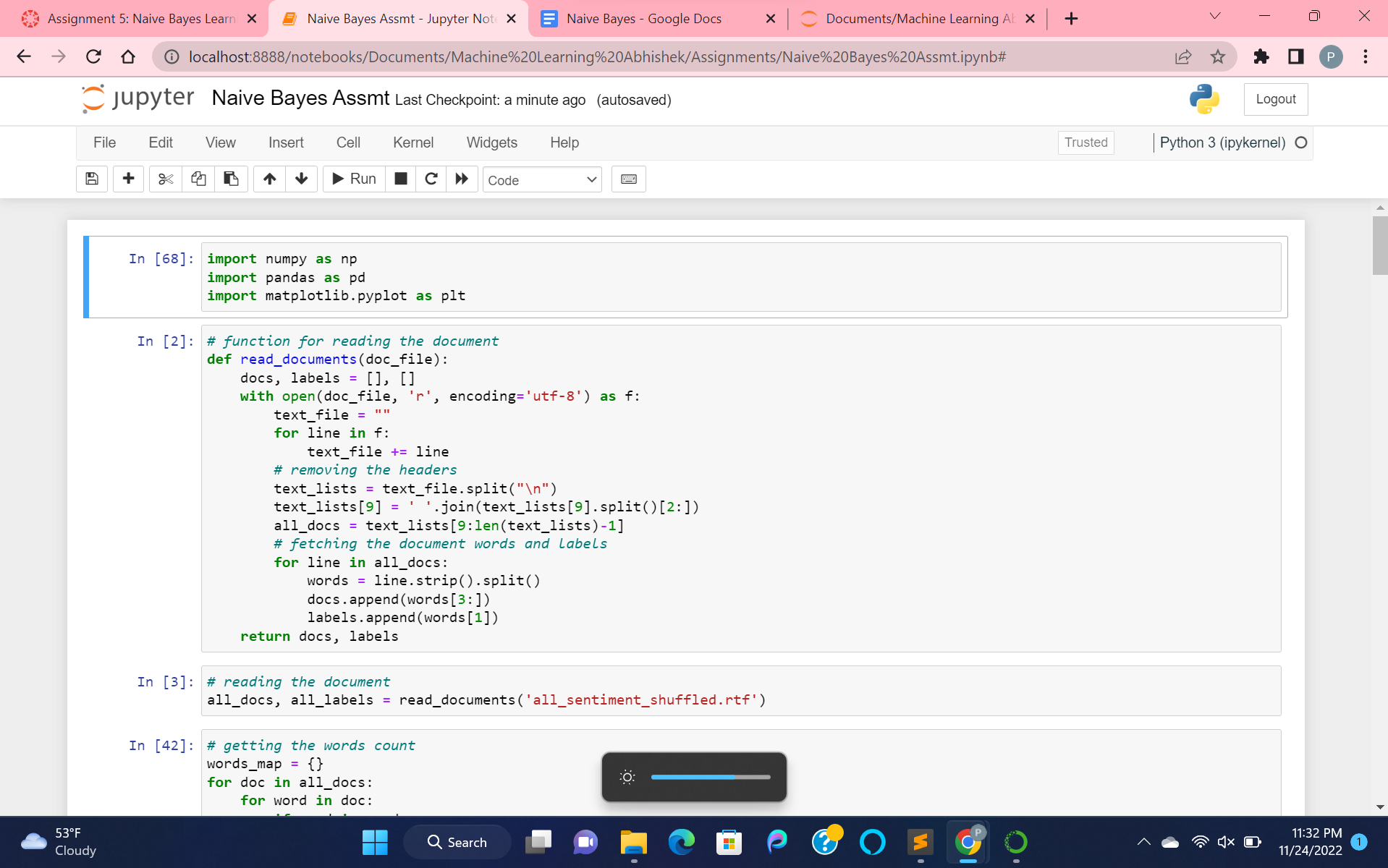
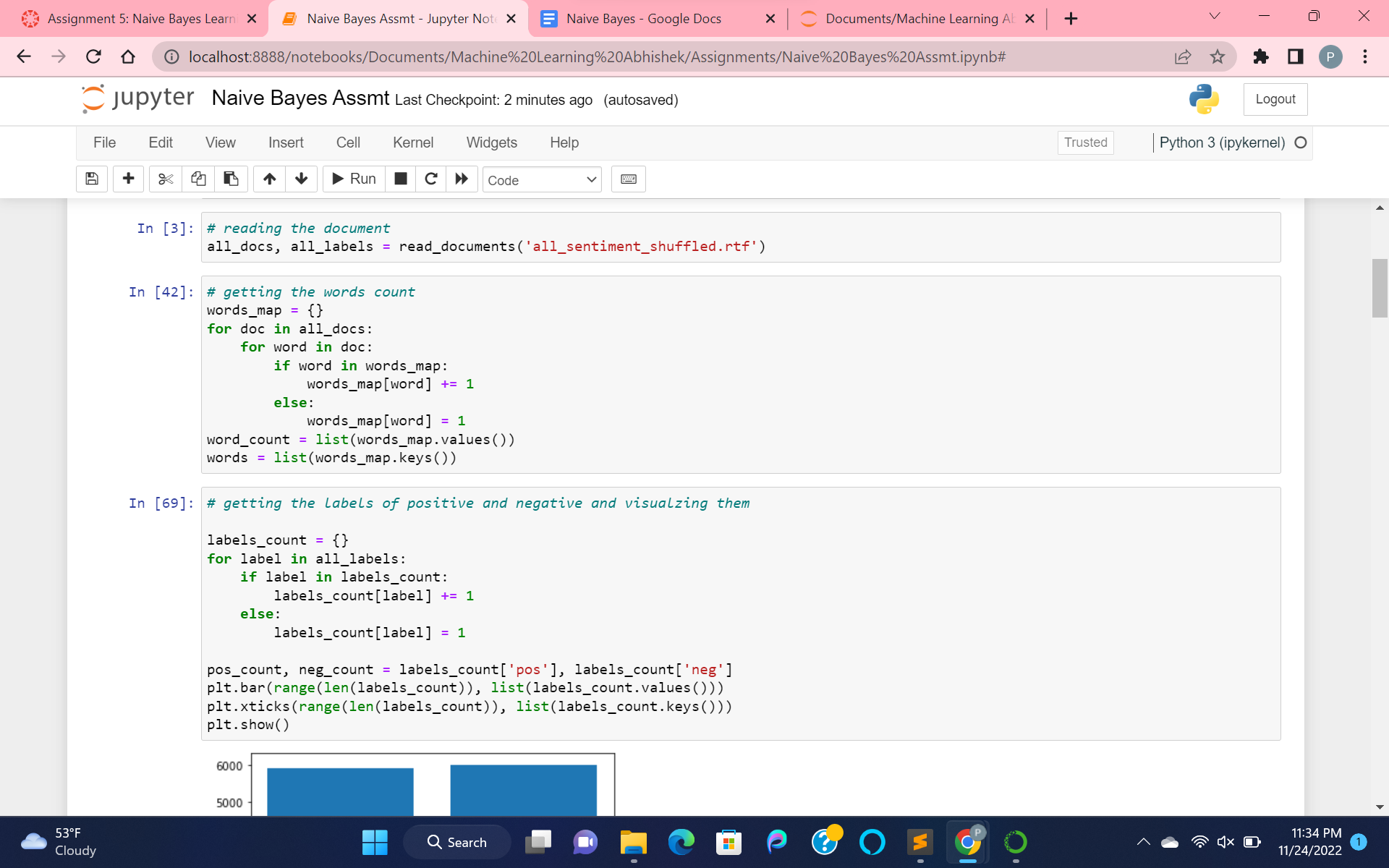
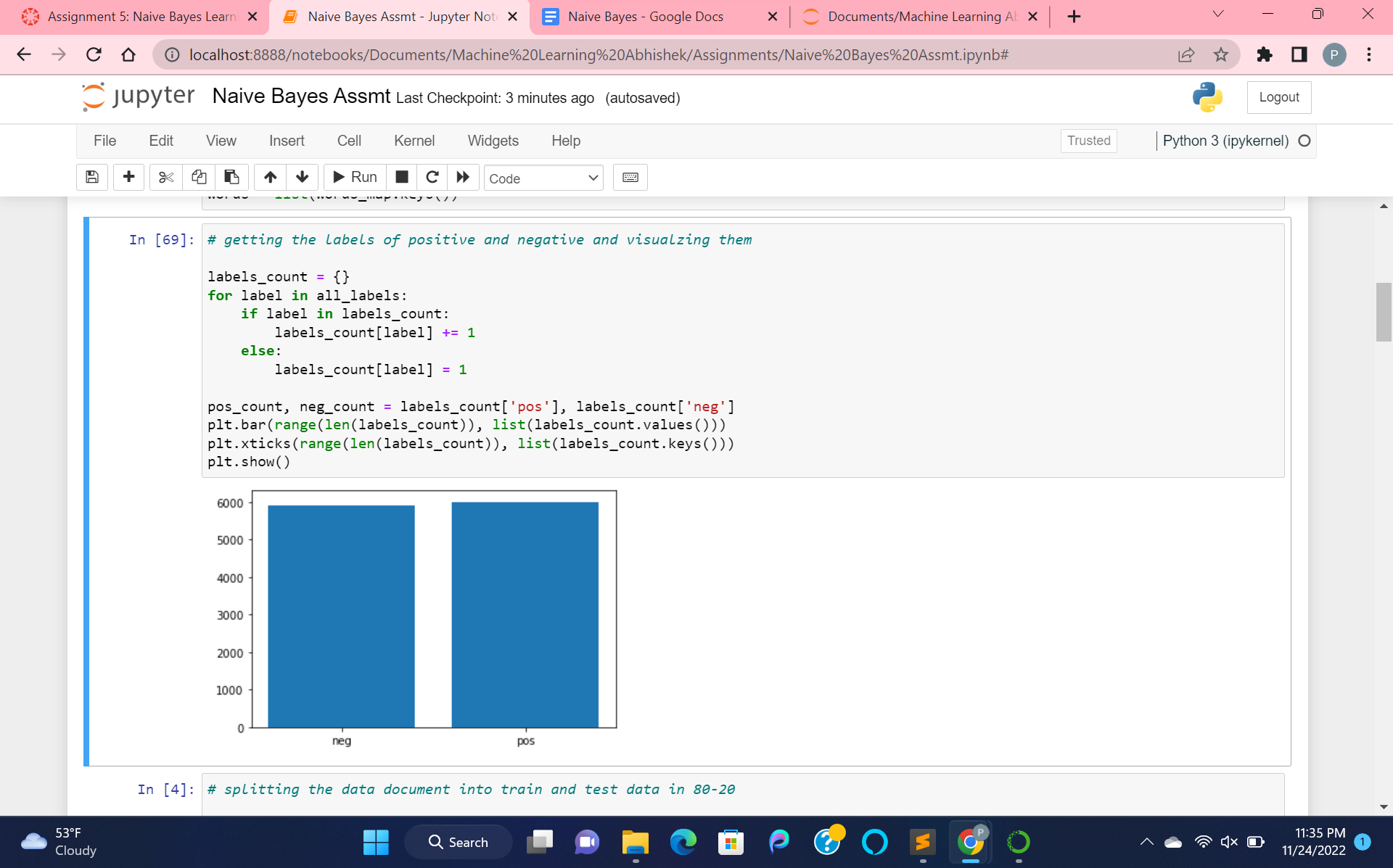
**NAIVE BAYES LEARNING**



