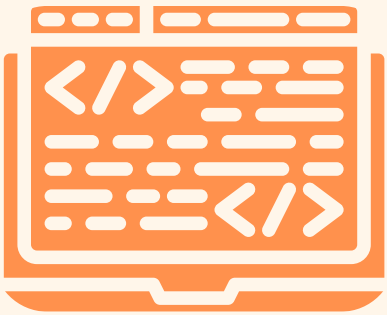
IRIS AND REUTERS

DS TOOLS LAB WORK



Supervised by
Dr. Mohamed
Eng. Sohaila

TEAM NAMES AND IDS

Name	ID
MAHMOUD ESSAM	20221460231
ABDELRAHMAN ASHRAF	20221374041
ABDULLAH HUSSIEN	20221427861
FARES MUHAMMED	20221461330
ZYAD ASHRAF	20221374025

IRIS DATA

This report delves into a comprehensive analysis of the Iris dataset, a well-known dataset in the field of machine learning. The dataset encompasses measurements of sepal length, sepal width, petal length, and petal width for three distinct species of iris flowers: setosa, versicolor, and virginica.

Data Attributes

Sepal Length (cm)

This column represents the length of the iris flower's sepal (the outermost part of the flower).

It is measured in centimeters.

Sepal Width (cm)

This column represents the width of the iris flower's sepal.

It is measured in centimeters.

Petal Length (cm)

This column represents the length of the iris flower's petal (the inner part of the flower).

It is measured in centimeters.

Petal Width (cm)

This column represents the width of the iris flower's petal.

It is measured in centimeters.

Species

This column represents the species of the iris flower.

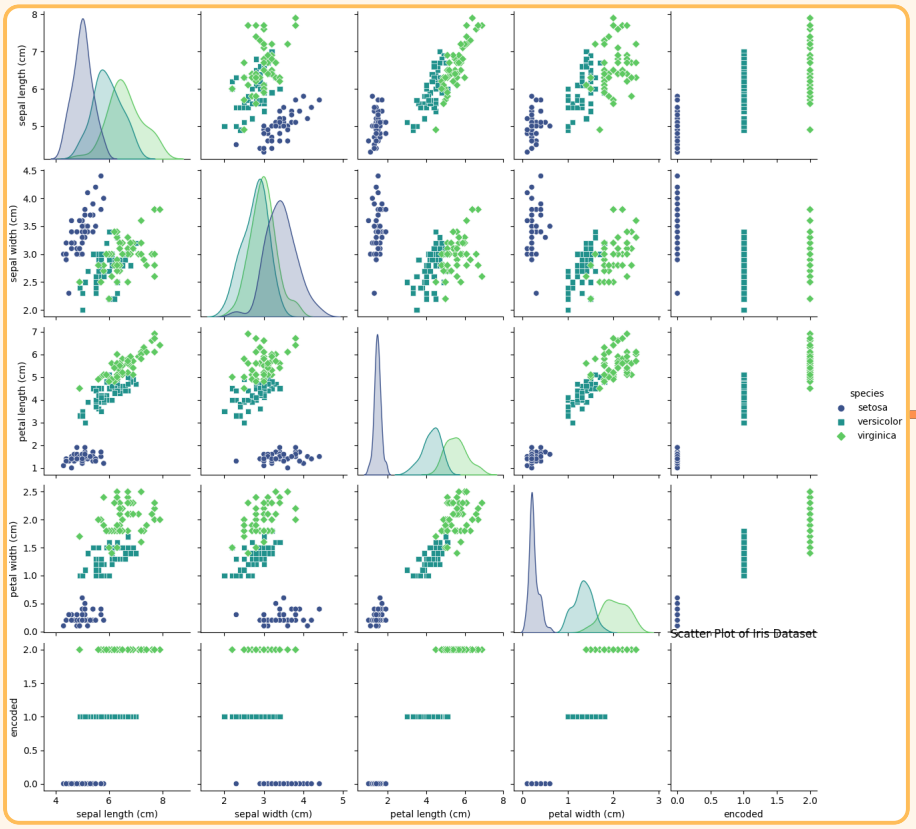
It is a categorical variable with three possible values: 'setosa', 'versicolor', and 'virginica'.

