Data and Artificial Intelligence

Cyber Shujaa Program

Week 11 Assignment
Natural Language Processing

Student Name: Denis Kombe

Student ID: CS-DA01-25094

Table of Contents

Objectives

1. Importing and loading the pre-trained BERT tokenizer.
2. Listing 10 sentences.
3. Compare sentence pairs using BERT embeddings.
4. Predicting semantic similarity using a defined threshold.
5. Evaluate prediction accuracy against ground truth.
6. Explaining the NLP concepts used in this assignment.

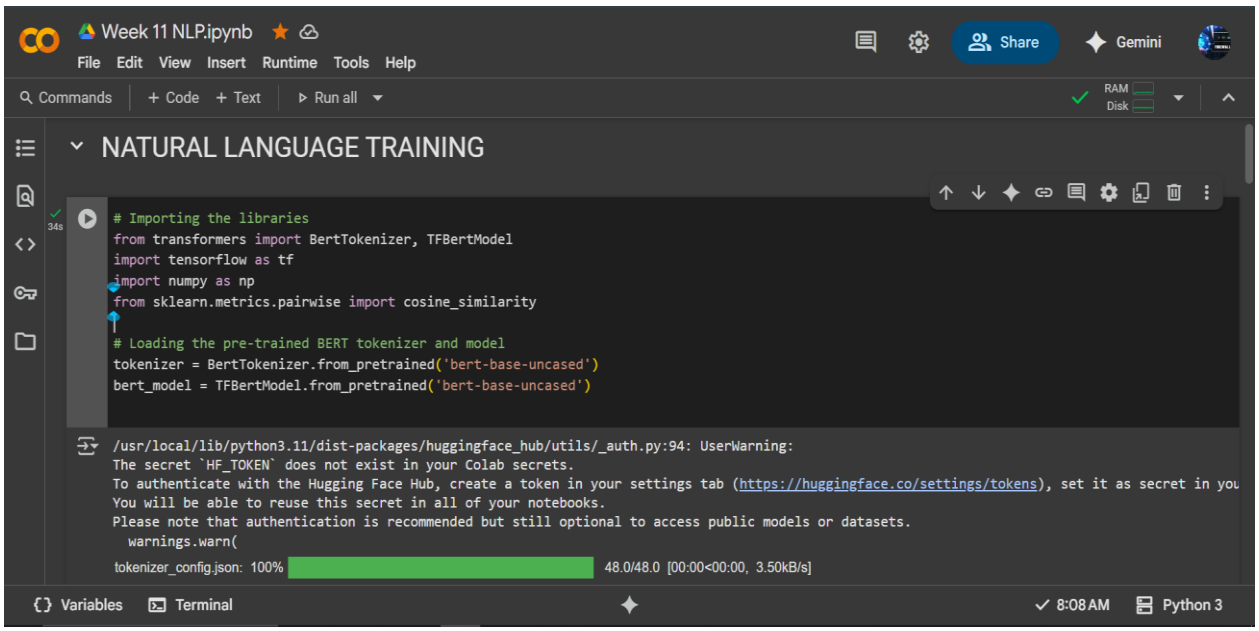
Introduction

This assignment focused on exploring sentence similarity using a pre-trained BERT model. The goal was to analyze how BERT understands the meaning of different sentences in context and how this can be used to measure how similar or different they are. By comparing 10 sentence pairs, I applied cosine similarity to the embeddings and used a set threshold to predict whether the sentences were similar or not. This assignment made me understand important NLP concepts and how they apply in real tasks.