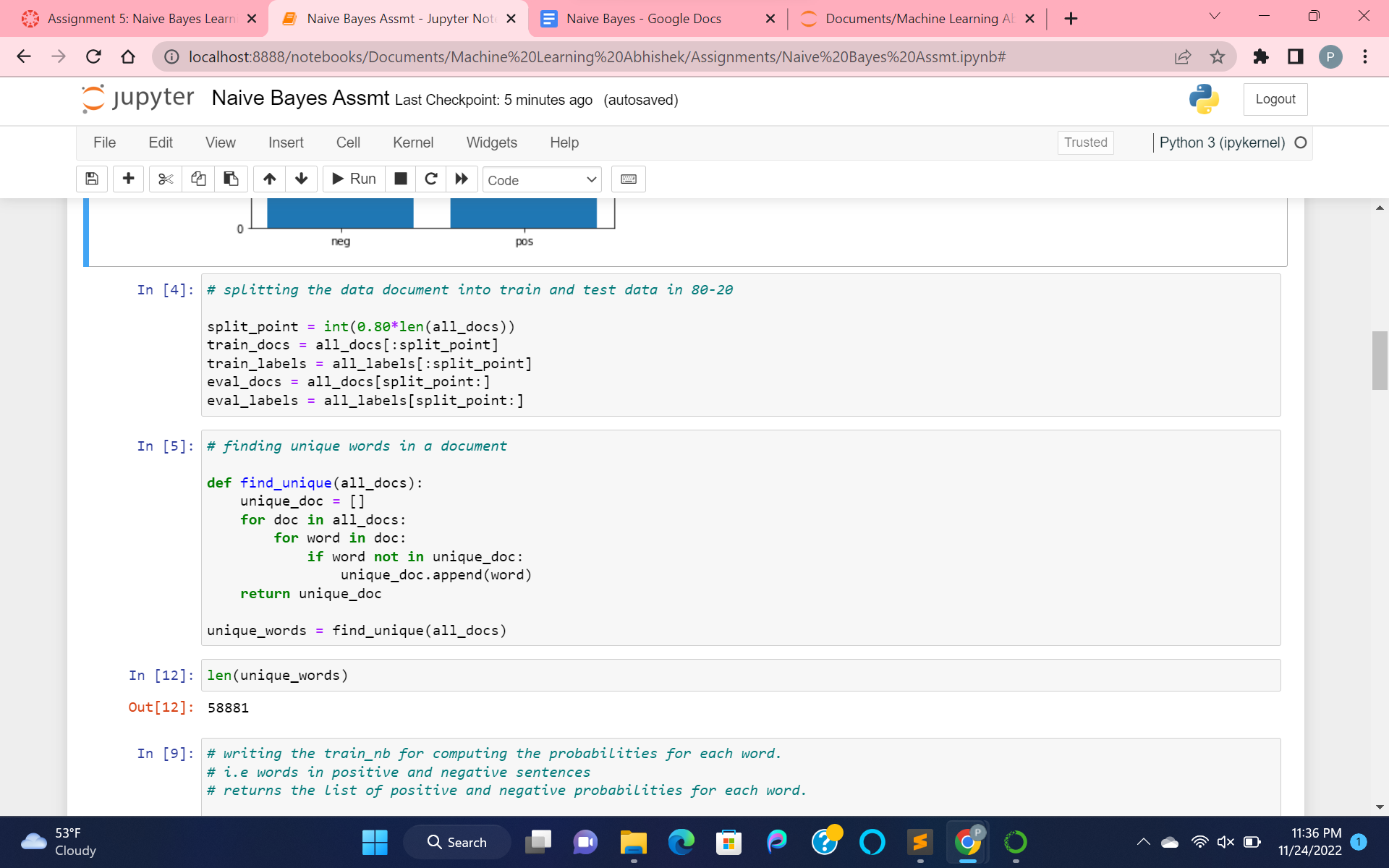
Importing the data by converting the rtf data to text data. By opening the file in read mode and encoding as utf-8, removing unnecessary headers. Then fetching the labels and words in the document.



