News Classification

Use Jupyter Notebooks to create a News Classifier with Multimodal Naive Bayes Classification.

Read in the Data: Use the Kaggle News Category Classification data set for this project. Note this is a JSON dataset and will be read in differently than a CSV.

Test Data: Before you start building your classifier remove ~~the first~~ 20% of the data and place it in a separate table, this will be your testing dataset. Once you have created your model (Steps 0 and 1) with the other 80% of the data you will test to see how accurate your model is by having the program categorize the records in your testing data set.

Stop Words: Just like in your first Project you should remove the Stop Words in all of the articles before building your model or testing.

Classifier: You may use whatever python libraries you wish but you should write the code for each of the four steps in the Multimodal Naive Bayes Classification yourself.

Step 0: Laplace Smoothing (remember this is for word counts not the categories' probabilities)

Step 1: Find probabilities for each word for each category

Step 2: Calculate the probability that a record in the testing data set is part of each category.

Step 3: Compare probabilities calculated in Step 2. Choose the largest probability to assign the category tag for the data.

Results: Since you have the categories that the test data was originally sorted into you can compare the predicted probabilities with the original classifications of the news articles. Report the overall effectiveness of your classifier as a percentage of news items categorized correctly. This should be done as a percentage across all the data in the test data set as well as the percent in each category that were categorized correctly.

Documentation: Make sure to comment your code and give credit to any sources you used in creating your code.

Reflection: Create a document with the answers to the following questions. This should be formatted with each numbered question as the start of its own section, with the answer below it. Make sure your document has a title and your name on it.

Technical:

1. How is JSON formatted? Give an example with an explanation.
2. Why were the stop words removed from the text? If you don't do this step in your code what changes? What other steps could you take along the same lines to improve your program?
3. Why was Laplace Smoothing done?
4. This is a machine learning algorithm. What is the purpose of training and testing data sets? Why might 20% of the data have been reserved for testing, not more, not less?
5. How did the size of data affected the time complexity? How did you manage this challenge?

Process:

1. How did you approach the assignment? Did you give yourself enough time?
2. What challenges did you have with the code?
3. What did you learn while working on this assignment?