WORCESTER POLYTECHNIC INSTITUTE
DATA SCIENCE PROGRAM

Case Study 3

Review Sentiment Analysis

SUBMITTED BY
Sirshendu Ganguly
Enbo Tian
Dang Tran

Date Submitted	**: 4/06/2022**
Date Completed	**: 4/06/2022**
Course Instructor	**: Prof. Ngan**

Motivation and Background

Before this case study, we have already talked about the scores of movies from 1995 to 2000. However, we can not decide if the score is effective or not. People may give a careless review to a movie when they don't pay attention, and whatever score they give to a movie, they don't have the duty on the score. Thus, if we want to drop the illusory score to movies from people, it is important to consider the written reviews of movies from them.

Furthermore , when we consider the reviews of movies, we noticed that people may have a lack of consideration for the real quality of a movie, but go with the tide. It is also important to drop these kind of written reviews to make the review of a movie more objectively.

Data Sources

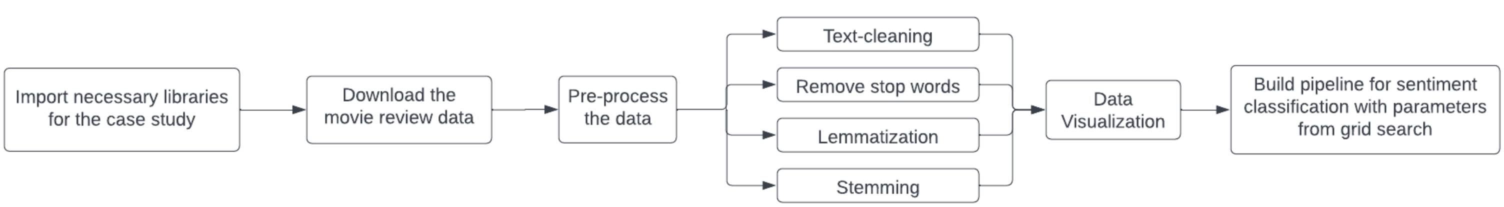
In this case study, we used the v2.0 polarity movie_reviews dataset from http://www.cs.cornell.edu/people/pabo/movie-review-data. This dataset contains 2000 movie reviews in a "txt_sentoken" folder with them being separated by sentiment. The reviews with a positive sentiment are placed in the "pos" subdirectory and the ones with negative sentiment were placed in the "neg" subdirectory.

Various methods were used to determine if a review was positive or negative. With a 5-star system, 3.5 stars and above are considered positive while 2 stars and below are considered negative. With a 4-star system, 3 stars and up are considered positive while 1.5 stars and below are considered negative. With a letter grade system, B or above is considered negative and C- or below is considered negative.

In the dataframe of the dataset, we have review and sentiment. "review" is the raw html text of a review and sentiment can be either 1 or 0 which corresponds to a positive or negative sentiment accordingly.

Methodology

Problem 1:



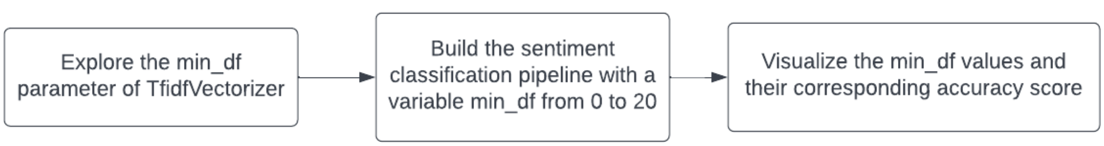
- To begin, we needed to import some necessary libraries. These include nltk, sys, re, numpy, pandas, matplotlib.pyplot, seaborn, and sklearn.
- Downloaded the movie review into a "txt_token" folder using the provided python script.
  - Loaded the dataset using sklearn load_files then put it into a panda DataFrame.
  - Set the columns to be "Review" and "Sentiment".
- Looking at a preview of the dataset, it's clear pre-processing is necessary.
  - Created a function to clean to clean the text
    - Removed backslash-apostrophe.
    - Removed everything except the letters in the alphabets.

        - Removed unnecessary whitespaces.
        - Converted the text to lowercase.
    - Created a function to remove the stopwords
        - Imported a list of english stopwords from nltk.corpus.
    - Created a function to lemmatize the dataset
        - Used the lemmatizer from nltk.stem.
    - Created a function to stem the dataset
        - Used the stemmer from nltk.stem.
- Visualized the word frequency of the dataset with a horizontal bar graph.
- Develop sentiment classification pipeline.
    - Split the dataset into training and testing data sets with sklearn train_test_split.
    - Create a panda dataframe for the training set for data exploration.
        - Made a countplot of the sentiments behind movies' reviews using seaborn and matplotlib.

```
plt.figure(figsize=(10,6))
sns.set_theme(style="darkgrid")
sns.countplot(x=train_df.sentiment)
plt.title("Countplot of classes of sentiments behind movie's review")
plt.show()
```

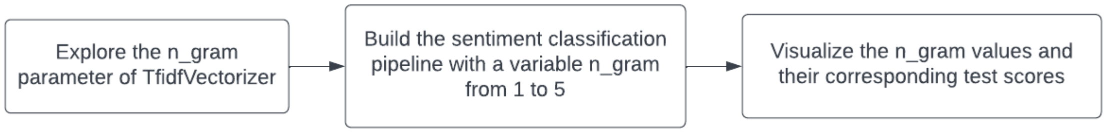
        - Created a new feature "Length" of the number of characters in each review.
            - Plotted a histogram of the reviews' length.
    - Build a pipeline with sklearn Pipeline.
    - Define the n_gram parameter.
    - Used grid search to find the optimal parameters.
    - Predict the test values with grid_search.predict().
    - Print out the classification report.
    - Plot out the confusion matrix.
    - Print out the accuracy score and F1 score of our pipeline.