Getting the words count and labels count. Then visualizing the count of labels.



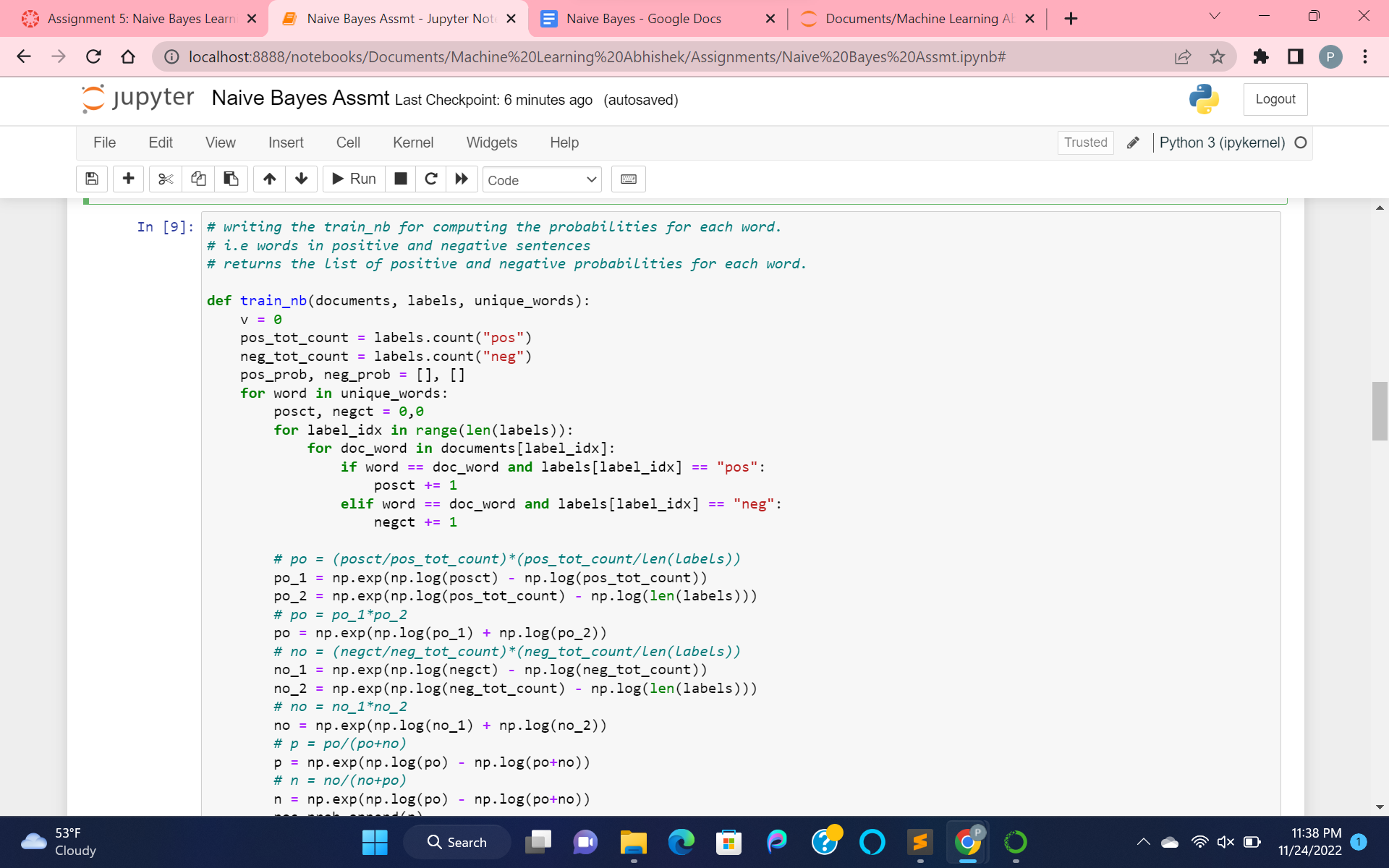
Visualizing the labels of data documents.



