1-Data Preprocessing

(Creating Complete Database file)

Importing Necessary Libraries

In [1]:
```
import pandas as pd
import numpy as np
import os

base_dir = '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAn
alytics/BigData-Programs/Mini-Project/Data'
```

Storing Country-wise Database file location into python list

In [2]:
```
dir_lst = []
for data_file in os.listdir(base_dir):
    dir_lst.append(os.path.join(base_dir,data_file))
```

In [3]:
```
dir_lst
```

Out[3]:
```
['/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-Australia.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-Canada.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-China.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-France.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-Germany.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-India.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-Iran.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-Italy.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-Japan.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-Netherlands.csv
',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-SouthKorea.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-Spain.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-Taiwan.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-UK.csv',
 '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAnalytics/Bi
gData-Programs/Mini-Project/Data/Artificial-Intelligence-US.csv']
```

Creating a Python list with country names

In [4]:
```
country_lst = ['Australia','Canada','China','France','Germany','India','I
ran','Italy','Japan','Netherlands','South Korea','Spain','Taiwan','United
Kingdom', 'United States']
```

Crossing checking number of database files and number of country names

In [5]:
```
len(dir_lst) == len(country_lst)
```

Out[5]:
```
True
```

Reading database of each country, removing unnecessary columns, adding country name column and storing those as a pandas dataframe into python list

In [6]:
```
df_lst = []

for data_file, country in zip(dir_lst, country_lst):
        print(country)
        df_tmp = pd.read_csv(data_file)

        df_tmp = df_tmp.drop(['Author(s) ID' ,'Source title','Volume', 'I
ssue', 'Art. No.', 'Page start', 'Page end', 'Page count', 'DOI', 'Link',
'Document Type', 'Publication Stage','Open Access', 'Source', 'EID'],axis
='columns')

        df_tmp = df_tmp.fillna(0)
        df_tmp['Country'] = country
        df_lst.append(df_tmp)
```

```
Australia
Canada
China
France
Germany
India
Iran
Italy
Japan
Netherlands
South Korea
Spain
Taiwan
United Kingdom
United States
```

In [7]:
```
print(country_lst[0])
df_lst[0].head()
```

```
Australia
```

Out[7]:

	Authors	Title	Year	Cited by	Country
0	Soares, J.V.B., Leandro, J.J.G., Cesar Jr., R....	Retinal vessel segmentation using the 2-D Gabo...	2006	1083.0	Australia
1	Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuch...	The graph neural network model	2009	1031.0	Australia
2	Karantonis, D.M., Narayanan, M.R., Mathie, M.,...	Implementation of a real-time human movement c...	2006	908.0	Australia
3	Mirjalili, S.	Dragonfly algorithm: a new meta-heuristic opti...	2016	865.0	Australia
4	Naseem, I., Togneri, R., Bennamoun, M.	Linear regression for face recognition	2010	768.0	Australia

Merging coutry-databases from python list into a single pandas dataframe

In [8]:
```
df = pd.concat(df_lst)
```

In [9]:
```
df.shape
```

Out[9]:
```
(67694, 5)
```

In [10]:
```
df.head()
```

Out[10]:

	Authors	Title	Year	Cited by	Country
0	Soares, J.V.B., Leandro, J.J.G., Cesar Jr., R....	Retinal vessel segmentation using the 2-D Gabo...	2006	1083.0	Australia
1	Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuch...	The graph neural network model	2009	1031.0	Australia
2	Karantonis, D.M., Narayanan, M.R., Mathie, M.,...	Implementation of a real-time human movement c...	2006	908.0	Australia
3	Mirjalili, S.	Dragonfly algorithm: a new meta-heuristic opti...	2016	865.0	Australia
4	Naseem, I., Togneri, R., Bennamoun, M.	Linear regression for face recognition	2010	768.0	Australia

Applying same data pipeline as above for databases with funding column

In [11]:
```
# Changin base_dir variable to path of database with funding column

base_dir = '/mnt/General_Stuff/Study Stuff/Documents/CDS/Sem-II/BigDataAn
alytics/BigData-Programs/Mini-Project/Data_with_sponsor_details'
```

In [12]:
```
dir_lst = []
for data_file in os.listdir(base_dir):
    dir_lst.append(os.path.join(base_dir,data_file))
```

In [13]:
```
df_lst = []

for data_file, country in zip(dir_lst, country_lst):

        print(country)

        df_tmp = pd.read_csv(data_file)

        df_tmp = df_tmp.drop(['Link','Publication Stage'],axis='columns')

        df_tmp = df_tmp.fillna(0)

        df_tmp['Country'] = country

        df_lst.append(df_tmp)
```

```
Australia
Canada
China
France
Germany
India
Iran
Italy
Japan
Netherlands
South Korea
Spain
Taiwan
United Kingdom
United States
```

In [14]:
```
df_addon = pd.concat(df_lst)
```

In [15]:
```
df_addon.head()
```

Out[15]:

	Authors	Title	Year	Cited by	Funding Details	Country
0	Tao F., Qi Q., Liu A., Kusiak A.	Data-driven smart manufacturing	2018	375.0	National Natural Science Foundation of China\n...	Australia
1	Zhang K., Gao X., Tao D., Li X.	Single image super-resolution with non-local m...	2012	362.0	National Natural Science Foundation of China\n...	Australia
2	Kristan M., Matas J., Leonardis A., Vojir T., ...	A Novel Performance Evaluation Methodology for...	2016	264.0	Seventh Framework Programme	Australia
3	Ding C., Choi J., Tao D., Davis L.S.	Multi-Directional Multi-Level Dual-Cross Patte...	2016	243.0	National Science Foundation\n\nAustralian Rese...	Australia
4	Celebi M.E., Kingravi H.A., Iyatomi H., Asland...	Border detection in dermoscopy images using st...	2008	241.0	National Cancer Institute	Australia

Adding Funding Column to previously merged database

In [16]:
```
df['Funding_Details'] = 0
```

Extracting titles from df and df_addon

In [17]:
```
titles_old = list(df.Title)
titles_new = list(df_addon.Title)
```

Adding funding values to df

In [18]:
```
count = 1
for i in range(len(titles_new)):
    for j in range(len(titles_old)):
        if titles_new[i]==titles_old[j]:
            df.iloc[j,-1] = df_addon.iloc[i,-2]
    # print(f'{count}\t{titles_new[i]}')
    count += 1
```

In [19]:
```
df.head()
```

Out[19]:

	Authors	Title	Year	Cited by	Country	Funding_Details
0	Soares, J.V.B., Leandro, J.J.G., Cesar Jr., R....	Retinal vessel segmentation using the 2-D Gabo...	2006	1083.0	Australia	0
1	Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuch...	The graph neural network model	2009	1031.0	Australia	0
2	Karantonis, D.M., Narayanan, M.R., Mathie, M.,...	Implementation of a real-time human movement c...	2006	908.0	Australia	0
3	Mirjalili, S.	Dragonfly algorithm: a new meta-heuristic opti...	2016	865.0	Australia	0
4	Naseem, I., Togneri, R., Bennamoun, M.	Linear regression for face recognition	2010	768.0	Australia	0

Saving final database to 'csv' file

In [20]:
```
df.to_csv('Complete_database.csv')
```