Problem 2:



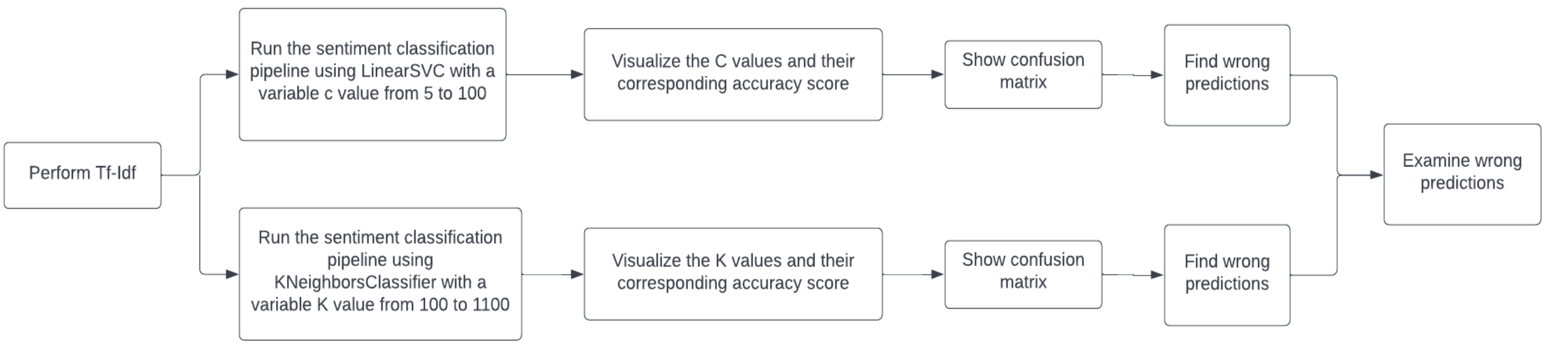
- Used a for loop to build the sentiment classification pipeline with various min_df values.
    - Record the min_df value and the corresponding accuracy score into separate arrays in each iteration.
- Create a panda DataFrame of min_df values and their corresponding accuracy scores.
    - Create a line plot of min_df values vs accuracy scores using matplotlib and seaborn.
    - Get the min_df value with the highest accuracy score.

- Used the same approaches above to explore the max_df parameter.



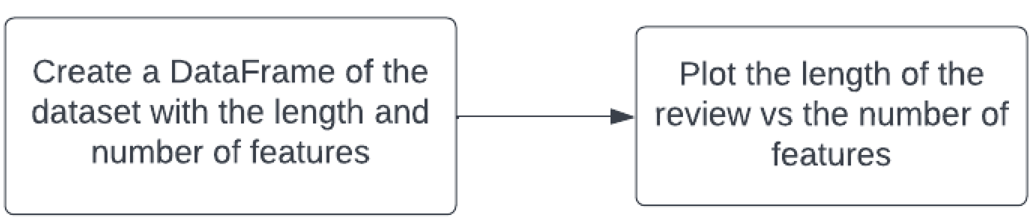
- Manually set 5 different n_gram parameters from 1, 1 to 1,5.
- Do a gridsearch with all the different parameters.
    - Print out all the n_gram parameters and their corresponding test scores.
- Print out the accuracy score using the best n_gram param.

Problem 3:



- Performed Tf-Idf
    - Used the best parameters found in Problem 2.
    - Computed Xtrain using fit_transform on docs_train.
    - Computed Xtest using transform on docs_test.
- Used the same methods in Problem 2 when exploring the min_df and max_df values on the C value when running the pipeline with LinearSVC.
    - Shows the accuracy and F1 score along with the confusion matrix for LinearSVC.
    - Find wrong predictions by getting the indices where the test dataset doesn't match with the prediction dataset
- Used the same method above for KNeighborsClassifier with different K values
- Examine wrong predictions
    - Shows the number of wrong predictions using SVC and KNN.
    - Find records where both classifiers failed to make a prediction.

Problem 4:



- Create a function that calculates the number of features each reviews
- Create a new array for the number of features for all reviews
    - Iterates through all the rows of the dataset and calculates the number of features using the function created above.
- Add this array as a new column to the DataFrame.
- Used seaborn and matplotlib to make a scatterplot of the length vs number of features.