Splitting the data into training as train\_docs, train\_labels and testing as eval\_docs, eval\_labels.

Then finding the unique words in find\_unique function for further execution.

The number of unique words is 58881.

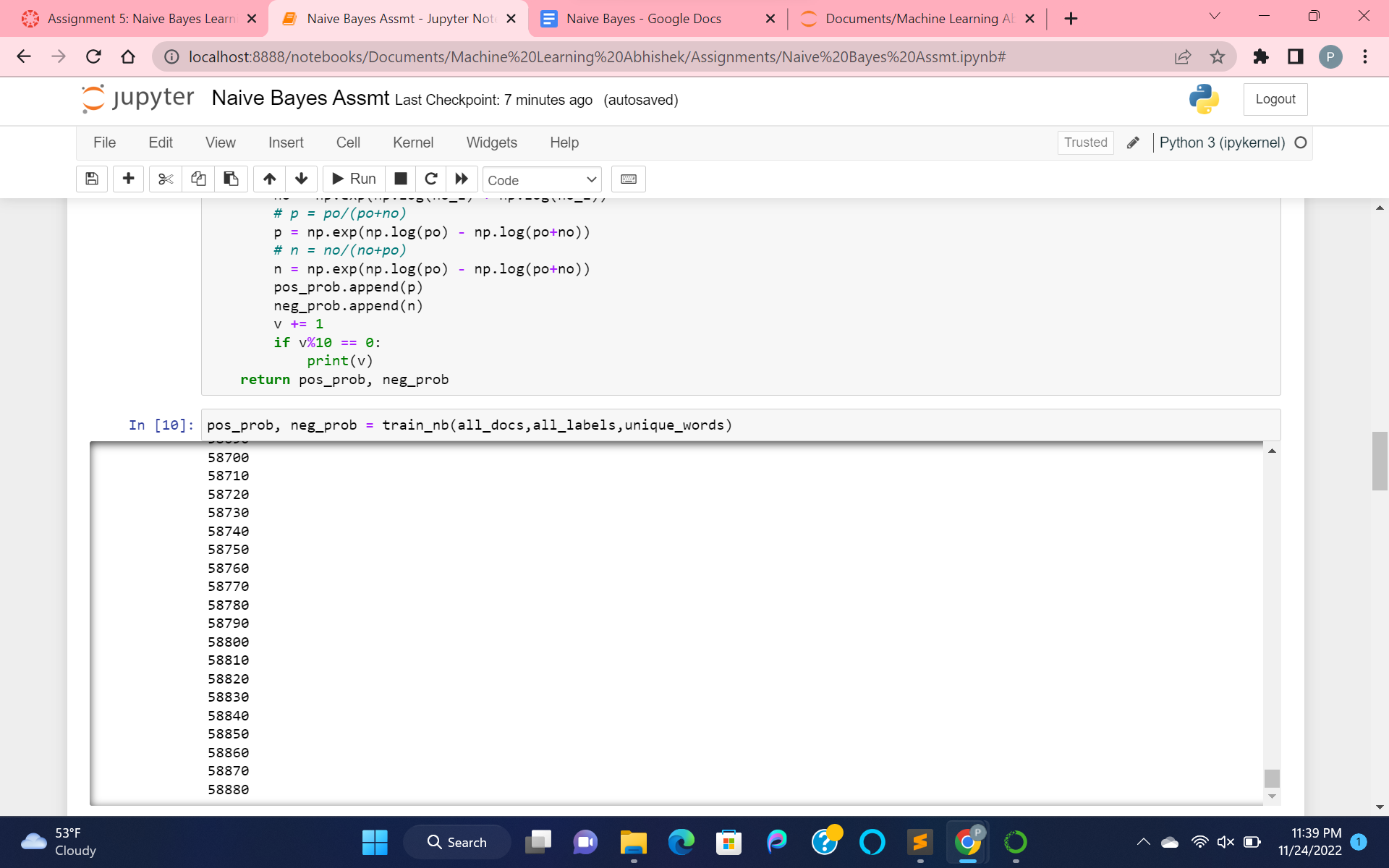


Written a function to train naive bayes for computing the probabilities for each word.

I.e words in positive and negative

Returns the list of positive and negative probabilities for each word.

