RQ 5.1: In the first half of the lab, you produced t-SNE plots of BERT embeddings. You plotted these embeddings in different colors depending on which category of text they came from. How do you interpret the results? What difference(s) did you observe between the plot in Problem 1 and that in Problem 2? Summarize your observations in a few sentences.

From the plot in Problem 2, we can see that the word embeddings for the same word "record" are more similar to each other if they belong to one class. Similary, the sentences embeddings are in the same situation.

The difference between the two plots is the boundary for different clusters are obscure in word embedding. However, that in sentences embeddings are much clear.

RQ 5.2: In Task 5.2, you saw the extractive & abstractive summaries side-by-side, as well as the ROUGE-2 scores computed for them. Which method obtained the higher ROUGE-2 score in your testing? Which method produced the "better" summaries in your opinion?

In my assignment, the score for abstractive is the highest.

In my opinion, the abstractive method produced the "better" opinion.

RQ 5.3: In Task 5.3, you ran the text generation with different temperature values. What happens when the "temperature" is close to zero? What happens for higher temperature values (> 1)? At which temperature setting did the model generate the "best" summaries, in your opinion?

When the "temperature" is close to zero, the summary is pretty close to the original context, which means the content generated is not like a summary.

When the "temperature" is larger than one, I found some random words are generated. Sentences cannot be understood. Too much randomness reduce the accuracy of the content generated.

In my opinion, when the temperature is set to be one, the result is the best.