The species is the target variable that we often want to predict in machine learning tasks.

What happened in preprocess?

a. **Duplicated Entries:** We checked for the same data appearing more than once and got rid of it.
b. **No Missing Stuff:** We looked for missing information, and there wasn't any.
c. **Check of outliers** and interval of data

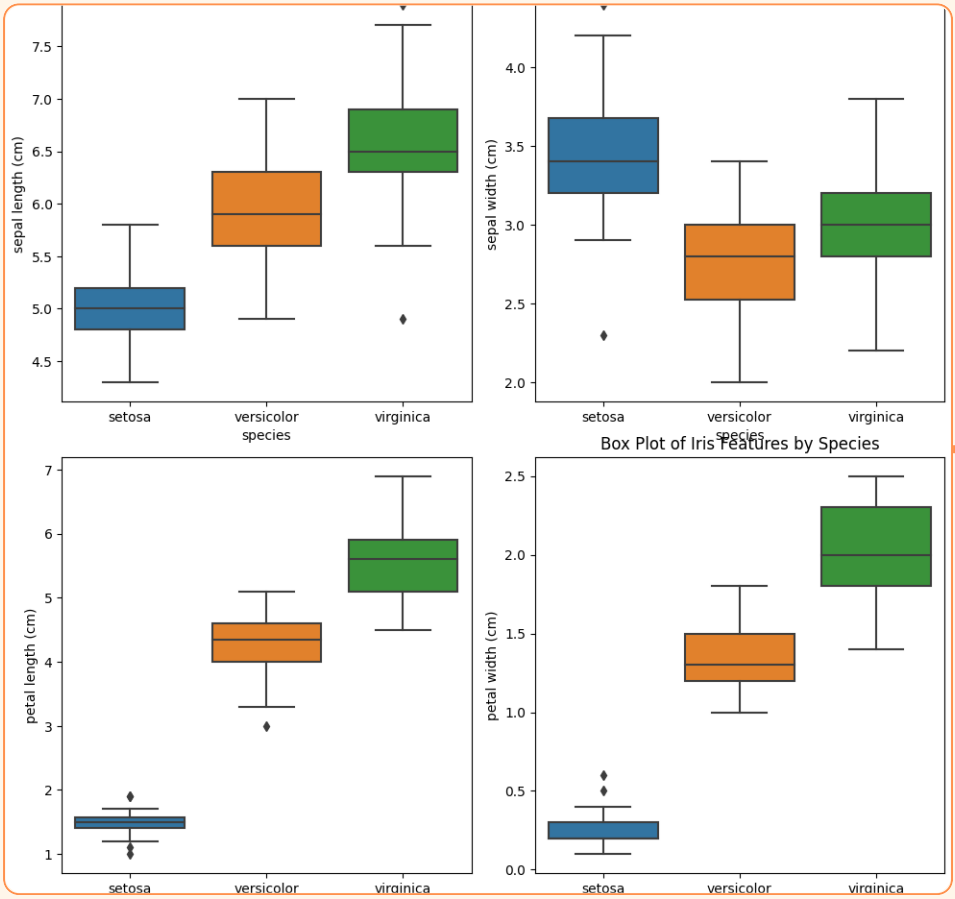
What about visualization?



