□ Implementation:

To implement this program, I used Java programming language. I used HashMap data structure for saving the words along with their frequencies for each class. For word parsing, I used Java's StringTokenizer, and I also ignored all of the digits and delimiters except the apostrophe (') while tokenizing. Then, I removed 1000 most frequent words from the vocabulary.

To avoid the underflow error, I calculated the logarithm of prior probability $P(c_i)$ and logarithms of each of the likelihoods $P(x_j|c_i)$. Then, finally I calculated the summation of all terms.

□ Dataset:

In this program, we have a dataset containing 19997 newsgroup messages (997 from soc.religion.christian class and (1000*19)=19000 from 19 other classes) drawn from the 20 newsgroups.

Training Dataset: For training, I used 9999 newsgroup messages (499 from soc.religion.christian class and (500*19)=9500 from 19 other classes) from the Dataset.

Testing Dataset: For testing, I used 9998 newsgroup messages (498 from soc.religion.christian class and (500*19)=9500 from 19 other classes) from the Dataset.

□ Result:

Among the **9998** newsgroup messages in testing dataset, my program accurately classified **7101** newsgroups messages.

Therefore, the classification accuracy is:

$$\frac{7101}{9998} * 100\% = \mathbf{71.02}\%$$

□ How to run the program:

In this program, there are 5 Java source files. Among those, NaiveBayesClassifier.java contains the main() method. Below is the procedure to run this program:

1. Let's assume, you have unzipped the folder to "C:/Downloads". Then, open Windows command prompt and navigate to:
   "C:/Downloads/NaiveBayesClassifier_1001382987/Code"

2. Type "**javac *.java**" and press enter. This will compile each java source files to java class files.

3. Finally, type "**java NaiveBayesClassifier**" and press enter to run the program. Please note that, you'll need the JDK (Java Development Kit) installed in your machine. Also, you'll need to setup your environment variables for running Javac and Java commands as stated here.