**Naïve Bayes**

After reading the datasets for training which were the questions.txt and topics.txt, we normalized the questions in the questions.txt file to remove stop words. After we normalised our data, we used the *countVectorizer* to convert it into a vector, then we separated our dataset into training and test data. After Normalization and Vectorization, we used the Multinomial Naïve Bayes classifier in sklearn to train and test the accuracy of our model.

**Logistic Regression**

After reading the datasets for training which were the *questions.txt* and *topics.txt*, we normalized the questions in the *question.txt* file to remove stop words. After Normalization, we use the *tfidVectorizer*to convert our normalised *question.txt* file in a vector which is basically an array of numbers, then we separated our dataset into training and test data by taking 33% of our entire dataset as our test data. After Normalization and Vectorization, we used the Logistic Regression classifier in sklearn to train our model and we use it to test the accuracy of our model.

We chose to use Logistic Regression in the implementation of our because it had a higher accuracy than that of the naïve bayes.

**LATENT DIRICHLET ALLOCATION**

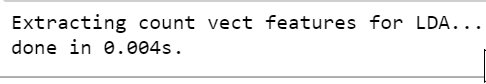
Latent Dirichlet Allocation (LDA) is an example of a probabilistic topic model. The LDA model utilities the Bag-of-words approach in representing text documents.

For topic modelling, the LDA model under Machine Learning is one of the most popular algorithms and therefore has extensive research and documentation on it. Also, for documents vectors that are sparse, have low-dimensional rate, LDA works well to identify topic similarities and extracts them.

For this program, Sckit Learn was also one main tool used in Vectorization

Vectorization

Under Vectorization, we used two different approaches to test speed and accuracy. The Count Vectorizer and the TFIDF vectorizer were chosen to perform tis task and their speed test were recorded accordingly using the same dataset.

Normalization

To ensure standardized dataset, stop words were removed from the dataset

LDA vs NMF