Scatter Plot

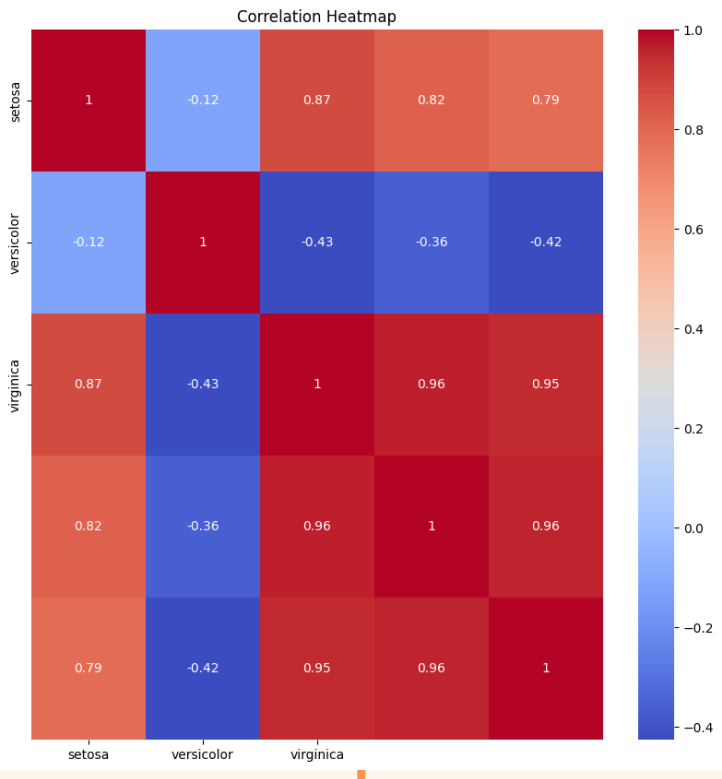
A scatter plot was constructed to visually represent the distribution of iris flowers based on their sepal and petal measurements.

Each species is differentiated by a distinct marker, providing an immediate visual understanding of the dataset's structure.



Box Plot

Box plots were utilised to present a summary of the distribution of sepal and petal measurements within each species. This aids in identifying potential variations and outliers across different features.



Correlation Heatmap

A heatmap of the correlation matrix was generated to elucidate the relationships between sepal and petal measurements. This visualisation helps in understanding the interdependencies among different features.

What machine Model used in the iris?

Decision Tree Classifier

A Decision Tree classifier was implemented with hyperparameter tuning using Grid Search. The best model was identified based on its performance metrics in accurately classifying iris species.

Neural Network

A Neural Network model was constructed using the TensorFlow Keras library. The model architecture involves three layers with ReLU activation in hidden layers and softmax activation in the output layer, facilitating multi-class classification.

USER INTERFACE FOR PREDICTION

A neural network model was constructed using the tensorflow keras library. The model architecture involves three layers with “relu” activation in hidden layers and softmax activation in the output layer, facilitating multi-class classification.

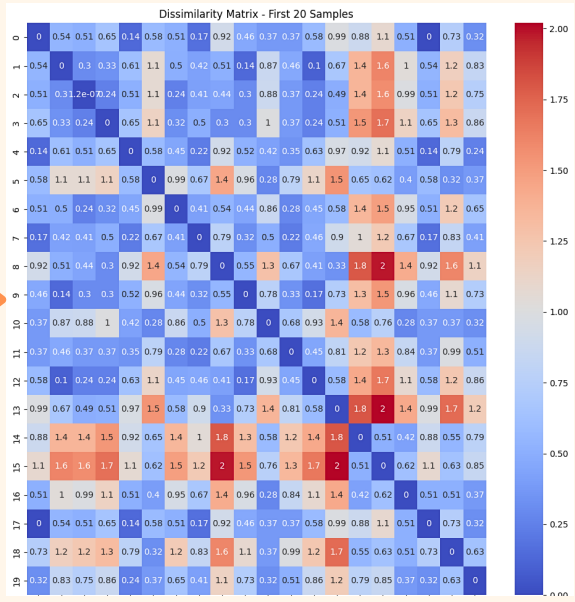
DISSIMILARITY ANALYSIS

DISSIMILARITY MATRIX

A dissimilarity matrix was computed using pairwise euclidean distances between the first 20 samples of the Iris dataset. The resulting matrix was visualised through a heatmap, offering insights into the dissimilarity relationships among the initial dataset entries.