Tasks Completed

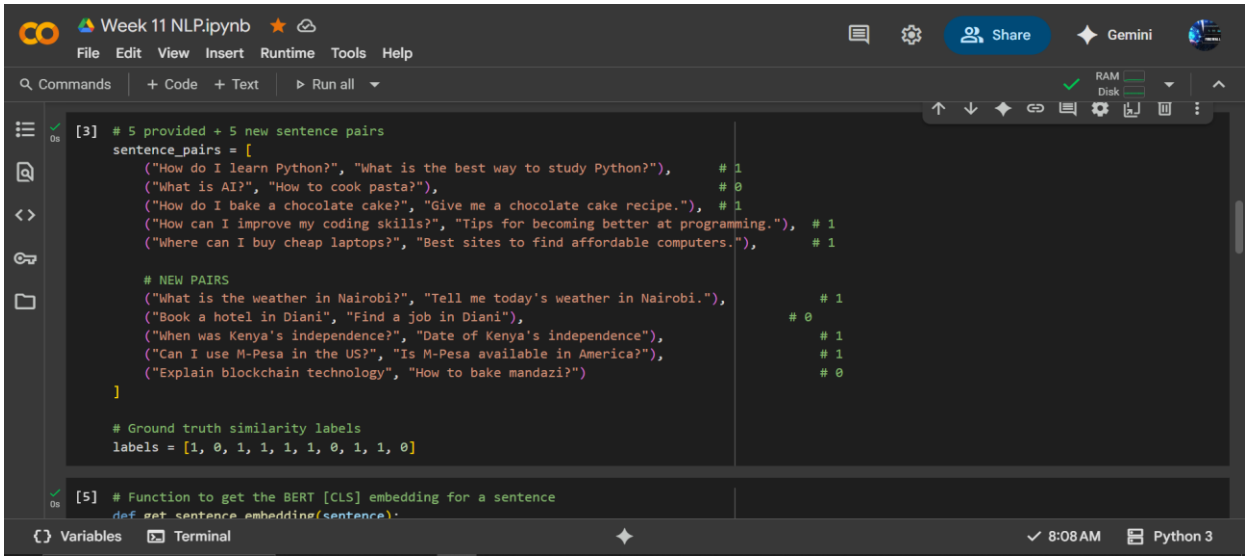
Step 1: Importing and loading the pre-trained BERT tokenizer.

Started by importing the necessary libraries needed for this project and loading the pre-trained BERT tokenizer and model.



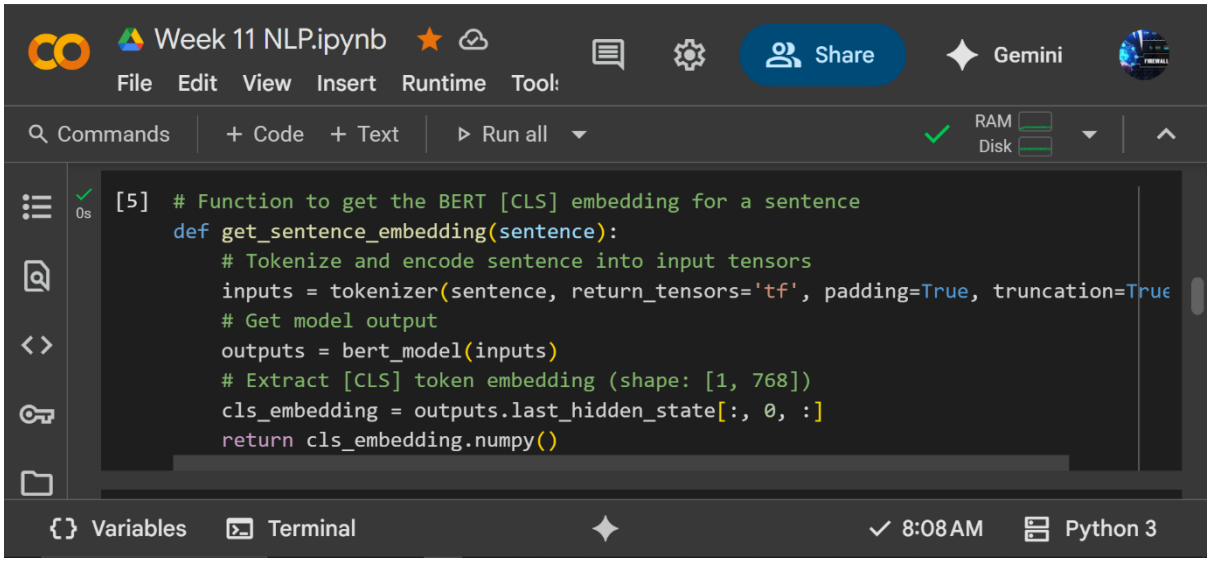
Step 2: Listing the 10 sentence pairs.

