**Assignment 3: Embedding with BERT**

By a negative silhouette score, we can say that the classes are badly grouped.

Accuracy value is 5%, which means that the model is not well capturing the relationship between product descriptions and their classes (subtitles).

You can also see it in the plot. Those in the same class are not very far from each other in the plot.

**Assignment 4: Similarities between texts**

**Cosine similarity:**

Cosine understood that the style of writing and theme content coincided, even though they are not the same words.

For the top 3, values are low (0.29, 0.23, 0.16 respectively). In these kinds of texts (product descriptions) is usual and expected, as the writing is creative and not repetitive.

**Jaccard similarity:**

It is lower than in Cosine as it expects an exact coincidence with words (0.22, 0.16, 0.12 respectively).

Even though phrases like “legendary classic” or “comfort” sound similar in a semantic level, if the words are not exactly the same, Jaccard doesn’t recognize them.

**Conclusion**

Even if the values may seem low, this doesn’t mean that products aren’t similar. Product descriptions and advertisements use creative and unique writing for each of the products, which reduces similarities in text.

TF-IDF + Cosine was the most efficient one to capture similarities and could be better to other tasks like grouping products by category or model, or even for personalized recommendations.

**Assignment 5: N-grams**

For bigram, after “the”, we got “first”, “same”, “most”, “other”, “``”.

These are plausible next words after “the”. They are common adjectives or determiners that follow “the”. The presence of opening quotation mark is a bit odd but can happen since punctuation wasn’t filtered out in the input text, or those quotes appear frequently in the corpus.

For trigram, after “in the”, we got “world”, “first”, “united”, “same”, “past”.

This looks very reasonable, as phrases like “in the world”, “in the first”, “in the united” (US), “in the same”, “in the past” are all very common expressions.

Qualitatively, your model is capturing common collocations and phrases quite well.

Quantitatively, we split data into 90% train data and 10% test data, and then evaluate accuracy for both n=2 and n=3.

For bigram, top 5 accuracy is 23.60%, so in 23.6% of cases, the next real term in the test train is within the top 5 predictions in our model. In other words, our model succeeds in predicting the next words in 1 over 4 or 5 cases, when there are 5 tries per prediction.

For a very simple model, it is a good performance marker. Top 5 Accuracy for bigrams without smoothing are between 15 and 25%. Hence, the model is working pretty well.

For trigram, top 5 accuracy is 14%. This can be due to data sparsity, as it takes a lot more possible combinations of 2 words. It means it has more contexts, so it succeeds less in predicting next words. These performance evaluation marks are normal in models without smoothing.