Python Lab - Education Or Cinema ?

In this segment, you will learn how to implement both Multinomial and Bernoulli Naive Bayes classifiers in python.

Please find the dataset of test data here, train data here and the code file here.

Now in the next lecture, we will be proceeding with the pre-processing steps required to fit a Naive Bayes model on it.

Stop Words:

We can see a few trivial words such as 'and','is','of', etc. These words don't make any difference in classifying a document. These are called stop words. So we would like to get rid of them. We can remove them by passing a parameter stop_words='english' while instantiating Countvectorizer().

Refer to the image below.

```
array([[0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1],
       [0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0],
       [0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0],
       [1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0],
       [0, 1, 0, 0, 2, 0, 0, 0, 1, 0, 1, 0]])
```

Sparse Matrix

Now the way to get rid of these is know as a **Compressed Sparse Row format.**

```
(0, 2)        1
(0, 5)        1
(0, 7)        1
(0, 11)       1
(1, 1)        1
(1, 2)        1
(1, 3)        1
(1, 6)        1
(2, 2)        1
(2, 3)        1
(2, 5)        1
(2, 6)        1
(2, 10)       1
(3, 0)        1
(3, 5)        1
(3, 9)        1
(4, 1)        1
(4, 4)        2
(4, 8)        1
(4, 10)       1
```

Compressed Sparse Matrix

This representation can be understood as follows:

Consider first 4 rows of the output: (0,2), (0,5), (0,7) and (0,11). It says that the first document (index 0) has 7th , 2nd , 5th and 11th 'word' present in the document, and that they appear only once in the document- indicated by the right hand column entry.

In the next lecture, we will finally build a Naive Bayes Classifier using the pre-processed data.

Note : At 1:19, Instructor mistakely told X_test is Compressed Sparse matric instead of X_test is not compressed sparse matric.

In the next segment, you will implement both Multinomial and Bernoulli Naive Bayes classifiers on a real dataset to classify SMSes as spam or ham.

Multinomial