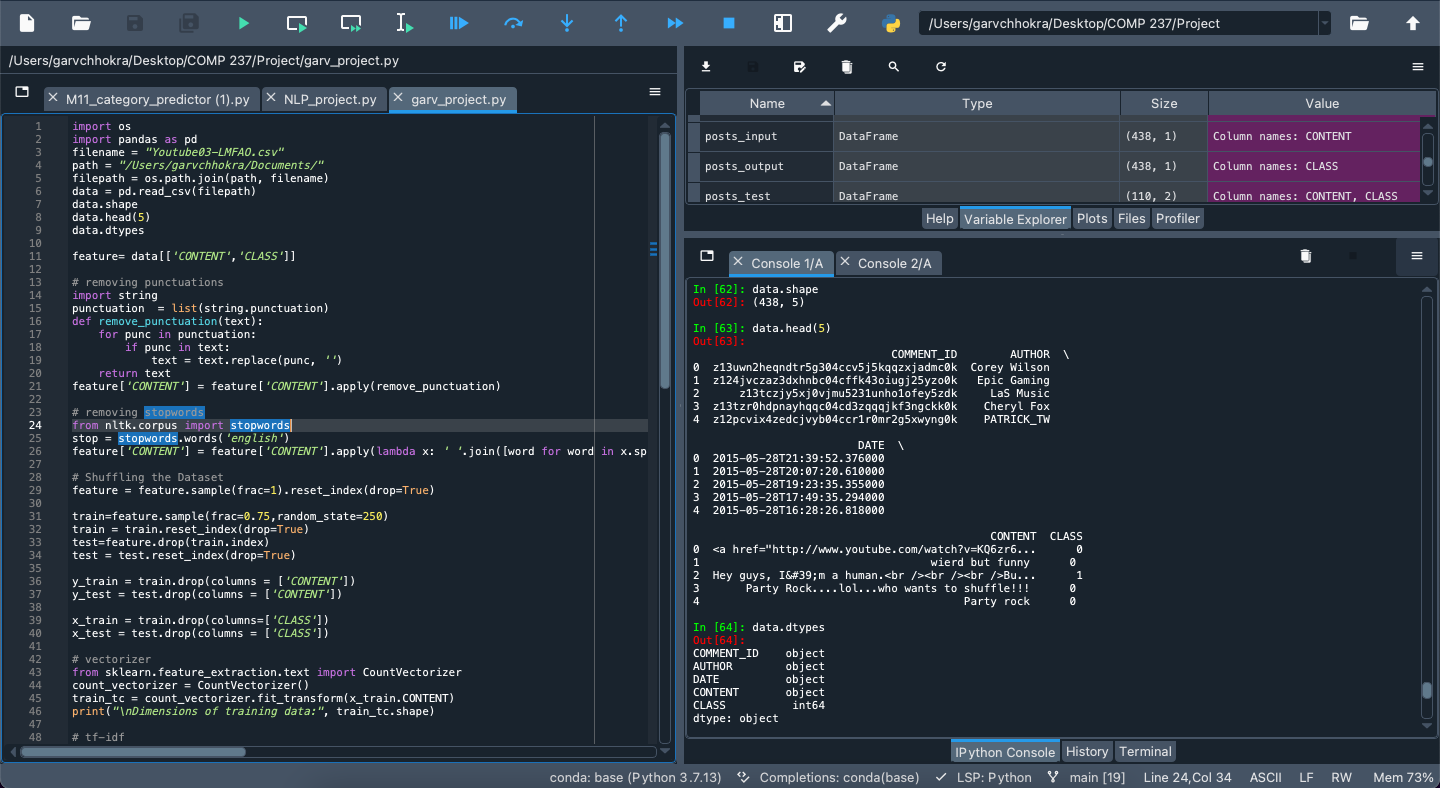
**Project Assignment - NLP**

**Data Exploration:**

Used shape, head, and Data types for the exploration of our data.



Text

Description automatically generated

**Selecting two columns for the model:**

We have selected two columns that are CONTENT and CLASS. We have removed Comment Id because it is not giving any pattern and all the unique so it won’t be beneficial if we put it in our model for training and testing. Apart from that, we have also removed the Author column because it wasn’t giving any specific pattern to identify whether that comment is spam. Hence, we only took CONTENT and CLASS columns in our model.

**Preparing the data for the model:**

We have removed stop words and punctuation from our Content data (Comments). So, that we can make our model more accurate.

**Creating Train and Test data for the model:**

Applying 75% for training and 25% for testing. It is obtained that the x test and train are having shapes of (110,2) and (328,2) respectively.

Graphical user interface, text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

**Dimension of training data after using CountVectorizer:**

After fitting the training data into the count vectorizer. The shape of training data came out to be: (328,775)

Graphical user interface, text

Description automatically generated



**Downscale the transformed data using tf-idf:**

After fitting data into TFidf Transformer the shape came out to be (328, 775).

Graphical user interface, text

Description automatically generated

Text

Description automatically generated