LDA Overview

For the topic modeling portion of our project we used the LDA algorithm. Specifically, we used the scikit-learn implementation of LDA. In our project workflow, the topic modeling came after the data cleaning process and before the Naive Bayes classifier training. To prepare the data for LDA topic modeling, we created a list containing all the questions in the dataset. Due to the preceding data cleaning step, these questions were clean, free of stop words, and lemmatized. The questions were then vectorized using the top 2000 tokens in the entire corpus of documents. Each token was either a unigram or a bigram. The result of vectorization was a sparse question term matrix that was then fed into the LDA algorithm which was set up to discover ten topics. After topic discovery, each question was then labelled with the discovered topic of highest likelihood. Because the LDA algorithm gave us the top words in each topic, we used ChatGPT to summarize these words into more human-comprehensible topics. That is, we told ChatGPT to summarize the top 100 tokens from each topic into a few concise terms. These terms were then used to label each question in the dataset. This labeled data was then exported and used to train the Naive Bayes Classifier.