Results

Problem 1:

The original review data we connected:

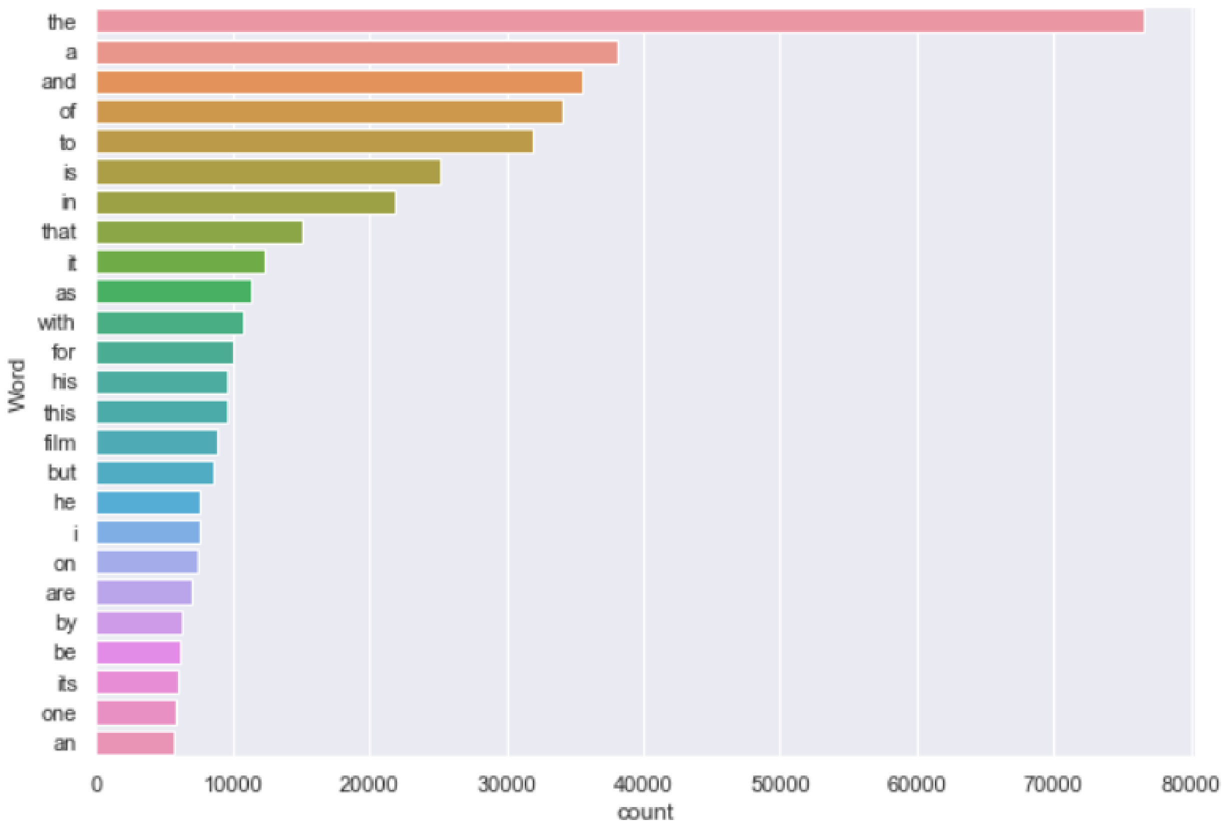
	review	sentiment
0	b'plot : two teen couples go to a church party...	0
1	b'the happy bastard\'s quick movie review \nda...	0
2	b"it is movies like these that make a jaded mo...	0
3	b' " quest for camelot " is warner bros . \' f...	0
4	b'synopsis : a mentally unstable man undergoin...	0

```
Snapshot of data['review']

0       b'plot : two teen couples go to a church party...
1       b'the happy bastard\'s quick movie review \nda...
2       b"it is movies like these that make a jaded mo...
3       b' " quest for camelot " is warner bros . \' f...
4       b'synopsis : a mentally unstable man undergoin...
                              ...
1995    b"wow ! what a movie . \nit's everything a mov...
1996    b'richard gere can be a commanding actor , but...
1997    b'glory--starring matthew broderick , denzel w...
1998    b'steven spielberg\'s second epic film on worl...
1999    b'truman ( " true-man " ) burbank is the perfe...
Name: review, Length: 2000, dtype: object
```

After cleaning the text:

```
Snapshot of data['review']

0       plot two teen couples go to a church party dri...
1       the happy bastards quick movie review damn tha...
2       it is movies like these that make a jaded movi...
3       quest for camelot is warner bros first feature...
4       synopsis a mentally unstable man undergoing ps...
                              ...
1995    wow what a movie its everything a movie can be...
1996    richard gere can be a commanding actor but hes...
1997    glory starring matthew broderick denzel washin...
1998    steven spielbergs second epic film on world wa...
1999    truman true man burbank is the perfect name fo...
Name: review, Length: 2000, dtype: object
```

Generate a graph to visualize the words and frequency in data's review:



Removing stop words:

```
0       plot two teen couples go church party drink dr...
1       happy bastards quick movie review damn k bug g...
2       movies like make jaded movie viewer thankful i...
3       quest camelot warner bros first feature length...
4       synopsis mentally unstable man undergoing psyc...
                              ...
1995    wow movie everything movie funny dramatic inte...
1996    richard gere commanding actor hes always great...
1997    glory starring matthew broderick denzel washin...
1998    steven spielbergs second epic film world war i...
1999    truman true man burbank perfect name jim carre...
Name: review, Length: 2000, dtype: object
```

