In [1]:

```
import pandas as pd
import wikipedia
articles=['Linear Algebra',
          'Data Science',
          'Artificial intelligence',
          'European Central Bank',
          'Financial technology',
          'International Monetary Fund',
          'Basketball',
          'Swimming',
          'Cricket']
lst=[]
title=[]
for article in articles:
   print("loading content: ",article)
   lst.append(wikipedia.page(article).content)
   title.append(article)
```

```
loading content:  Linear Algebra
loading content:  Data Science
loading content:  Artificial intelligence
loading content:  European Central Bank
loading content:  Financial technology
loading content:  International Monetary Fund
loading content:  Basketball
loading content:  Swimming
loading content:  Cricket
```

In [2]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(stop_words={'english'})
X = vectorizer.fit_transform(lst)
```

In [5]:

```
from sklearn.cluster import KMeans
k = 4
model = KMeans(n_clusters=k, init='k-means++', max_iter=200, n_init=10)
model.fit(X)
labels=model.labels_
wiki_cl=pd.DataFrame(list(zip(title,labels)),columns=['Document','Cluster'])
print(wiki_cl.sort_values(by=['Cluster']))
```

```
                     Document  Cluster
2      Artificial intelligence        0
3        European Central Bank        0
5  International Monetary Fund        0
6                   Basketball        0
7                     Swimming        0
8                      Cricket        0
0                Linear Algebra        1
1                  Data Science        2
4          Financial technology        3
```

In [6]:

```
k = 8
model = KMeans(n_clusters=k, init='k-means++', max_iter=200, n_init=10)
model.fit(X)
labels=model.labels_
wiki_cl=pd.DataFrame(list(zip(title,labels)),columns=['Document','Cluster'])
print(wiki_cl.sort_values(by=['Cluster']))
```

```
                    Document  Cluster
6                 Basketball        0
8                    Cricket        0
1                Data Science        1
7                    Swimming        2
0              Linear Algebra        3
4         Financial technology        4
2     Artificial intelligence        5
5  International Monetary Fund        6
3       European Central Bank        7
```

In [ ]:

```
# Out of the values of k = 4 and 8, k = 8 is better for the given data, since it doesn't in
# together, that have little correlation with each other (like "International Monetary Fund
# together, when k = 4). But when k = 8, they are well separated into their own categories.
# ("Basketball" and "Cricket") are grouped together in thise case.
```