BUS 243 Name:

Quiz 3.

Problem. We can use Latent Semantic Analysis to represent a bag-of-words or a TF-IDF vector as compact meaning vectors called "word-topic vectors" or "document-topic vectors." These vectors can be used to represent the meaning of a document in a lower-dimensional space and to identify the topics that are discussed in a document. However, there are some cases where these topic vectors may not be useful. Supply your intuition of such cases.

The degree of correlation between columns in a dataset affects the corresponding principal components' ability to explain the original data's variance. Specifically, less correlated columns tend to result in principal components that explain less variance in the original data.

Of course, we can address this question conceptually. One such case is when the documents are highly specialized or technical, and the vocabulary is domain-specific. In such cases, the words in the documents may not frequently appear in the training corpus, making it difficult for LSA to identify the underlying topics. For instance, if we use LSA to analyze research papers in a specific field, we may not get meaningful topic vectors if the vocabulary in the papers is not common in other domains.

Another case where topic vectors may be useless is when the documents are too short or sparse. LSA requires significant text data to identify the underlying topics accurately. If the documents are too short or sparse, LSA may not have enough information to identify the meaningful relationships between the words and the topics.