I added 5 more sentences including and labeled each sentence based on the similarities they had.



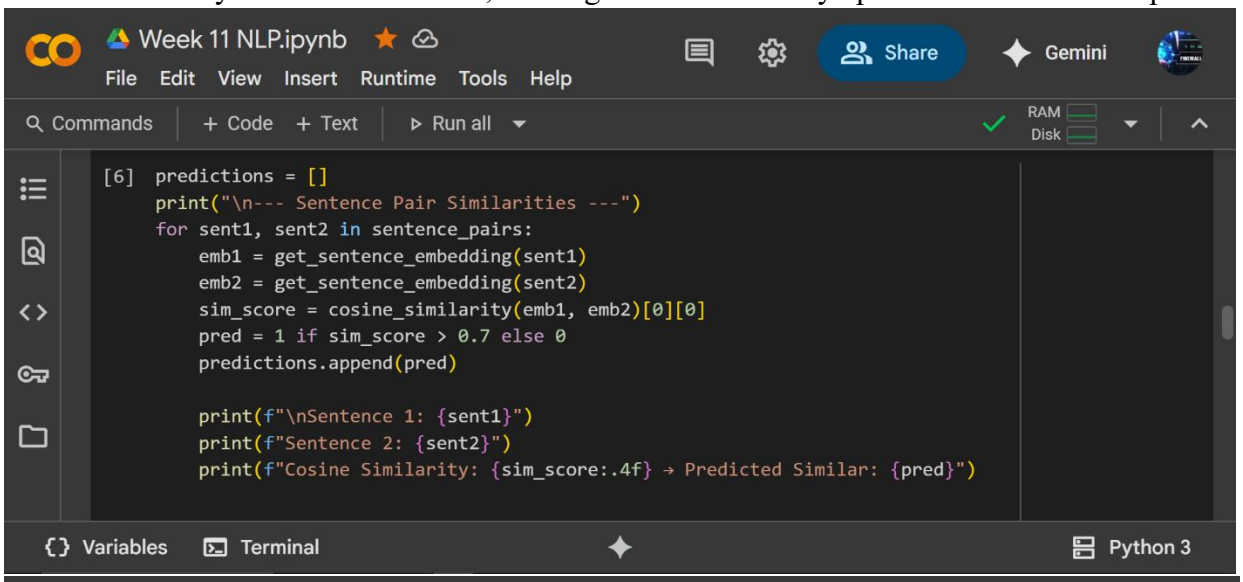
Step 3: Compare sentence pairs using BERT embeddings.

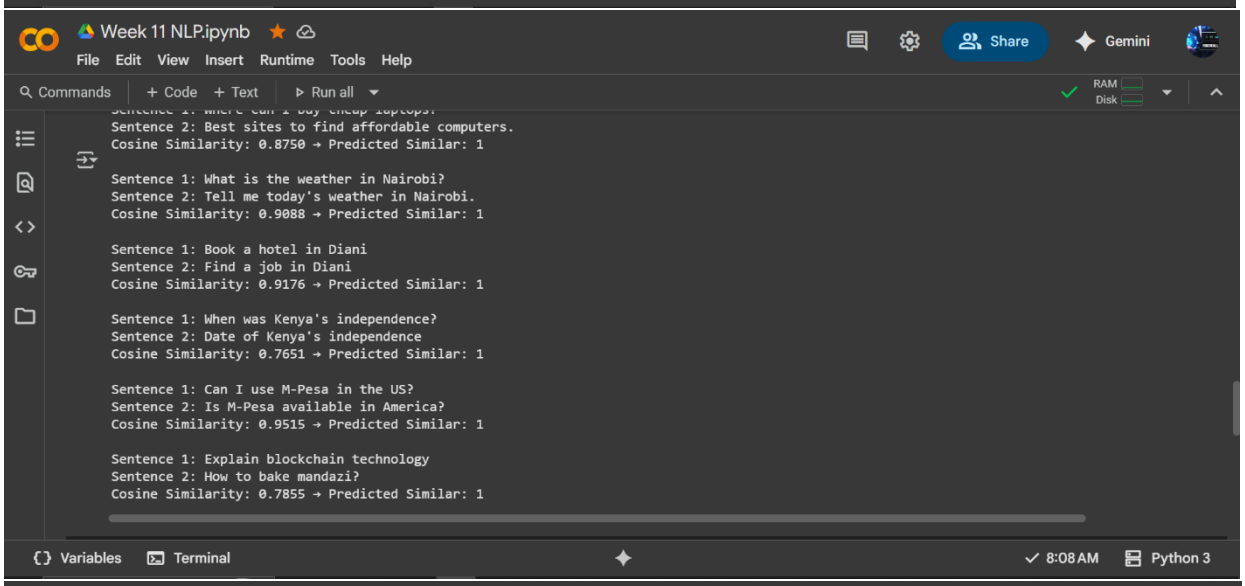
Applying a pre-trained BERT model for encoding a total of 10 sentence pair into contextual embeddings. Each pair is processed to extract sentence-level meaning using the CLS token.

