Document clustering using K Means algorithm

The following clusters are found with varying values of K for the given 9 documents:

1. K = 4

Cluster	Titles
1	Basketball, Cricket
2	Linear algebra, Data science, Artificial Intelligence
3	Financial technology, International Monetary fund, European Central Bank
4	Swimming

2. K = 6

Cluster	Titles
1	Linear Algebra
2	European Central Bank, International Monetary fund
3	Financial technology
4	Data Science
5	Basketball, Swimming, Cricket
6	Artificial intelligence

3. K = 8

Cluster	Titles
1	Cricket
2	Basketball
3	Swimming
4	Artificial Intelligence
5	Financial Technology, International Monetary Fund
6	Linear Algebra
7	Data Science
8	European Central Bank

4. K = 12

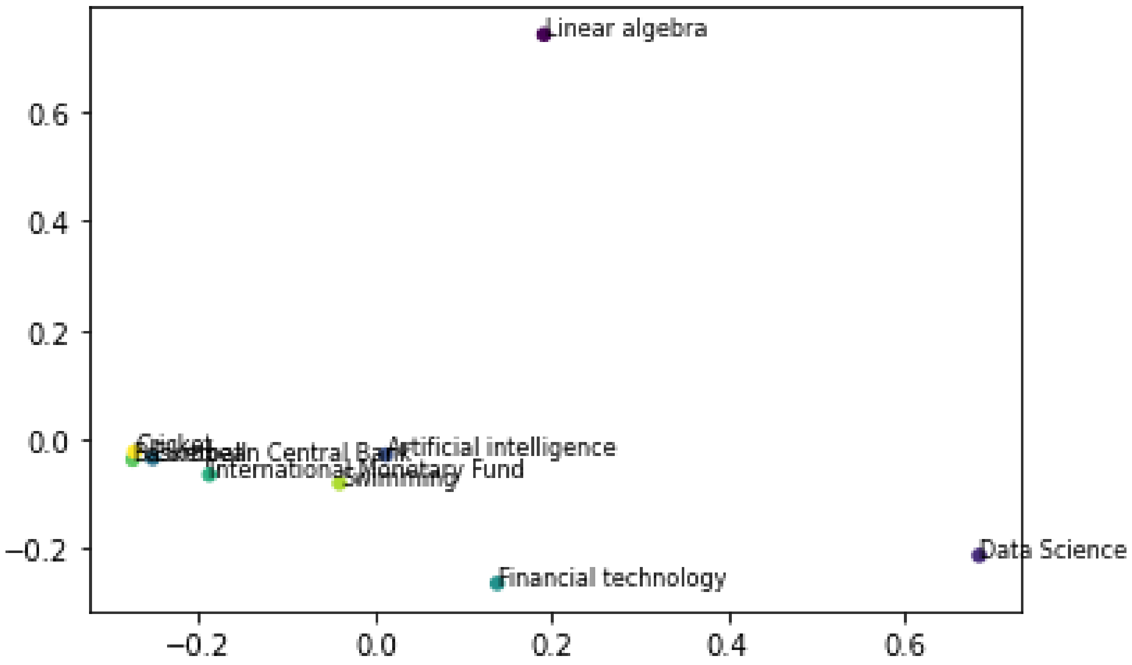
Cluster	Titles
1	Cricket
2	Basketball
3	Swimming
4	Artificial Intelligence
5	Financial Technology
6	Linear Algebra
7	Data Science
8	European Central Bank
9	International Monetary Fund
10	<Empty>
11	<Empty>
12	<Empty>

c) The optimal option for K from the above is k = 4 for the following reasons:

1) We can clearly see semantically similar concept documents from different categories, i.e AI, Finance and Sports,
2) For k = 6/8 we can see only very few concepts getting clustered together, Since we have only 9 concepts it is essential that we take k to be < N/2 at least since we can see clear semantics which are maintained by tf-idf vectorization as well

Note: All cluster centroids initializations are done from the dataset since it is very difficult to find vital seed initializations from random values as K-means converge to a local minima easily.

Even from PCA reduction to 2 principal compositions, we can clearly see good 4 clustering (Note: this may not be a perfect representation of a feature vector of more than 8K attributes but still has some clustering)


Linear algebra
Cricket
European Central Bank
Artificial intelligence
International Monetary Fund
Swimming
Data Science
Financial technology