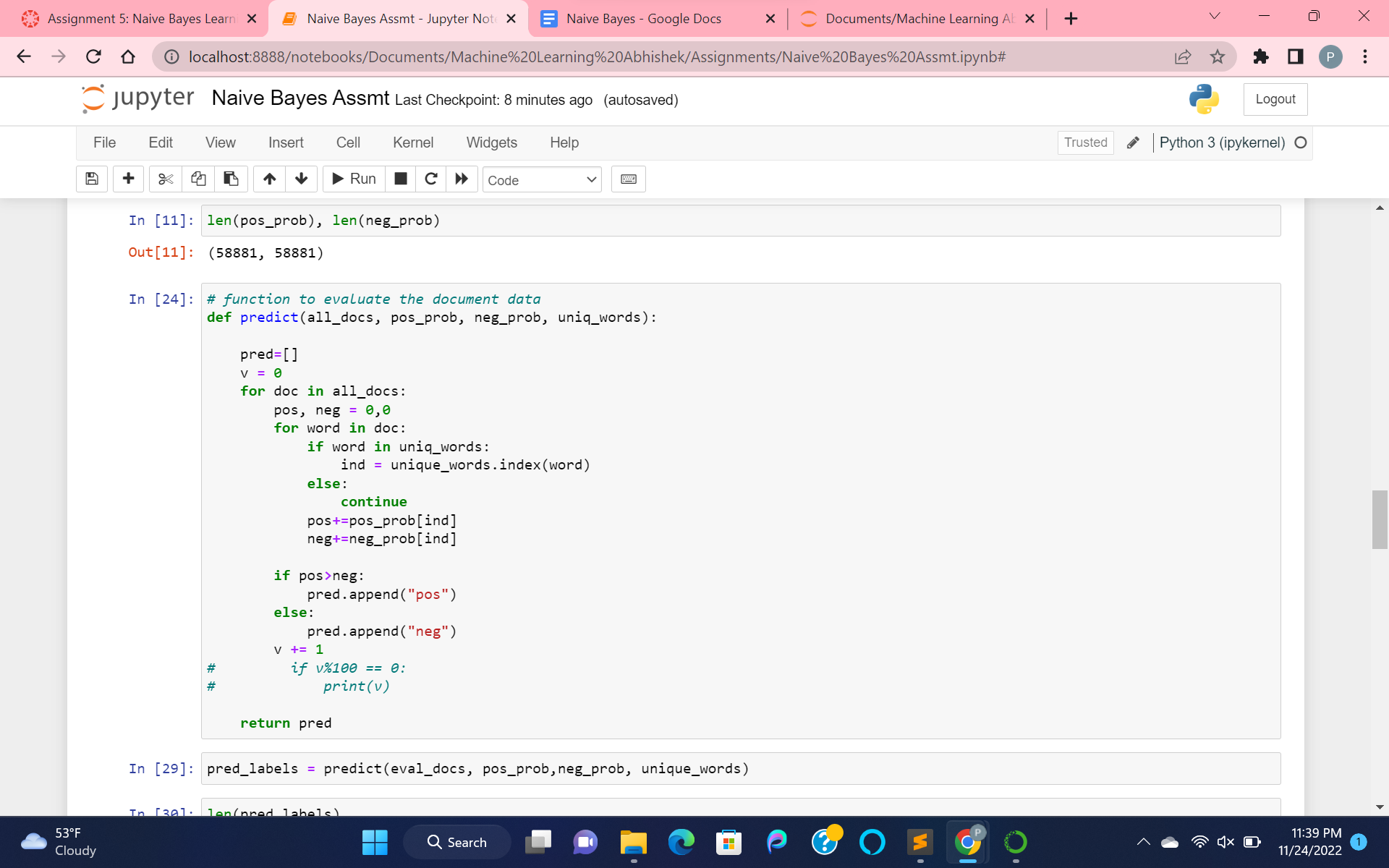
Computing the data docs by counting the positive and negative sentences and then processing them according to the formula and using np.log, np.exp for value calculations.



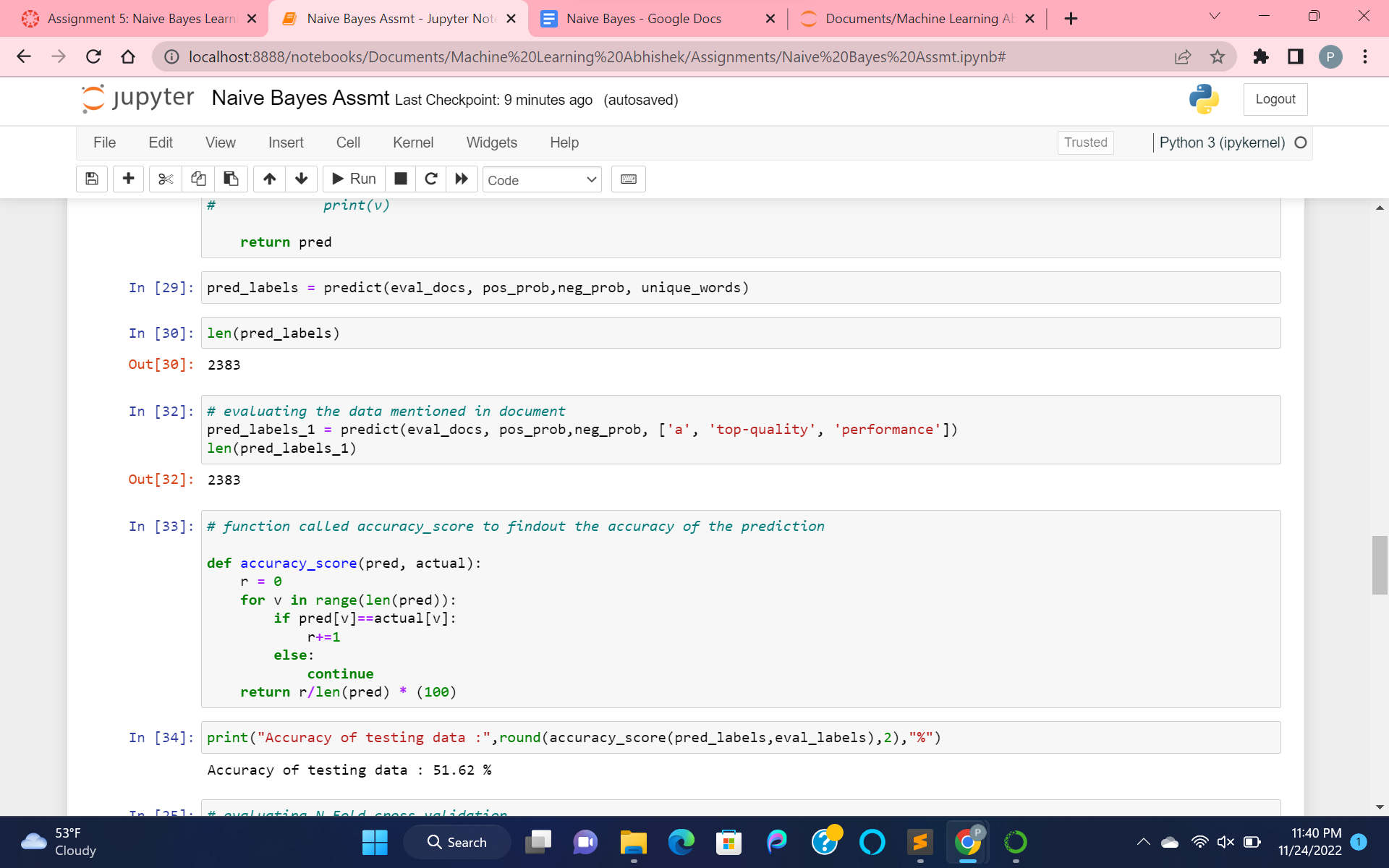
After training the data, we get the pos\_prob and neg\_prob with 58,881 records.