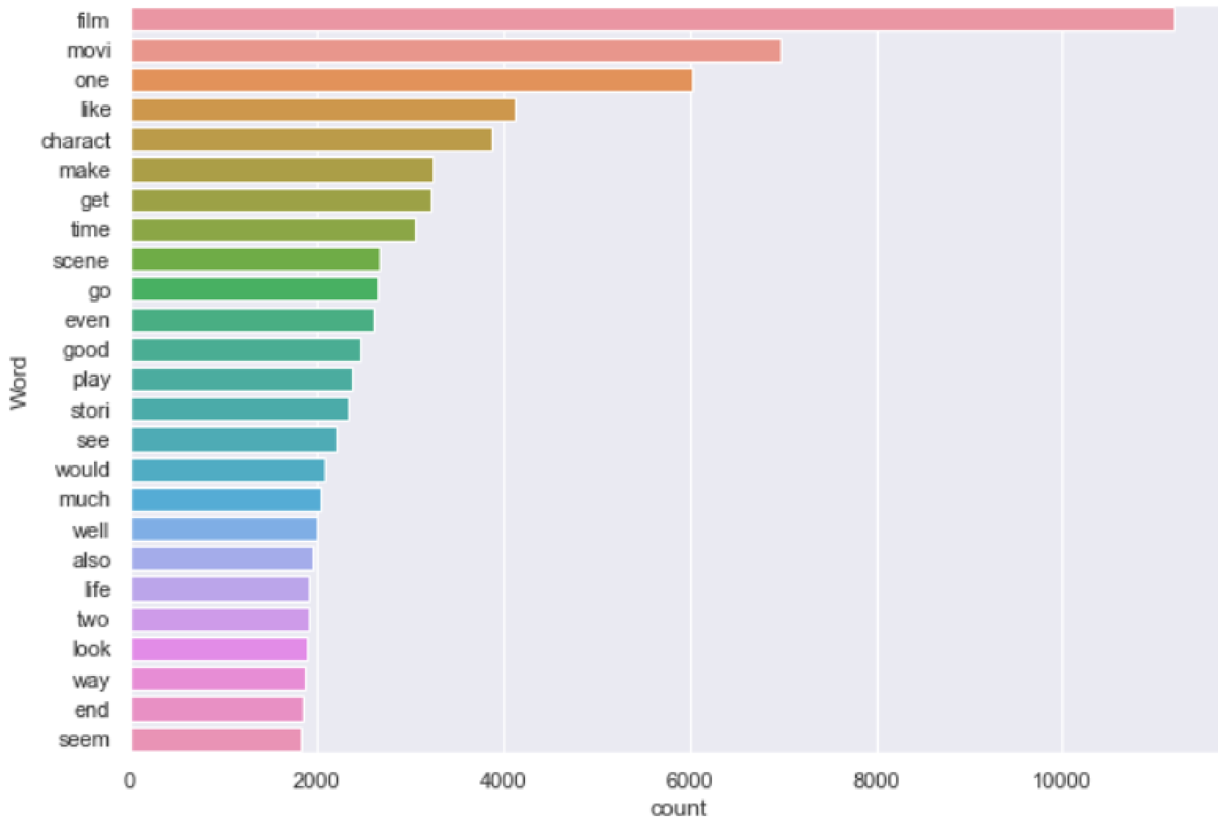
Lemmatization on Review:

```
0       plot two teen couple go church party drink dri...
1       happy bastard quick movie review damn k bug go...
2       movie like make jaded movie viewer thankful in...
3       quest camelot warner bros first feature length...
4       synopsis mentally unstable man undergoing psyc...
                              ...
1995    wow movie everything movie funny dramatic inte...
1996    richard gere commanding actor he always great ...
1997    glory starring matthew broderick denzel washin...
1998    steven spielberg second epic film world war ii...
1999    truman true man burbank perfect name jim carre...
Name: review, Length: 2000, dtype: object
```

Stemming on Review:

```
0       plot two teen coupl go church parti drink driv...
1       happi bastard quick movi review damn k bug got...
2       movi like make jade movi viewer thank invent t...
3       quest camelot warner bro first featur length f...
4       synopsi mental unstabl man undergo psychothera...
                              ...
1995    wow movi everyth movi funni dramat interest we...
1996    richard gere command actor he alway great film...
1997    glori star matthew broderick denzel washington...
1998    steven spielberg second epic film world war ii...
1999    truman true man burbank perfect name jim carre...
Name: review, Length: 2000, dtype: object
```

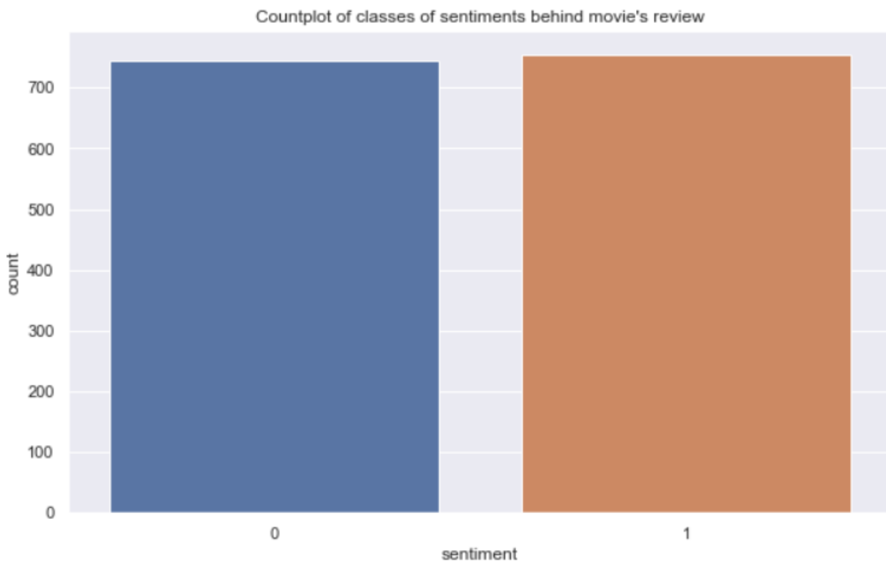
Generate a graph to visualize the words and frequency in data's review:



