Homework 9
Jeff Peng
3034340183

1.
Unigram without lemmatization and stemming
Training accuracy: 0.9999
Validation accuracy: 0.8216

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

Unigram with lemmatization but without stemming
Training accuracy: 1.0
Validation accuracy: 0.8184

```
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)
```

Unigram without lemmatization but with stemming
Training accuracy: 1.0
Validation accuracy: 0.82

```
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)
```

2.
Bigram without lemmatization and stemming
Training accuracy: 0.99995
Validation accuracy: 0.8208

```
original_clean_reviews=review_cleaner(train['review'],lemmatize=False,stem=False)
train_predict_sentiment(cleaned_reviews=original_clean_reviews,
y=train["sentiment"],ngram=2,max_features=1000)
```

Bigram with lemmatization but without stemming
Training accuracy: 0.99995
Validation accuracy: 0.8184

```
original_clean_reviews=review_cleaner(train['review'],lemmatize=True,stem=False)
```

```
train_predict_sentiment(cleaned_reviews=original_clean_reviews,
y=train["sentiment"],ngram=2,max_features=1000)
```

Bigram without lemmatization but with stemming
Training accuracy: 1.0
Validation accuracy: 0.8264

```
original_clean_reviews=review_cleaner(train['review'],lemmatize=False,stem=True)
train_predict_sentiment(cleaned_reviews=original_clean_reviews,
y=train["sentiment"],ngram=2,max_features=1000)
```

3.
max features=10
Training accuracy: 0.87125
Validation accuracy: 0.5648

```
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=10)
```

---

max features=100
Training accuracy: 0.99985
Validation accuracy: 0.7206

```
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=100)
```

---

max features=1000
Training accuracy: 1.0
Validation accuracy: 0.8184

```
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)
```

---

max features=5000
Training accuracy: 1.0
Validation accuracy: 0.8352

```
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=5000)
```

For original cleaned reviews using both unigram and bigram, the top ten most important features included similar words, ex: “worst” and “worse”; for lemmatized reviews for unigram and bigram, the most important features included more words, since “worst” and “worse” were lumped together; for stemmed reviews using unigram, the top ten most important features included words that are missing the ending letters, such as “wast” and “terribl”, as well as “aw” which should actually be “awful,” while using bigram resulted in “wast[e]” and “wast[e] time,” which is redundant. The validation accuracy for all scenarios under unigram and all scenarios under bigram is similar.

As the number of max_features increased, so did the validation accuracy. However, the difference between 1000 and 5000 max_features is insignificant, and the only difference between the top ten important features for each is the first has “wonderful” while the second has “worse” instead, which is actually less helpful as “worst” and “worse” practically mean the same thing.