Lab 5: LLMs and Text Summarization
Reflection Report

TDDE16 – Text Mining

Anonymous ID
96764

IDA
Linköping University

December 4, 2023

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.*

The plot shows a visualization of the high-dimensional embedding data, colored according to the category label. Results from problem 1 shows OK clustering between the labels, although some embeddings from different labels get mixed up. This might be reasonable, however, since all embeddings are "based" on the word 'record', while having different context. This should in return give embeddings that are closer to each other, i.e., more similar, even though the surrounding context is about either companies, athletes, or albums. Judging from the plot, it also seems that the word embeddings labeled as *Company* or *Athlete* is further away from each other in vector space, while for example *Company* and *Album* is closer. This should mean that the word record in a company context is more similar to the word record in an *Album* context rather than in an *Athlete* context.

Looking at the results from problem 2, where we instead plotted the sentence embeddings, much more clear clustering can be seen. This did not surprise us, since the sentence embeddings encapsulates the overall content of the input sentence. Some *Company* and *Album* sentences are still "mixed up", however, for the most part the clustering is very clear. The same observation is also made here, where judging from the plot *Company* sentences are more similar to *Album* sentences than *Athlete* sentences, even though all are about 'records' in some way.

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 our testing, the extractive method obtained the best ROUGE-2 score. We also agreed that the extractive method in our case produced "better" summaries. Our abstractive model often produced summaries in the form of numbered lists. This was probably due to how we prompted the model, and because of hardware constraints, we were not able to experiment as much with this as we would have liked. Playing around with the prompting as well as some model parameters would maybe yield better results for the abstractive model.

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?*

In basic terms, the temperature is a parameter for "model creativity". Setting the temperature closer to zero makes the model more precise and less "creative", while a higher temperature makes the model more creative, not always picking the next word by highest probability.

In our testing, we found that as we lowered the temperature, the model got more and more "extractive" rather than "abstractive" in its summaries. I.e., it rather took words and phrases from the original prompt, than building its own sentences. In this task, we actually thought the summaries got better this way, since it would be closer to the original text, with lower risk of "hallucinations". However, the problem with itemized list still persisted, which again was probably due to how we prompted the model. The best temperature was in our opinion quite low, at 0.2.

Increasing the temperature to values above 1 made the model very confusing, and at times it was unable to produce sentences that made any sense. For example, a temperature value of 1.5 could yield text that at first glance looks good, but when reading it, it made no sense. Increasing the temperature thus increase the risk of hallucination.