EXPORT TO CSV

The dissimilarity matrix was exported to a CSV file, facilitating further analysis and providing a tangible record of dissimilarity values.



Reuters

The dataset under consideration originates from Reuters a renowned international news agency celebrated for its extensive and credible news coverage. Comprising a diverse array of news articles, this dataset serves as a valuable resource for natural language processing (NLP) and text classification tasks within the context of machine learning.

Data Attributes

topics

The 'TOPICS' column categorises news articles into distinct topics or subjects, providing a categorical label for each piece. This column acts as the target variable for the machine learning model, indicating the primary focus or theme of the articles.

body

The 'BODY' column encompasses the primary textual content of the news articles. Through preprocessing and analysis, this text becomes the focal point for feature extraction and model training.

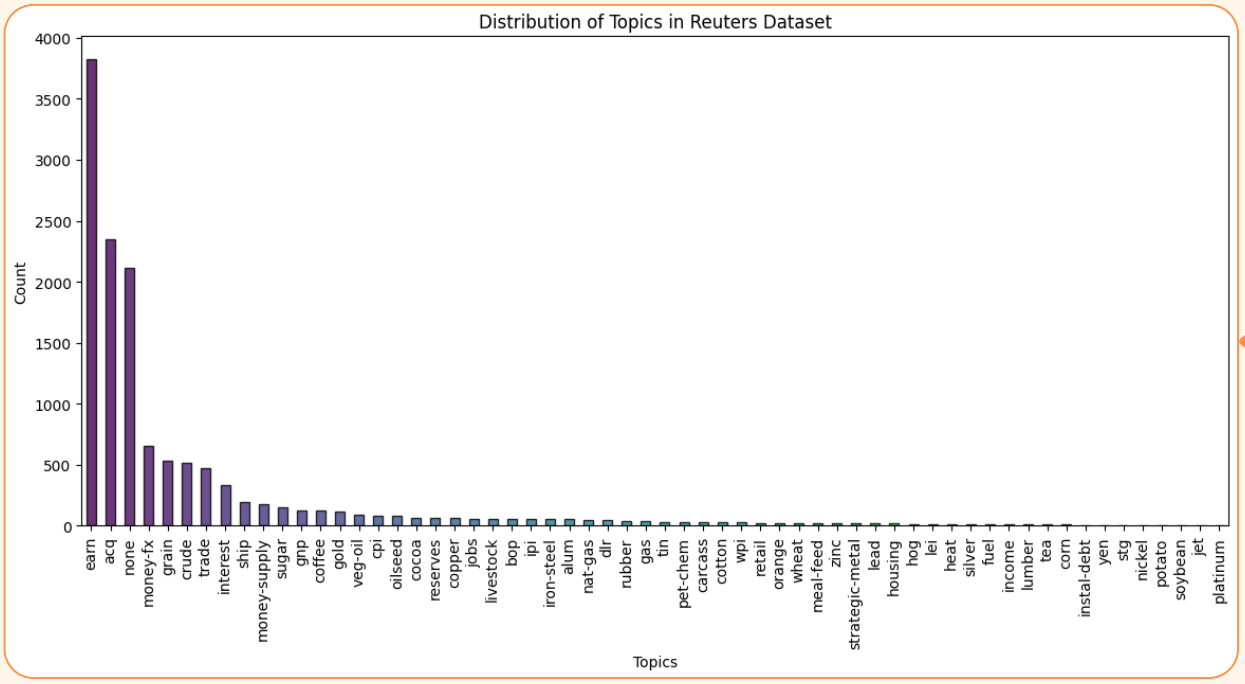
we need to make person if add sentence, then the result which topic suppose this sentence to be in which topic

What happened in preprocess?