We split the data into training and test set, and shows the first three review in the training set:

	review	sentiment
0	well ill damn excel surpris confus campi far i...	1
1	gere willi poitier chase around world jackal f...	1
2	well guess time year one time year movi crafto...	0

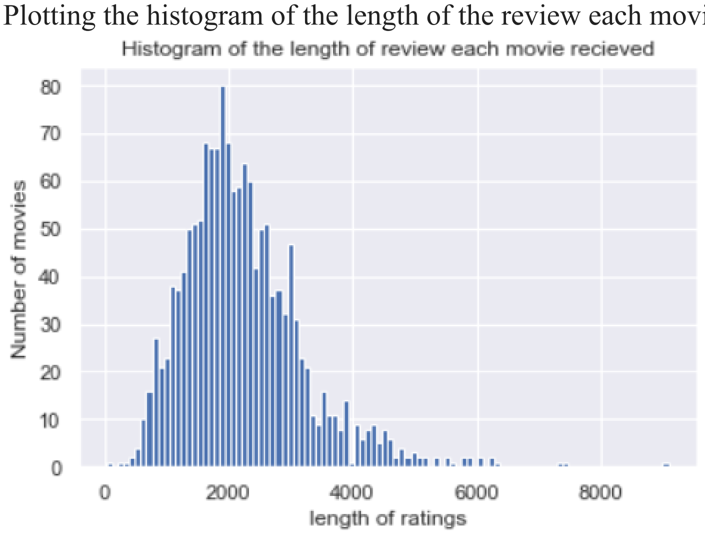
We plot the count plot of the different classes of reviews present:



We created a new feature 'length' which is the length of each movie review:

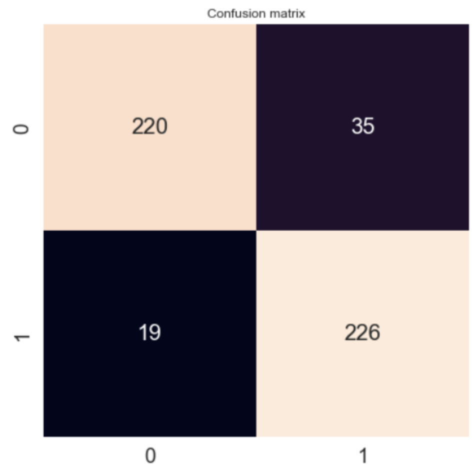
	review	sentiment	length
0	well ill damn excel surpris confus campi far i...	1	1872
1	gere willi poitier chase around world jackal f...	1	1940
2	well guess time year one time year movi crafto...	0	2904
3	filmcrit com colleagu norm schrager nail sessi...	1	2864
4	plot someth bunch kid go haunt hous play parod...	0	1816
5	american pie acknowledg cold hard fact movi do...	1	1805
6	note may consid portion follow text spoiler fo...	1	4359
7	high school comedi seem hot genr moment she va...	0	1889
8	stephen pleas post appropri mafia crime isnt f...	0	1338
9	sequel dont theyr suppos like toy stori far ma...	1	3278

Plotting the histogram of the length of the review each movie received:



We then predicted the test values and showed the confusion matrix:

```
              precision    recall  f1-score   support

         neg       0.92      0.86      0.89       255
         pos       0.87      0.92      0.89       245

    accuracy                           0.89       500
   macro avg       0.89      0.89      0.89       500
weighted avg       0.89      0.89      0.89       500
```



The accuracy score is: 89.2 %

The F1 score is: 89.32806324110672 %

Problem 2:

min_df:

	min_df value	accuracy score
0	0	0.874
1	1	0.874
2	2	0.890
3	3	0.884
4	4	0.878

max_df:

	max_df value	accuracy score
0	250	0.864
1	300	0.874
2	350	0.872
3	400	0.884
4	450	0.886

