Chapter 5. Text Clustering

there are many methods for text clustering, from graph-based neural networks to centroid-based clustering techniques, a common pipeline that has gained popularity involves three steps and algorithms:

1. **Convert the input documents to embeddings with an *embedding model*.**
2. **Reduce the dimensionality of embeddings with a *dimensionality reduction model*.**
3. **Find groups of semantically similar documents with a *cluster model*.**

Embedding Documents

The first step is to convert our textual data to embeddings, as illustrated in Figure 5-3. Recall from previous chapters that embeddings are numerical representations of text that attempt to capture its meaning.

Choosing embedding models optimized for semantic similarity tasks is especially important for clustering as we attempt to find groups of semantically similar documents. Fortunately, most embedding models at the time of writing focus on just that, semantic similarity.