Investigating Models

1. Running semantic_prox_W2V (Provided by professor Carlo)

Output

```
Distance between the sentence -- Obama speaks to the media in Illinois -- and -- The president greets the press in Chicago -- is:
    0.347666557390307

Distance between the sentence -- Obama speaks to the media in Illinois -- and -- The hawk flies high in the sky -- is:
    0.30257127500320263
```

Initially, I experimented with the file semantic_prox_W2V, which was given by Professor Carlo. The model failed to distinguish between the first and third sentences, and even indicated that these two sentences were more alike than the first and second sentences. The WMD between the first and third sentences is approximately 0.303. Interestingly, this distance is smaller than the one between the first and second sentences, suggesting that, according to the Word2Vec model, these two sentences are more semantically similar. This might be a bit counter-intuitive, since we can tell, the second sentence seems more similar to the first one. This discrepancy might be due to the training data (Wikipedia page of Barack Obama), which might not have provided enough diverse contexts for the model to accurately learn the semantic relationships between words.

2. Running a sentence transformer from hugging face -

Link - [sentence-transformers/all-MiniLM-L6-v2 · Hugging Face]

This is very similar to the 2nd file provided by the professor with the exception that I am not using pinecone for the storing and retrieving vectors and I am not comparing with the embeddings from a corpus but with the two sentence embeddings created by the encoder of sentence tokenizer. This model intended to be used as a sentence and short paragraph encoder. Given an input text, it outputs a vector which captures the semantic information. The sentence vector may be used for information retrieval, clustering or sentence similarity tasks. So I used this to check for semantic similarity between sentences which will be useful for tokenization

This was the output that I got -

```
Cosine similarity between 'Obama speaks to the media in Illinois' and 'The president greets the press in Chicago' is:  0.6002549
Cosine similarity between 'Obama speaks to the media in Illinois' and 'The hawk flies high in the sky' is:  0.077405974
```

The embeddings for the sentences are used to measure the semantic similarity between them. This is a direct application of the embeddings to a problem, demonstrating how they capture semantic meaning. I used cosine similarity to compare the two vectors of the two

sentences and compare similarities. We can observe that the model is able to identify similarity between sentence 1 and 2 accurately and also identify that sentence 1 and 3 are very different in terms of semantic similarity of their embeddings.

My code -

```
from sentence_transformers import SentenceTransformer
from sklearn.metrics.pairwise import cosine_similarity

model = SentenceTransformer('all-MiniLM-L6-v2')

sentences = ['Obama speaks to the media in Illinois', 'The president greets the press in Chicago', 'The hawk flies high in the sky']

embeddings = model.encode(sentences)

similarity_1_2 = cosine_similarity(embeddings[0].reshape(1, -1), embeddings[1].reshape(1, -1))
similarity_1_3 = cosine_similarity(embeddings[0].reshape(1, -1), embeddings[2].reshape(1, -1))

print(f"Cosine similarity between '{sentences[0]}' and '{sentences[1]}' is: ", similarity_1_2[0][0])
print(f"Cosine similarity between '{sentences[0]}' and '{sentences[2]}' is: ", similarity_1_3[0][0])
```

3. Using Pinecone vector Database on RoBERTa Model

```
Query: obama was born in honolulu, hawaii
Matched sentence: , Obama was born in Honolulu, Hawaii.
   --- score: 0.9118343
Matched sentence: Of his years in Honolulu, Obama wrote: "The opportunity that Hawaii offered — to experience a variety of cultures in a climate of mutual r
asis for the values that I hold most dear."
   --- score: 0.626075506
Matched sentence: The couple married in Wailuku, Hawaii, on February 2, 1961, six months before Obama was born.
   --- score: 0.625392675

Query: the president graduated from columbia university
Matched sentence: Later in 1981, he transferred to Columbia University in New York City as a junior, where he majored in political science with a specialty
ived off-campus on West 109th Street.
   --- score: 0.55831486
Matched sentence: ",Presidency
   --- score: 0.525664806
Matched sentence: After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.
   --- score: 0.501388371

Query: obama is an american politician
Matched sentence: A member of the Democratic Party, Obama was the first African-American  president of the United States.
   --- score: 0.653675735
Matched sentence: [Barack Hussein Obama II ( (listen) bə-RAHK hoo-SAYN oh-BAH-mə; born August 4, 1961) is an American politician, lawyer, and author who ser
o 2017.
   --- score: 0.644973934
Matched sentence: "Obama is frequently referred to as an exceptional orator.
   --- score: 0.641505778

Time to evaluate the matching: 0.0292 mins
(newenv) (base) naveenrenji@Naveens-MacBook-Air-3 BackEnd %
```

Output -

Report:

After carefully studying our current approach using the RoBERTa model with the Pinecone vector database through the code provided by Professor Carlo, I found that the model was able to understand queries and give relevant answers from the vector database. I also noticed that it worked much faster because it could quickly find the right information in the Pinecone database.

Proposed Model Methodology

These findings are very promising for our project. My plan is to build a vector database for a specific subject (SSE). Then, I will use the RoBERTa model to get accurate information from this database. This information will be used as context and put into a question-answering model called which I experimented with recently, to generate precise responses to chat questions. This model will also be able to have the ability to create a pipeline that separates statistical and SSE queries and pass the query to the right model.

This method combines finding information using a pinecone vector database and RoBERTa model and generating responses using the context with a QnA model, which will make our conversational AI model have high-quality interactions. Combining this model with the RoBERTa and Pinecone-powered information retrieval system will result in accurate and context-based responses, making conversational responses more accurate and efficient. I am currently working on this method and finding ways to make it even better.