

The training accuracy is: 0.99995

The Validation accuracy is : 0.8314

Ten important fetures:

['bad', 'worst', 'great', 'waste', 'awful', 'excellent', 'best', 'terrible', 'boring', 'nothing']

Unigram stemmed Reviews :

```
“original_clean_reviews=review_cleaner(train['review'],lemmatize=False,stem=True)
train_predict_sentiment(cleaned_reviews=original_clean_reviews,y=train[“sentiment"],ngram=1,max_features=1000)"
```

The training accuracy is: 0.99995

The Validation accuracy is : 0.8238

Top ten important features:

['worst', 'bad', 'wast', 'great', 'aw', 'bore', 'excel', ‘love', 'terribl', 'noth']

Bigram Lemmatised Reviews :

The training accuracy is: 1.0

The Validation accuracy is : 0.8238

Top ten important features:

['bad', 'worst', 'great', 'waste', 'awful', 'terrible', 'boring', 'excellent', 'nothing', ‘worse']

A1: When it comes to unigram approach we have noticed that when it comes to train accuracies stemmed reviews have higher accuracy than both the original and lemmatised reviews but that does not nessaraly mean anything since when it comes to validation accuracy lemmatised accuracy is higher than the other 2 and stemmed is less than the original that means lemmatising has actually made the model better and stemming hasn't really done any good.

A2: Now bigram models are actually better and have improved the models but there is no significant difference in either lemmatising or stemming and hence both of them give nearly similar accuracies. However, there is a change in the top 10 important features that means that now different words have more weight in the sentiment.

A3: From changing the max_features from 10 to 5000, we have noticed that there is always a increase in the accuracy with increase in max-features but while comparing it is noticed that there is a significant rise from 10 to 100 and 100 to 1000 but only a 2% increase in the accuracy from 1000 to 5000. and hence i suppose that 1000 will be the most optimum value of max_features