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.

- we can see that in the projected vector space, vectors from different classes tend to cluster up exclusively from each other. Such clustering can be interpreted as that those embedded vectors from the same class(i.e. with similar meanings) are closer with each other than with those from other classes.
- And such clustering behaves better in the embedding of an entire sentence than the word "record". In Problem 2, we have different classes that clearly make their own clusters. While in Problem 1, it's more like splitting bigger clusters into subclusters. This indicates that the embedding of the word "record" from the same class is similar to each other, but not dissimilar enough with other classes in some cases.

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?

- abstractive_rouge_2 is higher in our testing
- the extractive summaries are better in my opinion. The abstractive model we are using is a simple one. Thus not as powerful as its counterparts such as ChatGPT. Simple methods might be more reliable under such case.

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 we have a lower temperature close to zero, the output tends to be fixed when we rerun
- when we have a higher temperature(> 1), the output tends to vary when we rerun.
- seems lower temperature will make the model fall into local optimal easily, but local optimal is good enough here because there is no "standard summarization" in we humans' view. Being not far away from optimal is good enough to yield a decent output. While high temperature values make model searching too arbitrarily and seem hard to converge.
- temperature=0.1 yields the best result in my view.