Then predict the model, by using a list of pos\_prob and neg\_prob with help of unique words.

By doing the sum of the pos\_prob and neg\_prob for a word, then comparing positive and negative values.

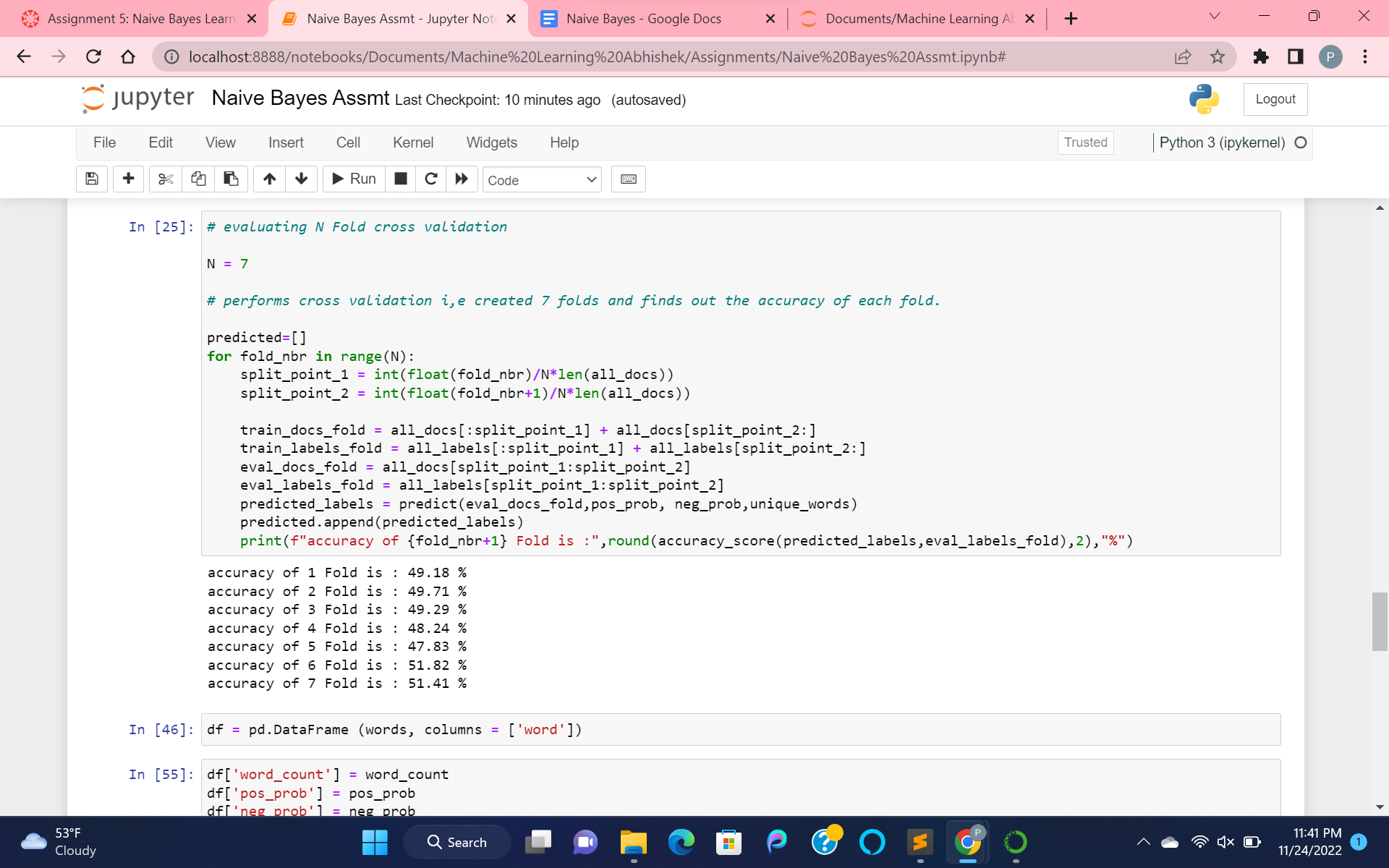


Then predicting the label data using the eval\_docs, list of pos\_prob and neg\_prob, unique words.