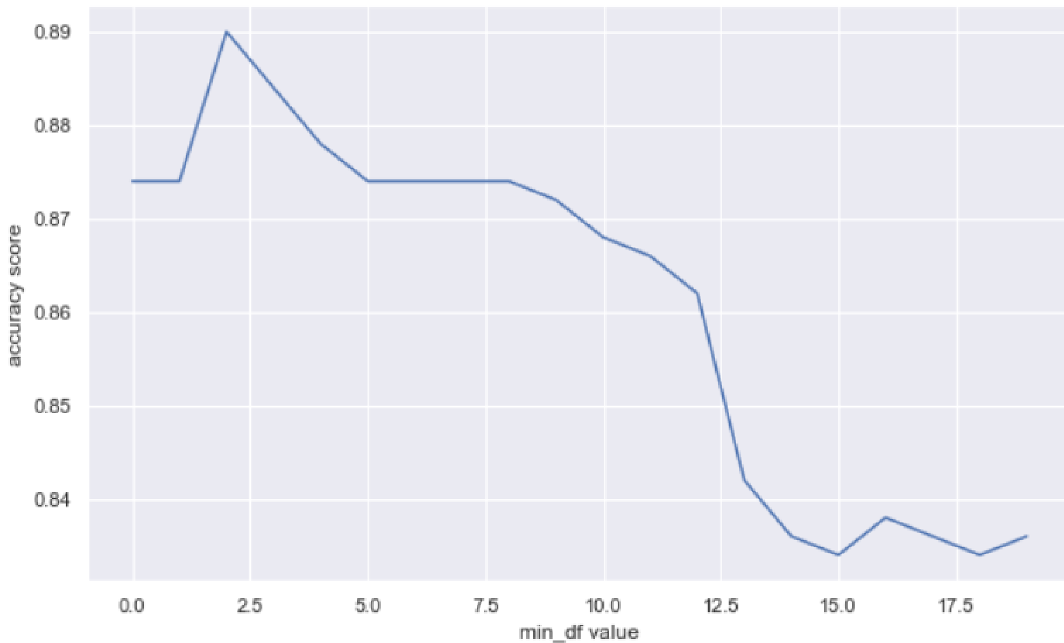
Finding the max value of accuracy for a given min_df value

	min_df value	accuracy score
2	2	0.89

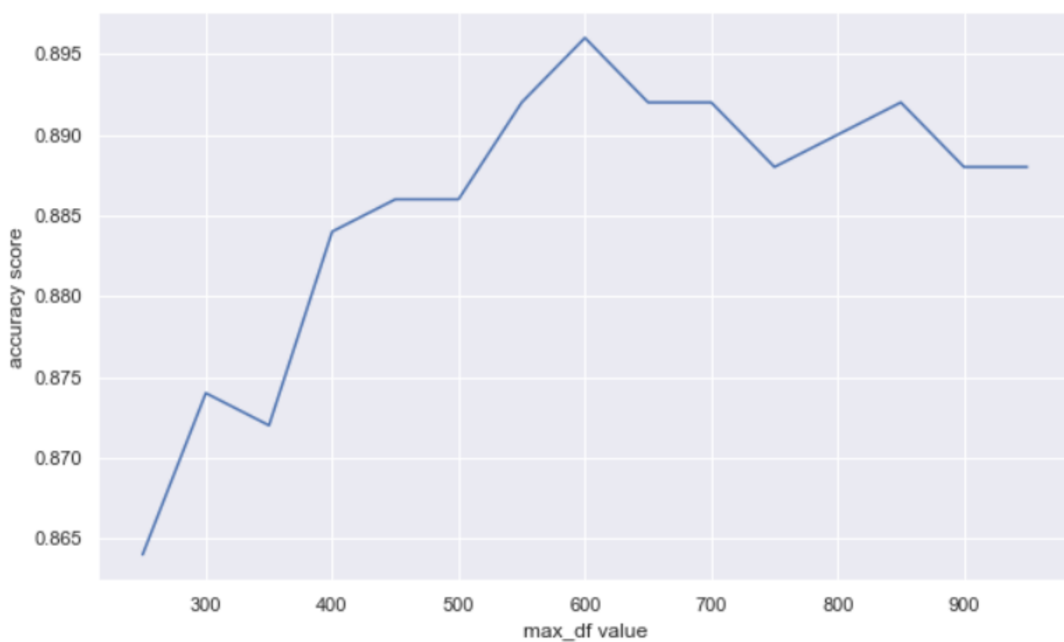
Finding the max value of accuracy for a given max_df value

	max_df value	accuracy score
7	600	0.896

Line plot of min_df vs accuracy score:



Line plot of min_df vs accuracy score:



Exploring the n_gram parameter of TfidfVectorizer:

```
0 params - {'vect__ngram_range': (1, 1)}; mean - 0.80; std - 0.02
1 params - {'vect__ngram_range': (1, 2)}; mean - 0.82; std - 0.02
2 params - {'vect__ngram_range': (1, 3)}; mean - 0.82; std - 0.02
3 params - {'vect__ngram_range': (1, 4)}; mean - 0.82; std - 0.02
4 params - {'vect__ngram_range': (1, 5)}; mean - 0.82; std - 0.02


The best accuracy score by n-gram param is:  89.60000000000001 %
```