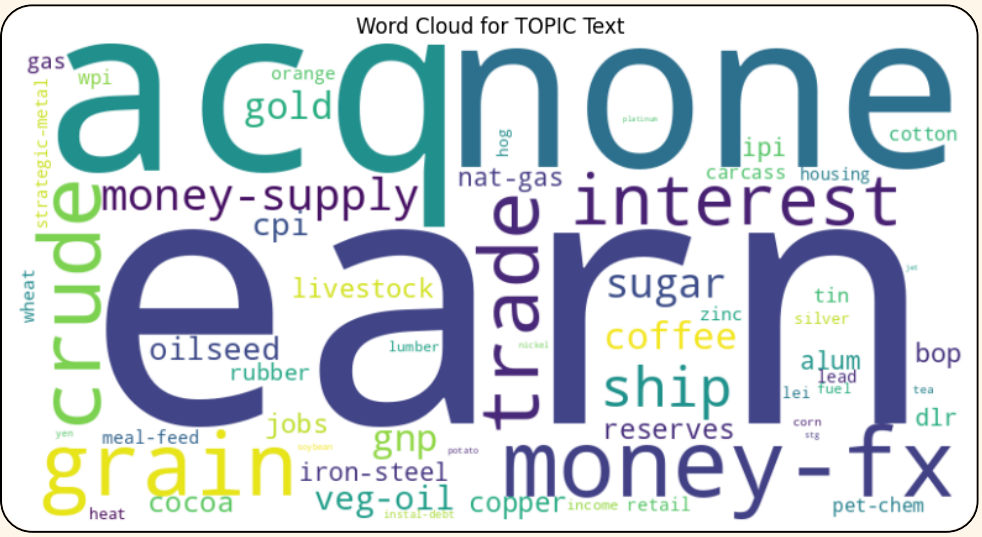
The project begins by loading the Reuters dataset from a CSV file, where columns like 'TOPICS' and 'BODY' are identified.
Basic statistical insights are provided, and missing values in the 'BODY' column are handled through filling with empty strings.
Text preprocessing involves converting the 'BODY' text to lowercase for consistency.

WHAT ABOUT VISUALIZATION?

THE CLASS DISTRIBUTION OF TOPICS IS PRESENTED USING A BAR CHART, FILTERING OUT TOPICS WITH COUNTS LESS THAN 3 FOR CLARITY.

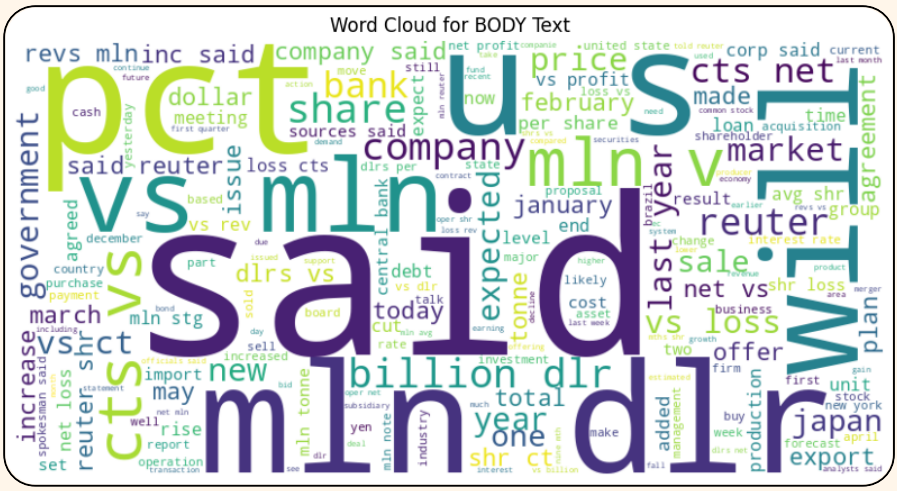


WORD CLOUDS ARE GENERATED FOR BOTH 'TOPIC' AND 'BODY' TEXTS, OFFERING A VISUAL REPRESENTATION OF THE MOST FREQUENT WORDS IN EACH CATEGORY.



