Comparison and Analysis:

In this lab, we tried two embedding approaches (Doc2Vec vs. Word2Vec) across three tested vector dimensions (50, 100, 150).
In the cluster_visualizations folder, we have 6 images, and we'll compare the 2 images with the same dimensions one by one.

50-Dimension Embeddings

Doc2Vec: 3 Clusters exist, though there is quite some overlap around the boundary in the PCA projection. Keywords revolve around broad finance or market terms. This method tends to produce coherent clusters but might miss some nuanced subtopics since the vector size is relatively small.
Word2Vec: shows a separation between two clusters. Because it relies on word embeddings averaged together, shorter posts or posts with fewer unique words can be less distinctly positioned. Generally, if the vocabulary is large, 50 dimensions may feel limiting—some nuance is compressed.

100-Dimension Embeddings

Doc2Vec: Increasing the dimensionality helps Doc2Vec capture more subtle variations among posts than 50 dimension. Clusters become more distinctly separated (visually in the PCA plots, we see less overlap), which suggests a better internal consistency of clusters.
Word2Vec: With 100 dimensions for word embeddings, the model typically captures richer word-level semantics. Supposedly, with more dimensions we will have a better chance of distinguishing certain topic-specific terms. The resulting clusters in the PCA visualization is slightly more compact compared to 50D, reflecting an improved representation of subtle differences.
Both Doc2Vec and Word2Vec saw a performance improvement from 50D to 100D. Word2Vec might especially benefit if the dataset contains a wide range of vocabulary. Meanwhile, Doc2Vec can better encode full-post context in 100 dimensions. Often, 100D is a sweet spot for both methods —enough capacity to capture variation without overfitting.

150-Dimension Embeddings

Doc2Vec: With 150 dimensions, Doc2Vec can capture even more document-level nuances. However, since our dataset is not large or varied enough, high-dimensional embeddings makes it harder to cluster cleanly (the "curse of dimensionality" degrade performance or cause sparser distributions). The PCA plots show more overlaps.

Word2Vec: Similarly, Word2Vec with 150 dimensions may capture more nuanced word relationships. In our plots, we don’t observe any significant improvement compared to 100D.
For both methods, 150 dimensions either bring out worse cluster coherence or lead to a diminishing return.

Note the actual document vectors exist in a much higher-dimensional space, but when reduced to 2D for visualization, some of the clustering patterns get compressed or distorted.

Conclusion

- Doc2Vec: Better at capturing overall document context, which can lead to coherent “theme-based” clusters.
- Word2Vec: More granular, capturing nuances at the word level; however, averaging can dilute context if posts differ greatly in length or structure.

The “best” choice depends on:

1. Nature and size of the data (short vs. long posts, large vs. small corpus).
2. Focus of the analysis (broad thematic grouping vs. keyword-based grouping).

3. Empirical metrics (e.g., silhouette scores, cluster interpretability).

Doc2Vec at 100 dimensions often yields solid results for clustering entire Reddit posts on broad topics, while Word2Vec at 100 or 150 dimensions can be strong if the emphasis is on capturing more fine-grained distinctions in word usage.