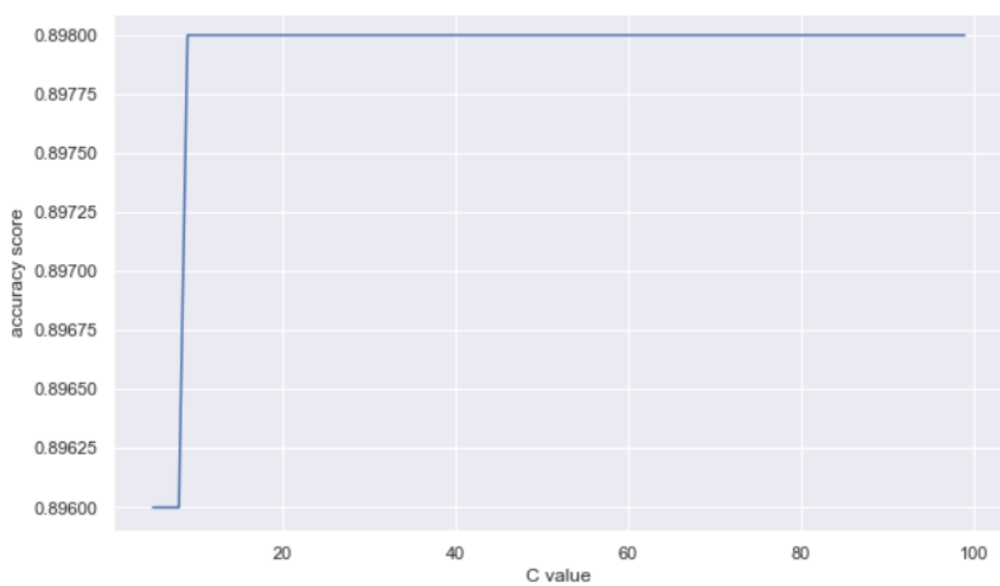
The first 5 n_gram are:

	n_gram	mean accuracy score
0	{'vect__ngram_range': (1, 1)}	0.798
1	{'vect__ngram_range': (1, 2)}	0.819333
2	{'vect__ngram_range': (1, 3)}	0.821333
3	{'vect__ngram_range': (1, 4)}	0.820667
4	{'vect__ngram_range': (1, 5)}	0.821333

Problem 3:

For LinearSVC

Plotting a line plot of k value vs accuracy score:



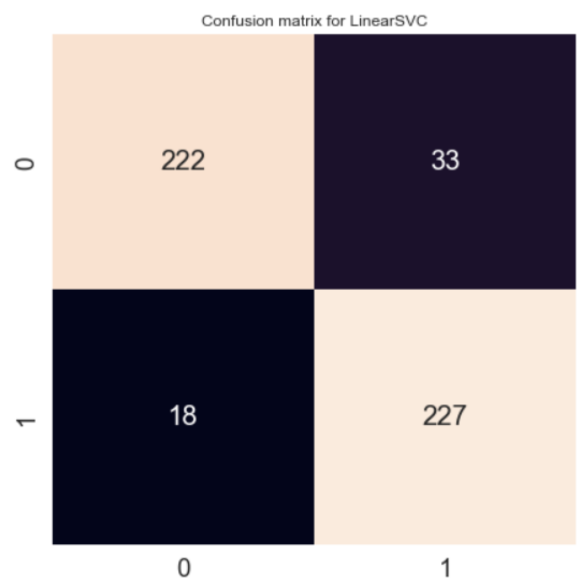
Finding the max accuracy score of k value:

	C value	accuracy score
4	9	0.898
5	10	0.898
6	11	0.898
7	12	0.898
8	13	0.898
...	...	...
90	95	0.898
91	96	0.898
92	97	0.898
93	98	0.898
94	99	0.898

The accuracy score is: 89.8 %.

The F1 score is: 89.72332015810277 %.

The confusion matrix for LinearSVC is:

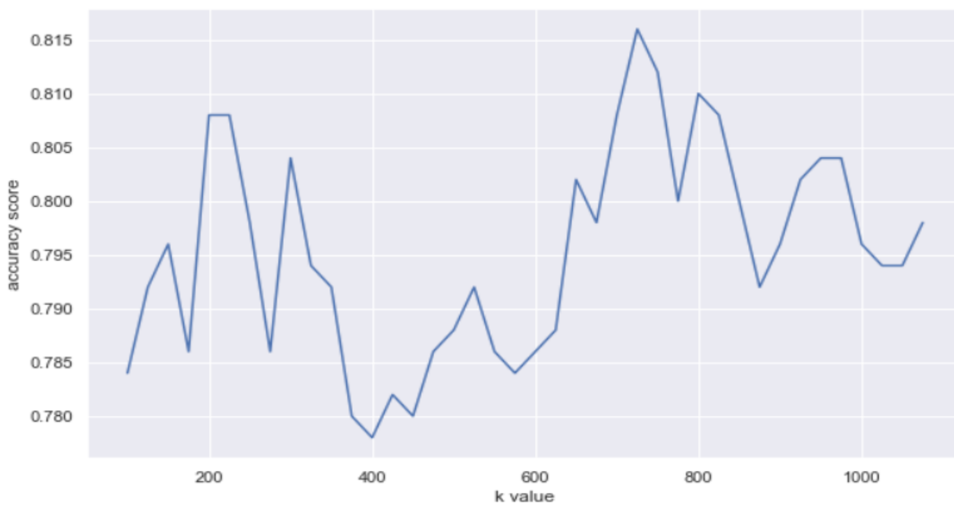


We find the wrong prediction:

	review	sentiment
7	that exactli long movi felt werent even nine l...	0
15	john carpent make b movi alway halloween escap...	0
36	among multitud erot thriller releas earli woma...	0
40	lengthi lousi two word describ bore drama engl...	0
42	pseudo intellectu film pseudo intellectu world...	0

For KNeighborsClassifier

Plotting a line plot of k value vs accuracy score:



