DOCUMENT CLUSTERING

Link to colab notebook:
https://colab.research.google.com/drive/172jglmxRx-J6QMwA6znZ7KQI3sCyEhGS?usp=sharing

1. k = 4

Cluster Associations:

Linear algebra - 1
Data Science - 1
Artificial intelligence - 1
European Central Bank - 2
Financial technology - 1
International Monetary Fund - 2
Basketball - 0
Swimming - 3
Cricket - 0

2. k = 8

Cluster Associations:

Linear algebra - 2
Data Science - 2
Artificial intelligence - 0
European Central Bank - 5
Financial technology - 4
International Monetary Fund - 5
Basketball - 1
Swimming - 3
Cricket - 1

3. k = 12

Cluster Associations:

Linear algebra - 1
Data Science - 2
Artificial intelligence - 5
European Central Bank - 0
Financial technology - 3
International Monetary Fund - 0
Basketball - 10
Swimming - 4
Cricket - 6

There are a few things to be observed in this problem. First of all, the dataset size is very small. So the algorithm converges quickly. However, this introduces a few problems. We need to be careful with the selection of k-values. As we can see for k=12 significant overfitting is observed. We only have 9 training examples and selecting k=12 will almost assign separate classes to each one of them. Looking at the training examples and considering the small dataset k=4 looks good enough to classify the given documents into broad categories. k=8 also seems to work fine but probably requires a little more training data to identify the clusters better.