Q.2

CODE : NLP/NewsGroupCategorizationNLP.py

EXPLANATION FOR DATA PREPROCESSING STEPS:

1. **Removal of stop words:**

Stop words do not contribute to the document classification as they are words that are generally used to complete the sentence without imparting much information regarding the document. They add unnecessary noise taking up space in our database and valuable processing time.

2. **Removal of numbers and special characters:**

Similar to stop words, they are not relevant for document classification and add unnecessary noise taking up space in our database and valuable processing time.

3. **Conversion to lowercase :**

Converting to lowercase would shrink the size of vocabulary and reduction in processing time. In most of the cases, the words should be considered similar irrespective of the case, eg (Apple and apple should not be considered differently)

4. **Stemming of words:**

Stemming of words is an important step in document classification as it reduces the vocabulary and focuses on the sense or sentiment of a document rather than deeper meaning.

FEATURE EXTRACTION:

1. **Feature (CountVectorizer) -** Convert a collection of text documents to a matrix of token counts. Builds a matrix with count of each word in a document.
2. **Feature : Tf-idf scores**
   Some words should have low scores despite having a high count as they do not impart any information being present in all the documents. For taking this into consideration, tf-idf scores are used as feature.

ACCURACY :

A. Feature : CountVectorizer

After preprocessing the data, splitting the data into training and test sets (0.3) and feature extraction, accuracy and confusion matrix are as below

Model : Gaussian Naive Bayes

Accuracy: 0.88

Confusion matrix:

```
[[257  22   5  24]
 [ 27 256   2   9]
 [  6  10 272  14]
 [  4  10  16 260]]
```

['rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey']

As per the confusion matrix, motorcycles are being confused with hockey. In order to improve accuracy, we can take look at the words that be causing the confusion.

Model : Logistic Regression

Accuracy: 0.85

Confusion Matrix

```
[[242  30  32   4]
 [ 28 243  20   3]
 [  7  10 275  10]
 [ 11  11  19 249]]
```

['rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey']

As per the confusion matrix, motorcycles are being confused with autos and baseball with hockey.

B. Feature : Tf-idf scores

After preprocessing the data, splitting the data into training and test sets (0.3) and feature extraction, accuracy and confusion matrix are as below

Model : Gaussian Naive Bayes

Accuracy: 0.87

Confusion matrix:

```
[[255  24   5  24]
 [ 29 254   2   9]
 [  7  11 268  16]
 [  4  11  13 262]]
```

['rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey']

As per the confusion matrix, motorcycles are being confused with autos.

Model : Logistic Regression

Accuracy: 0.87

Confusion Matrix

```
[[251  50   5   2]
 [ 26 263   4   1]
 [  5  21 264  12]
 [  4  19   8 259]]
```

['rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey']

As per the confusion matrix, motorcycles are being confused with autos.

OUTPUT OF SOME STEPS :

 # PREPROCESSING

- print(news_group.target_names)
  ['rec.autos', 'rec.motorcycles', 'rec.sport.baseball',
  'rec.sport.hockey']

- print(news_group.target.shape)
  (3979,)
- print(news_group.data[0])
  IMO any good player should score on power plays because of the man
  advantage. Very good power play scorers tend to become overrated
  because their point totals are inflated by power play points.
  +/- tends to expose these overrated players such as Brett Hull,
  John Cullen and Dave Andreychuck.

  Given the opportunity to play power play consistently, any player can
  inflate his totals.

 # AFTER PREPROCESSING

- print(news_group.data[0])
  imo good player score power play man advantag good power play scorer
  tend becom overr point total inflat power play point tend expos
  overr player brett hull john cullen dave andreychuck given opportun
  play power play consist player inflat total