IT AIDS IN IDENTIFYING AND EXPLORING THE FREQUENT TOPICS COVERED IN THE NEWS ARTICLES, ALLOWING FOR A QUICK AND INTUITIVE UNDERSTANDING.

THE CODE COMBINES NEWS ARTICLE CONTENT. IT MAKES A WORD CLOUD TO SHOW FREQUENT WORDS VISUALLY. THIS HELPS QUICKLY SEE COMMON TERMS IN THE DATA. THE VISUALISATION IS SIMPLE AND AIDS CONTENT EXPLORATION.



What machine Model used in the Reuters?

Preparing the Data (Encoding body)

The code begins by preparing the text data from the Reuters dataset. It limits the number of words to consider and converts the text into numerical sequences, this will ease the process of making Text to number

Encoding Topics

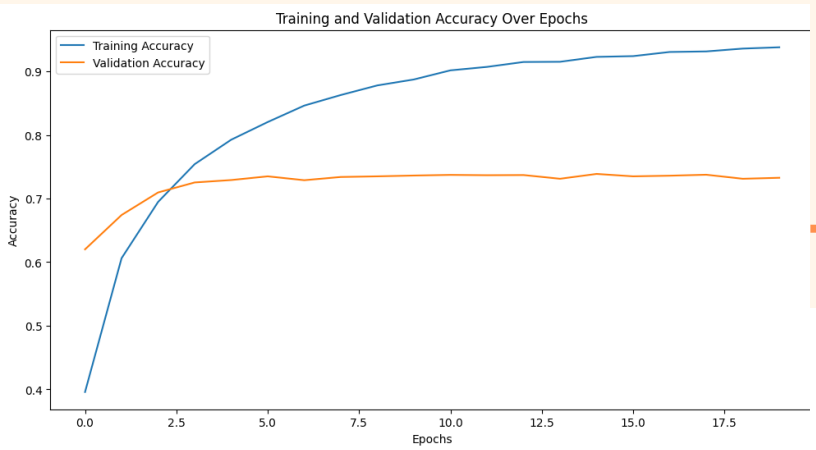
It encodes the topic labels into numerical values so that the model can understand and learn from them.

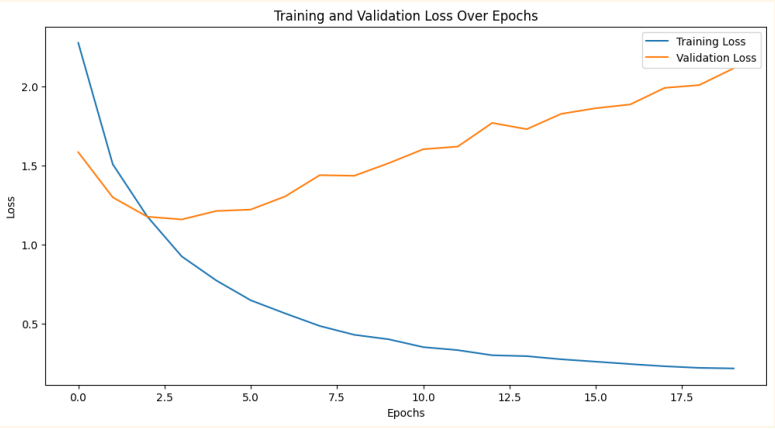
Building the Model

The neural network model is created.
It involves layers that help the model understand the relationships within the data.
The model is designed to learn patterns and connections in the text.

