Assignment 10 - Recommendation System

In [1]:
```
# Import Libraries
import pandas as pd
import numpy as np
from sklearn.metrics import pairwise_distances
from scipy.spatial.distance import cosine,correlation
```

In [7]:
```
# Import Dataset
books=pd.read_csv('book (4).csv',encoding='Latin1')
books
```

Out[7]:

	Unnamed: 0	User.ID	Book.Title	Book.Rating
0	1	276726	Classical Mythology	5
1	2	276729	Clara Callan	3
2	3	276729	Decision in Normandy	6
3	4	276736	Flu: The Story of the Great Influenza Pandemic...	8
4	5	276737	The Mummies of Urumchi	6
...	...	...	...	...
9995	9996	162121	American Fried: Adventures of a Happy Eater.	7
9996	9997	162121	Cannibal In Manhattan	9
9997	9998	162121	How to Flirt: A Practical Guide	7
9998	9999	162121	Twilight	8
9999	10000	162129	Kids Say the Darndest Things	6

10000 rows × 4 columns

In [8]:
```
books2=books.iloc[:,1:]
books2
```

Out[8]:

	User.ID	Book.Title	Book.Rating
0	276726	Classical Mythology	5
1	276729	Clara Callan	3
2	276729	Decision in Normandy	6
3	276736	Flu: The Story of the Great Influenza Pandemic...	8
4	276737	The Mummies of Urumchi	6
...	...	...	...
9995	162121	American Fried: Adventures of a Happy Eater.	7
9996	162121	Cannibal In Manhattan	9
9997	162121	How to Flirt: A Practical Guide	7
9998	162121	Twilight	8
9999	162129	Kids Say the Darndest Things	6

10000 rows × 3 columns

In [9]:
```
# Sort by User IDs
books2.sort_values(['User.ID'])
```

Out[9]:

	User.ID	Book.Title	Book.Rating
2401	8	Wings	5
2400	8	The Western way: A practical guide to the West...	5
2399	8	Ancient Celtic Romances	5
2402	8	Truckers	5
2405	8	The Art Of Celtia	7
...	...	...	...
2395	278854	La crÃ³nica del PerÃº (CrÃ³nicas de AmÃ©rica)	7
2398	278854	Celtic Mythology (Library of the World's Myths...	8
2393	278854	A corrente de Trewis Scott	7
2394	278854	As valkÃrias	7
2397	278854	A Treasury of Irish Myth, Legend, and Folklore	6

10000 rows × 3 columns

In [10]:
```
# number of unique users in the dataset
len(books2['User.ID'].unique())
```

Out[10]: 2182

In [11]:
```
# number of unique books in the dataset
len(books2['Book.Title'].unique())
```

Out[11]: 9659

In [12]:
```
# converting long data into wide data using pivot table
books3=books2.pivot_table(index='User.ID',columns='Book.Title',values='Book.Rating').reset_index(drop=True)
books3

```

Out[12]:

Book.Title	Jason, Madison &amp	Stories;Merril;1985;McClelland &amp	Other Repairing PC Drives &amp	'48	'O Au No Keia: Voices from Hawai'I's Mahu and Transgender Communities	...AND THE HORSE HE RODE IN ON : THE PEOPLE V. KENNETH STARR	01-01-00: A Novel of the Millennium	1,401 More Things That P*Ss Me Off	10 Commandments Of Dating	100 Great Fantasy Short, Short Stories	...	Zora Hurston and the Chinaberry Tree (Reading Rainbow Book)	Mc Fa	Jap Pr
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		
...	...	...	...	...	...	...	...	...	...	...	...	...		
2177	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		
2178	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		
2179	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		
2180	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		
2181	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN		

2182 rows × 9659 columns

In [13]:
```
# Replacing the index values by unique user Ids
books3.index=books2['User.ID'].unique()
books3
```

Out[13]:

Book.Title	Jason, Madison &amp	Stories;Merril;1985;McClelland &amp	Other Repairing PC Drives &amp	'48	'O Au No Keia: Voices from Hawai'I's Mahu and Transgender Communities	...AND THE HORSE HE RODE IN ON : THE PEOPLE V. KENNETH STARR	01-01-00: A Novel of the Millennium	1,401 More Things That P*Ss Me Off	10 Commandments Of Dating	100 Great Fantasy Short, Short Stories	...	Zora Hurston and the Chinaberry Tree (Reading Rainbow Book)	Mc Fa Jap Pr
276726	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
276729	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
276736	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
276737	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
276744	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	
162107	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
162109	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
162113	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
162121	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
162129	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	

2182 rows × 9659 columns

In [14]:
```
# Impute those NaNs with 0 values
books3.fillna(0,inplace=True)
books3
```

