**CS5560 Knowledge Discovery Management**

**In Class Assignment # 10**

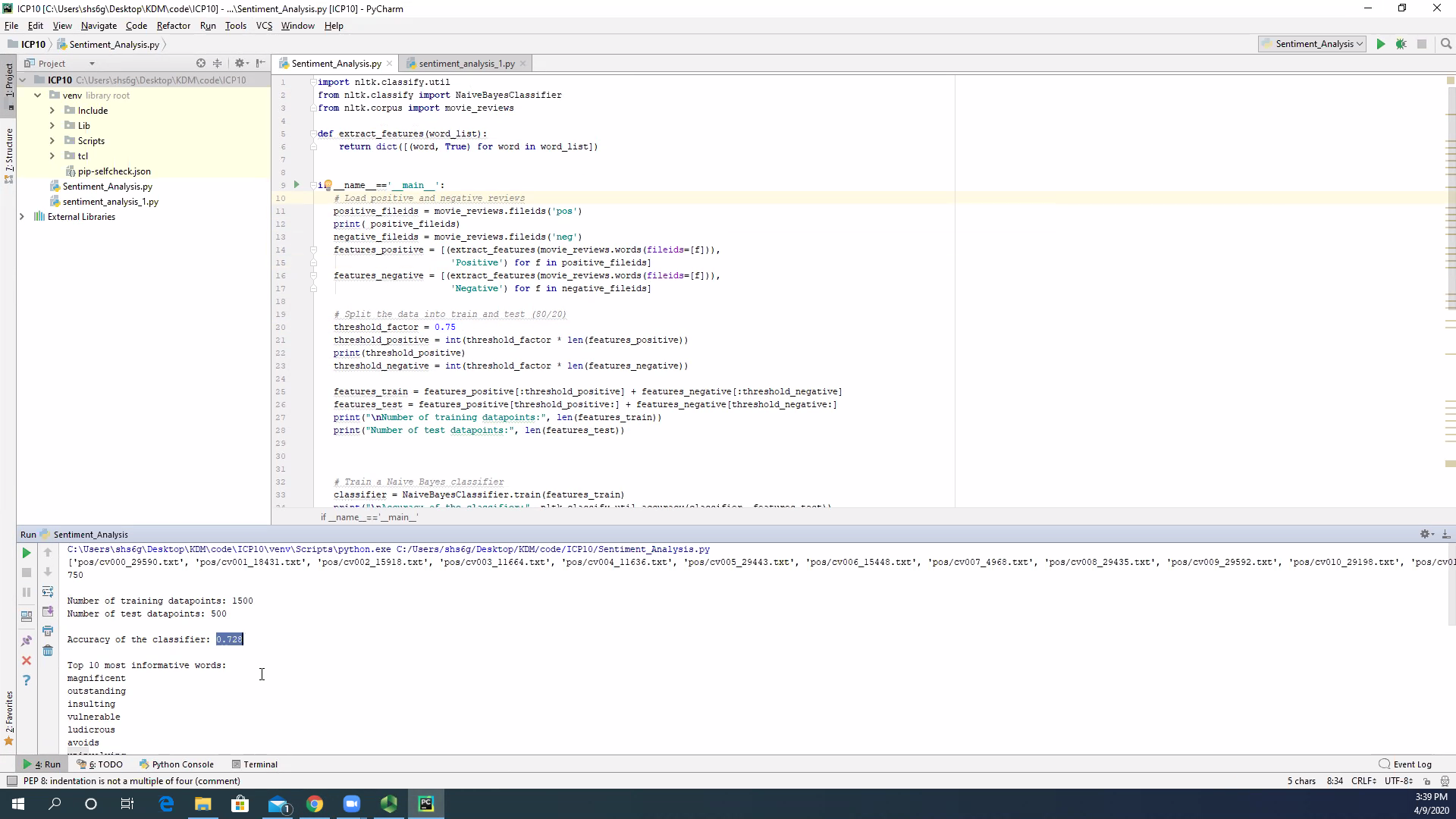
**9th April 2020**

Name: Srichakradhar Reddy Nagireddy

Class ID: 11

Perform the following:

1. Use the source code and change the train/test ratio and compare your findings with original train/test split (80/20). Explain why you are observing these changes (if any).



For 0.8:

‘Number of training datapoints:', 1600  
'Number of test datapoints:', 400  
'Accuracy of the classifier:', 0.735

For 0.75:

'Number of training datapoints:', 1500  
'Number of test datapoints:', 500  
'Accuracy of the classifier:', 0.728

For 0.9:

'Number of training datapoints:', 1800  
'Number of test datapoints:', 200  
'Accuracy of the classifier:', 0.73

For 0.6:

'Number of training datapoints:', 1200  
'Number of test datapoints:', 800  
'Accuracy of the classifier:', 0.766

For 0.5:

'Number of training datapoints:', 1000  
'Number of test datapoints:', 1000  
'Accuracy of the classifier:', 0.811

For 0.4:

'Number of training datapoints:', 1000  
'Number of test datapoints:', 1000  
'Accuracy of the classifier:', 0.78

**Observation:** The accuracy of the classifier is the highest at 50-50 split.

**Possible explanation**: Accuracy below 50% split is low because there is not enough data for the classifier to judge.

And, accuracy falls after the 50% split because, probably since the classifier distributes the probability over more words (as it considers more words) to judge, there can be misleading words that creep in.

1. Also change the validation reviews and run the model again. Report any changes you observe and explain your results.

**'Review**:', 'Star Wars is a junkyard of cinematic gimcracks not unlike the Jawas heap of purloined, discarded, barely functioning droids.'

('Predicted sentiment:', 'Positive')

('Probability:', 0.68)

**Possible explanation**: The classifier misclassified the review as positive because I guess, it couldn’t find any offensive words.

**'Review**:', "I'd rather wake up next to a severed horse head than ever watch Gotti again.")

('Predicted sentiment:', 'Negative')

('Probability:', 0.81)

**Possible explanation**: It has correctly classified this review may be because it found the word “severed”.

(**'Review**:', 'App seems to get worse with each update. Lots of bugs. Very good newspaper but very bad app.')

('Predicted sentiment:', 'Negative')

('Probability:', 0.73)

**Possible explanation**: Since this review has lot of negative words – like bad, worse, bugs, etc. It has identified the correct sentiment.