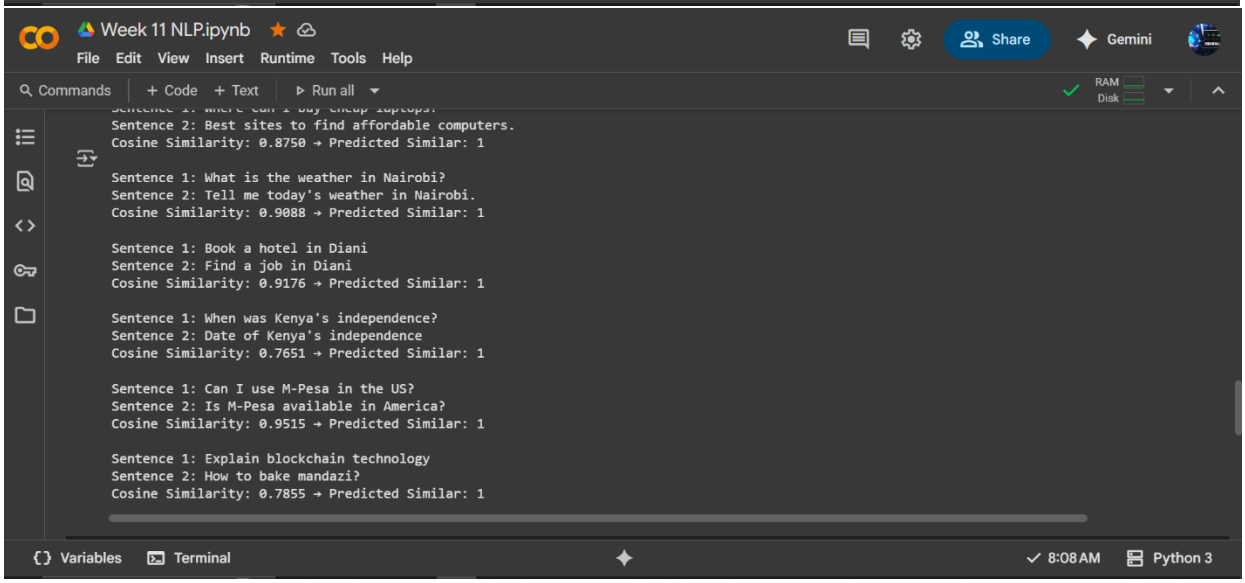


Step 4: Predicting semantic similarities using a defined threshold.

To classify each sentence pair as either similar (label = 1) or not similar (label = 0) using a cosine similarity threshold of 0.7, and generate a binary prediction for each pair.