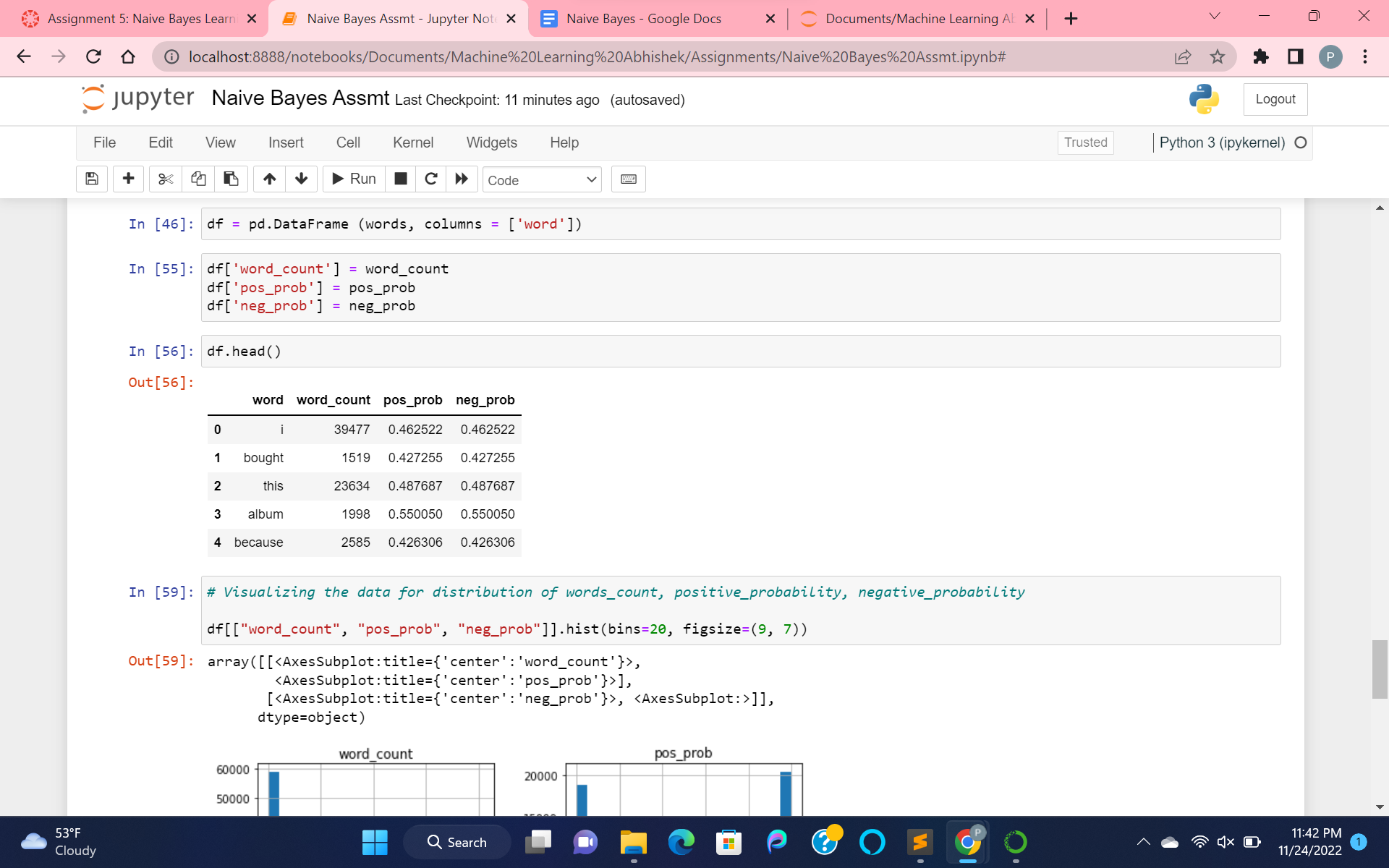
Then calculating the accuracy of testing data using predicted\_labels and eval\_labels.

We get 51.62% accuracy.