Out[14]:

Book.Title	Jason, Madison &amp	Stories;Merril;1985;McClelland &amp	Other Repairing PC Drives &amp	'48	'O Au No Keia: Voices from Hawai'I's Mahu and Transgender Communities	...AND THE HORSE HE RODE IN ON : THE PEOPLE V. KENNETH STARR	01-01-00: A Novel of the Millennium	1,401 More Things That P*Ss Me Off	10 Commandments Of Dating	100 Great Fantasy Short, Short Stories	...	Zora Hurston and the Chinaberry Tree (Reading Rainbow Book)	Mor Fall T (Japa Prov
276726	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
276729	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
276736	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
276737	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
276744	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
162107	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
162109	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
162113	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
162121	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
162129	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	

2182 rows × 9659 columns

In [15]:
```
# Calculating Cosine Similarity between Users on array data
user_sim=1-pairwise_distances(books3.values,metric='cosine')
user_sim
```

Out[15]:
```
array([[1., 0., 0., ..., 0., 0., 0.],
       [0., 1., 0., ..., 0., 0., 0.],
       [0., 0., 1., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 1., 0., 0.],
       [0., 0., 0., ..., 0., 1., 0.],
       [0., 0., 0., ..., 0., 0., 1.]])
```

In [16]:
```
# Store the results in a dataframe format
user_sim2=pd.DataFrame(user_sim)
user_sim2
```

Out[16]:

	0	1	2	3	4	5	6	7	8	9	...	2172	2173	2174	2175	2176	2177	2178	2179	2180	2181
0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2177	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
2178	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
2179	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
2180	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2181	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

2182 rows × 2182 columns

In [17]:
```
# Set the index and column names to user ids
user_sim2.index=books2['User.ID'].unique()
user_sim2.columns=books2['User.ID'].unique()
user_sim2
```

Out[17]:

	276726	276729	276736	276737	276744	276745	276747	276748	276751	276754	...	162085	162091	162092	162095	162103	162107	162109	1621
276726	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
276729	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
276736	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
276737	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
276744	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
162107	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0
162109	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0
162113	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
162121	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
162129	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

2182 rows × 2182 columns

In [18]:
```
# Nullifying diagonal values
np.fill_diagonal(user_sim,0)
user_sim2
```

Out[18]:

	276726	276729	276736	276737	276744	276745	276747	276748	276751	276754	...	162085	162091	162092	162095	162103	162107	162109	1621
276726	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
276729	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
276736	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
276737	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
276744	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
162107	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
162109	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
162113	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
162121	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
162129	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

2182 rows × 2182 columns

In [19]:
```
# Most Similar Users
user_sim2.idxmax(axis=1)
```

Out[19]:
```
276726    276726
276729    276726
276736    276726
276737    276726
276744    276726
           ...
162107    276726
162109    276726
162113    161453
162121    276726
162129    276726
Length: 2182, dtype: int64
```

In [ ]:
```
# extract the books which userId 162107 & 276726 have watched
books2[(books2['User.ID']==162107) | (books2['User.ID']==276726)]
```

In [20]:
```
# Extract the books which userId 162107 & 276726 have watched
books2[(books2['User.ID']==162107) | (books2['User.ID']==276726)]
```

Out[20]:

	User.ID	Book.Title	Book.Rating
0	276726	Classical Mythology	5
9987	162107	What's Bred in the Bone	7

In [21]:
```
# extract the books which userId 276729 & 276726 have watched
books2[(books2['User.ID']==276729) | (books2['User.ID']==276726)]
```

Out[21]:

	User.ID	Book.Title	Book.Rating
0	276726	Classical Mythology	5
1	276729	Clara Callan	3
2	276729	Decision in Normandy	6

In [22]:
```
user_1=books2[(books2['User.ID']==276729)]
user_2=books2[(books2['User.ID']==276726)]
```

In [23]:
```
user_1['Book.Title']
```

Out[23]:
```
1            Clara Callan
2    Decision in Normandy
Name: Book.Title, dtype: object
```

In [24]:
```
user_2['Book.Title']
```

Out[24]:
```
0    Classical Mythology
Name: Book.Title, dtype: object
```

In [25]:
```
pd.merge(user_1,user_2,on='Book.Title',how='outer')
```

Out[25]:

	User.ID_x	Book.Title	Book.Rating_x	User.ID_y	Book.Rating_y
0	276729.0	Clara Callan	3.0	NaN	NaN
1	276729.0	Decision in Normandy	6.0	NaN	NaN
2	NaN	Classical Mythology	NaN	276726.0	5.0

In [ ]:
```

```