Naive Bayes Classifier Documentation

The JeopardyTopicNaiveBayesClassification module constructs a Naive Bayes model to classify the Jeopardy questions by one of ten categories generated by LDA. In the web app, the script runs when the user has submitted a Jeopardy-style question of their choice. Before running the web app, make sure that the following Python packages are installed: nltk, scikit-learn, pandas, numpy, and matplotlib by running **pip install -r requirements.txt**.

The CSV containing the labeled topics for each question is preprocessed using pandas. Two dataframes are created in the process: one containing the entire CSV data and the second containing everything but the topics. The data is then split into training and test datasets using the train_test_split method from the sklearn.model_selection package, where X is the list of preprocessed Jeopardy questions and y is the topic prediction.

The Jeopardy questions from each dataset are then vectorized using the top 2000 tokens in their entire corpus where each token was a unigram or a bigram. The result was a sparse question term matrix: one for the training data and another for the test data. The training question-term matrix was then fed into a multinomial Naive Bayes classification model with the hyperparameter alpha set to 10. The test question-term matrix was used to predict the topics, which were then compared with the actual topics from the test dataset to assess the model's accuracy, which turned out to be around 77%.