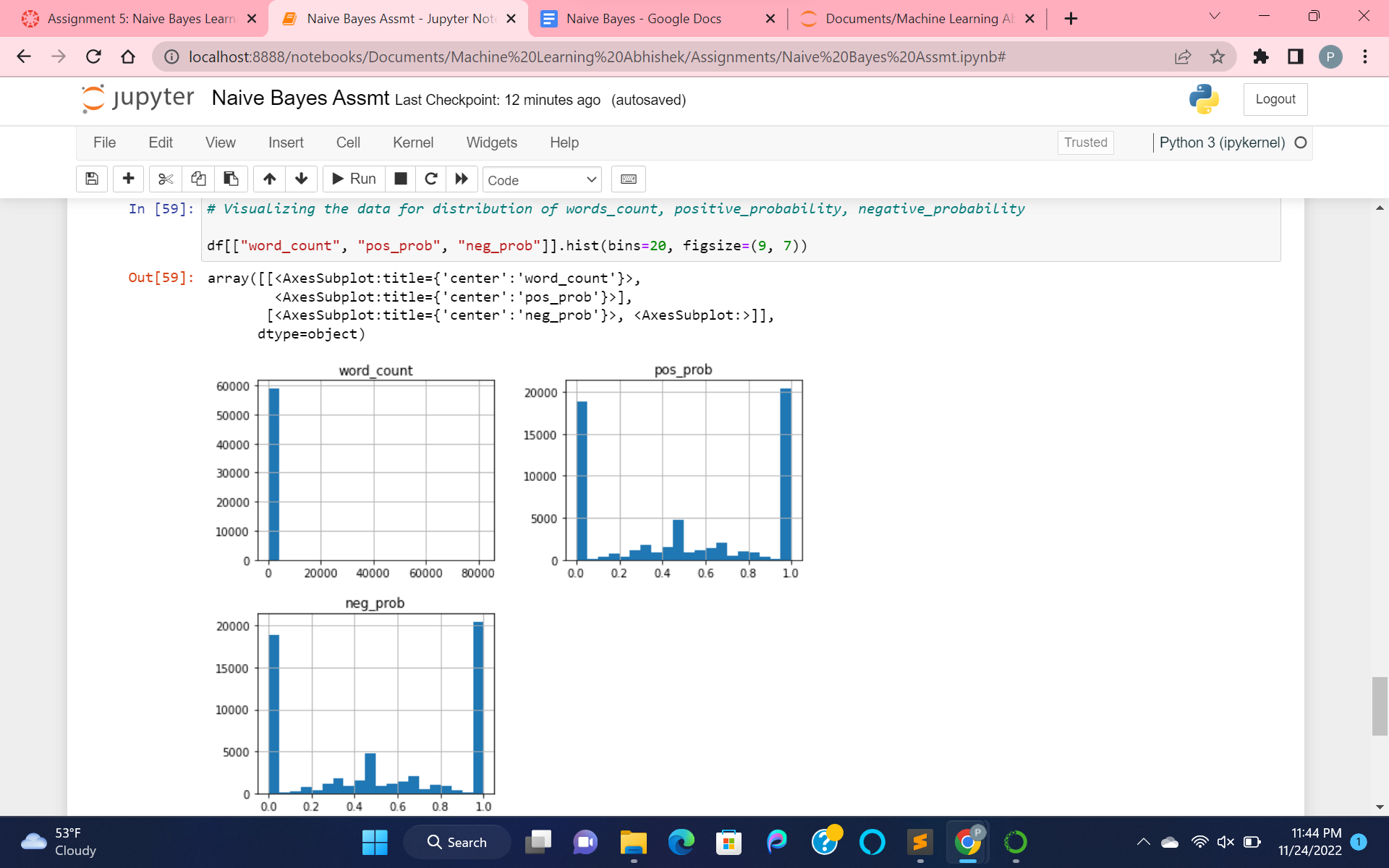


Now, performing the N Fold cross validation by taking the value of N as 7. Here by performing the cross validation for each fold and then finding out the accuracy.

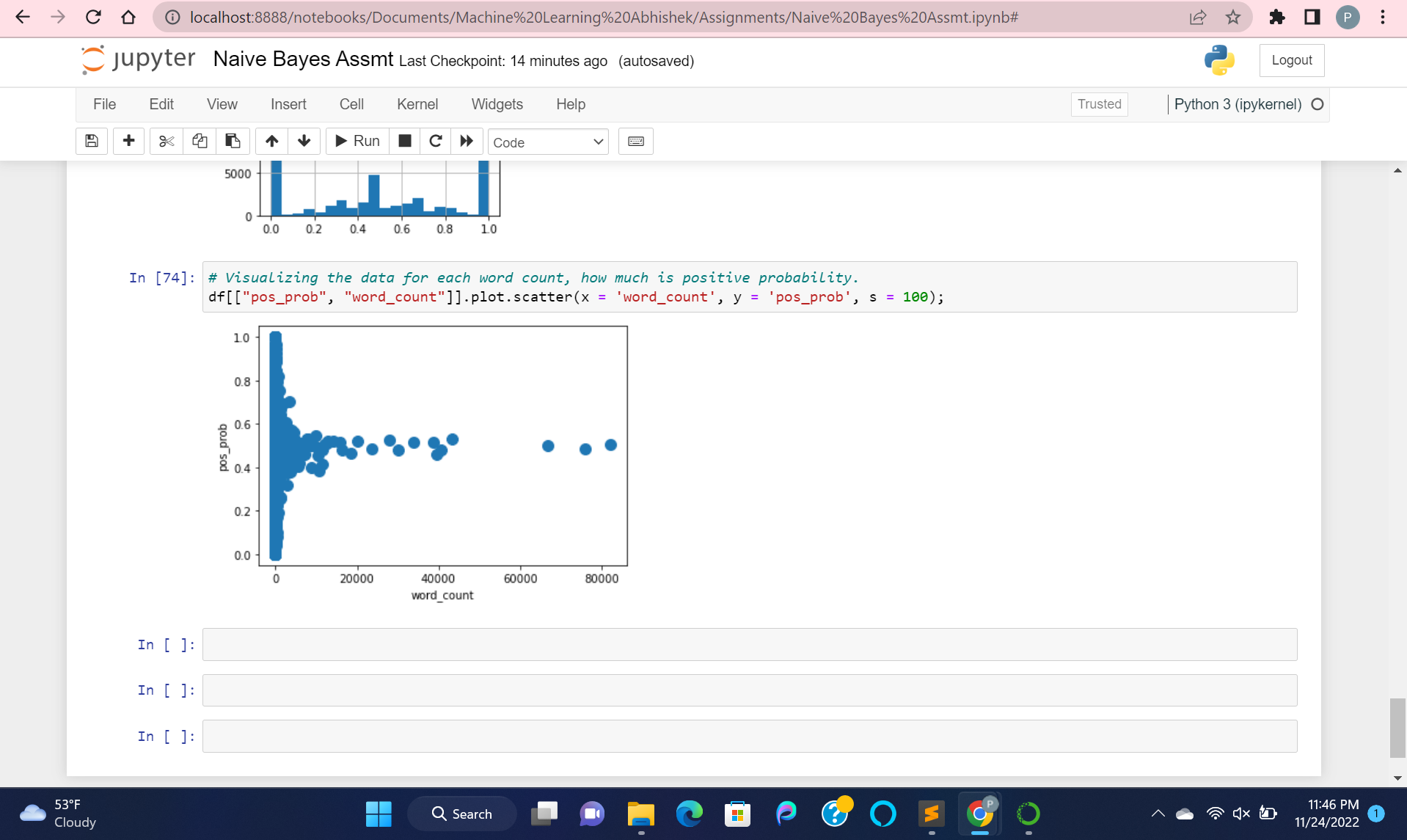
The accuracies for 7 folds that we got are all of them are around 50%.



Creating the data frame, for visualizing the data documents by considering the word count, positive probability, negative probabilities.



Visualizing the hist plot for word\_count, pos\_prob, neg\_prob.



Visualizing the scatter plot for pos\_prob, word\_count, where distribution is done in such a way as, for each word count, calculated positive probability.