Visualizing Training Progress

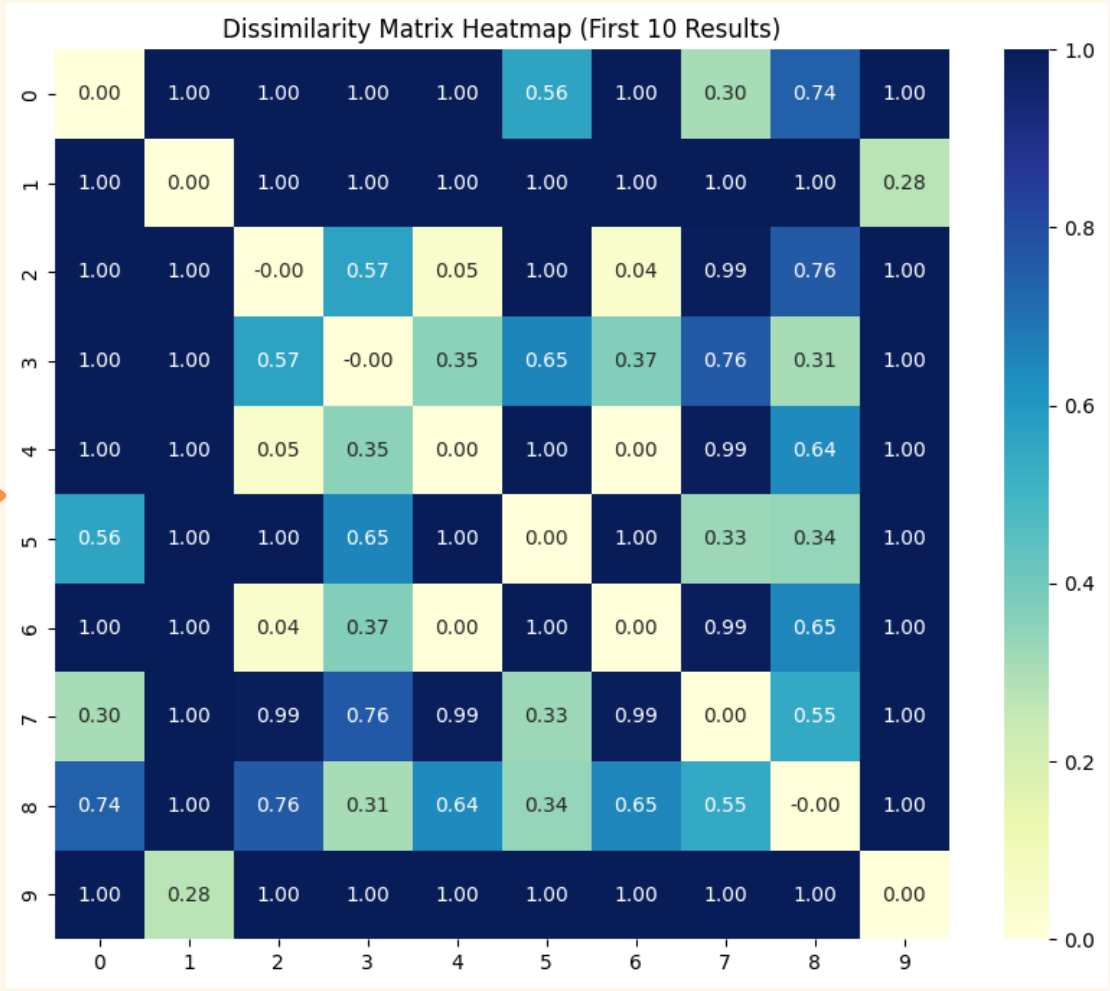
Graphs are created to show how well the model is learning over time. This helps to understand if the model is improving or if there are areas that need adjustment.





Analyzing Model Predictions and Similarity

Assuming the model is trained, the code aims to understand how similar or dissimilar the predictions are for the test data. It uses a concept called "cosine similarity," a mathematical measure to assess the similarity between two non-zero vectors.



Interactive Prediction

An interactive function is created to allow users to input new text and see the model's prediction for the topic of that text.

```
Enter Text: Global demand for dollars is on th
1/1 [==============================] - ETA: 0s
1/1 [==============================] - 0s 21ms/step
Predicted Topic: earn
```

Thats' All!