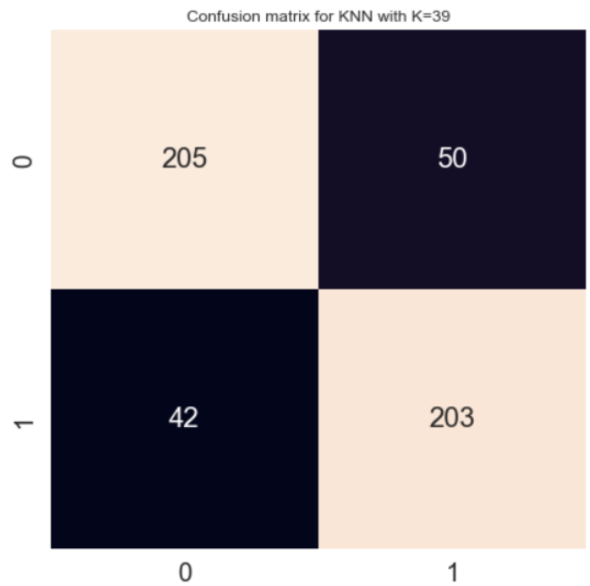
Finding the max accuracy score of k value:

	k value	accuracy score
25	725	0.816

The accuracy score is: 81.6 %.

The F1 score is: 81.52610441767068 %

The confusion matrix for KNeighborsClassifier is:



The wrong prediction:

	review	sentiment
7	that exactli long movi felt werent even nine l...	0
12	high fli hong kong style filmmak made way clas...	0
15	john carpent make b movi alway halloween escap...	0
28	follow disney live action dalmatian that bette...	0
34	your watch near two hour bore foul mouth flori...	0

We Find records where both classifiers failed to make prediction:

```
No of wrong predictions by SVC:  51
No of wrong predictions by KNN:  92
No of wrong predictions by both:  36
```

We can see that for all of these records where sentiment = 0 the predictions failed.

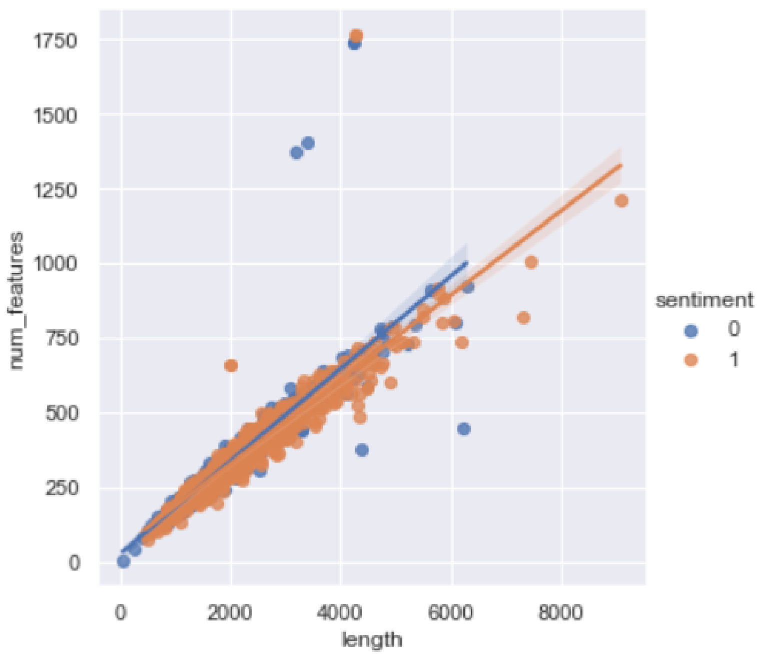
	review	sentiment
7	that exactli long movi felt werent even nine l...	0
15	john carpent make b movi alway halloween escap...	0
36	among multitud erot thriller releas earli woma...	0
40	lengthi lousi two word describ bore drama engl...	0
47	instinct kind movi inexperienc moviego undoubt...	0

Problem 4:

Creating a dataframe of features and length of review:

	review	sentiment	length	num_features
0	well ill damn excel surpris confus campi far i...	1	1872	330
1	gere willi poitier chase around world jackal f...	1	1940	336
2	well guess time year one time year movi crafto...	0	2904	512
3	filmcrit com colleagu norm schrager nail sessi...	1	2864	477
4	plot someth bunch kid go haunt hous play parod...	0	1816	344

Plotting the length of the review versus the number of features in that review:



Conclusion

In conclusion for problem 2, min_df vs accuracy score is a decreasing plot, which means the word not related to the review because of too infrequently is decreasing. The max_df vs accuracy is an increasing plot, which means the word not related to the review because of too frequently is increasing. The n-gram is a contiguous sequence of *n* items from the review. Since the accuracy is about 0.89, Using N-gram to keep multiplicity is useful to the reviews. In conclusion for problem 3, We get a larger value of accuracy score and F1 score on LinearSVC than K Neighbors Classifier. Also we have less type I and II error in LinearSVC than K Neighbors Classifier. Thus the method of LinearSVC is better than K Neighbors Classifier. The mistake of predictions might be because of Type II error, since all of the predictions are negative. In conclusion for problem 4, the plot of length and number of features are a very linear plot. We can say that they have a high relationship so we're unable to separate them using this method.