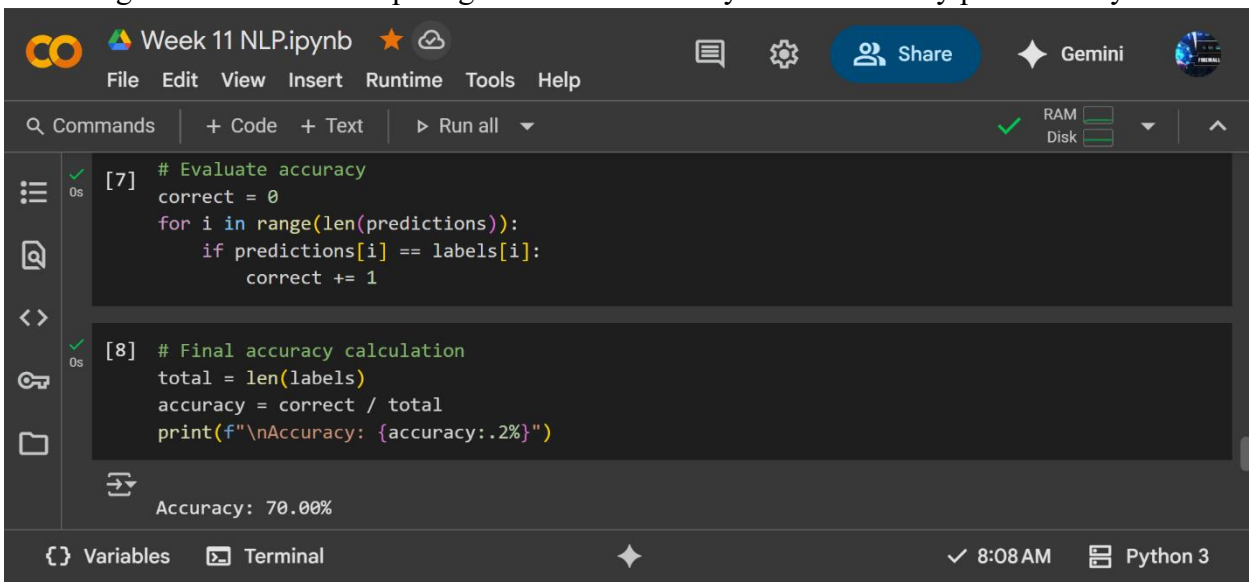






Step 5: Evaluating prediction accuracy

Assigning similarity labels for all 10 sentence pairs, then comparing the model's predictions with the ground truth and computing the overall accuracy of the similarity prediction system.



Step 6: Explaining the NLP concepts in this assignment

1. How does BERT differ from traditional NLP approaches like Bag of Words or TF-IDF??
   - Traditional NLP approaches like Bag of Words or TF-IDF create fixed-size vectors based only on word frequenct hence ignoring the meaning or order of words.
   - BERT understands context and word meaning in a sentence using its attention-based architecture. It generates contextual embeddings which differ based on sentence meaning.

2. What is the role of the encoder in the BERT model, and how is it used in this assignment?
   - The encoder in BERT learns the relationships between words in a sentence.
   - In this assignment, sentences are fed into BERT and the encoder outputs embeddings representing the meaning of each token. The CLS token's output from the encoder is used as a representation of the entire sentence.

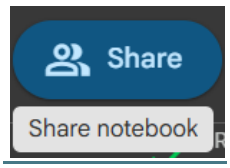
3. What are contextual embeddings? How are they generated and used in this code?
   - Contextual embeddings are word vectors that change depending on surrounding words. For example, "bank" in "river bank" vs "financial bank" will have different vectors.
   - BERTS generated these using self-attention and multiple transformer layers.
   - In this assignment, the CLS token embedding is used as a context-aware sentence representation to compare similarity between sentence pairs.

4. Why is the [CLS] token used for sentence similarity in this code?
   - To compare sentence meaning using cosine similarity.

5. What is cosine similarity, and why is it useful in comparing embeddings?
   - Cosine similarity measures the angle between two vectors and ranges from -1 (opposite) and 1 (identical).
   - It is useful because it helps quantify how similar two sentence embeddings are.

Link to Google Colab Notebook

Notebook Link to Code:



Conclusion

From this assignment, I have learned how powerful BERT is when it comes to understanding the actual meaning of words depending on how they are used in a sentence. Unlike traditional methods, BERT gives context to words, and that makes a big difference when measuring similarity. I also got to work with cosine similarity and understand how it helps us decide how close two sentences are in meaning. Overall, the experience gave me confidence in using pre-trained models and strengthened my understanding of contextual embeddings.