Github: https://github.com/sachinbiradar9/News-Classification

What is your data and task?
Data: https://data.world/elenadata/vox-articles
Task: Classify new news into various categories based on their headline.

What ML solution did you choose and, most importantly, why was this an appropriate choice?
I tried various classifiers- Decision Tree, Support Vector Classifier, Multinomial Naive Bayesian Classifier, Multilayered Perceptron, Random Forest. Multinomial Naive Bayesian Classifier worked the best. It is logical for Multinomial Naive Bayesian to work the best as even we as humans classify based on keywords. We are likely to predict "Politics" is we see keywords like Obama, election, republic and we are likely to predict "Criminal" if we see keywords like drugs, jail and so on. Naive bayesian scans whole dataset and finds the probabilities of each word in headline being associated with a class and then find the probability for whole headline hence it works good.
Steps - Load data > Split train/dev/test > Remove stop words from title > vectorize title using bag of words and convert category to numbers > feature reduction using variance > oversampling of data to make distribution uniform > train various classifier on training data > use dev data to check f1-score > choose naive bayesian model to predict the test data.

How did you choose to evaluate success?
I used F1-score to evaluate the models.

What software did you use and why did you choose it?
Language: Python as all the libraries are prewritten in python
Software: *scikit-learn* has all the machine learning algorithms prewritten, *numpy* has all math operation and list operation prewritten, *nltk* has all natural language tasks such as word tokenizer prewritten.

what are the results?

```
Predicting test data using Multinomial Naive Bayesian
                    precision    recall  f1-score   support

Business & Finance       0.52      0.47      0.50       108
  Criminal Justice       0.52      0.60      0.56        97
       Health Care       0.50      0.62      0.56        90
 Politics & Policy       0.75      0.68      0.71       333
  Science & Health       0.60      0.63      0.62       164
       avg / total       0.63      0.62      0.63       792
```

Show some examples from the development data that your approach got correct and some it got wrong: if you were to try to fix the ones it got wrong, what would you do?
Correct:

```
Title: 9 charts that explain the history of global wealth
True Category: Business & Finance, Predicted Category: Business & Finance
Title: remember when legal marijuana was going to send crime skyrocketing?
True Category: Criminal Justice, Predicted Category: Criminal Justice
```

Incorrect:

```
Title: bitcoin is down 60 percent this year here's why i'm still optimistic.
True Category:  Business & Finance, Predicted Category: Criminal Justice
Tile: maryland set to decriminalize marijuana
True Category: Criminal Justice, Predicted Category: Health Care
```

Some were classified incorrectly because they had only one keyword to associate with some category among many other tokens hence the product of other token overweighted the keyword. To improve the classification it would be very helpful to improve data preprocessing step where all such keywords which are *almost* uniformly distributed over all classes are removed hence they wont affect in classification.

Credits: *scikit-learn, numpy, nltk*