An Assignment Report

On

Recommender system

by

Aditya Sharma
AU18B1009

Under the guidance of

Ashish Bansal

School of Engineering

Avantika University, Ujjain

2020-2021

Content: **Page**

Problem statement:

Bookrent is the largest online and offline book rental chain in India. The company charges a fixed rental fee for a book per month. Lately, the company has been losing its user base. The main reason for this is that users are not able to choose the right books for themselves. The company wants to solve this problem and increase its revenue and profits. Different students will be required to work on different components of the broad problem statement and the relevant dataset will be made available to students.

1. Read the books dataset and explore it.

```
data_user = pd.read_csv('BX-Users.csv',nrows =10000,encoding='latin-1')

data_books = pd.read_csv('BX-Books.csv',nrows =10000,encoding='latin-1')

data_books_ratings = pd.read_csv('BX-Book-Ratings.csv',nrows =10000,encoding='latin-1')
```

```
data_user
```

	user_id	Location	Age
0	1	nyc, new york, usa	NaN
1	2	stockton, california, usa	18.0
2	3	moscow, yukon territory, russia	NaN
3	4	porto, v.n.gaia, portugal	17.0
4	5	farnborough, hants, united kingdom	NaN
...	...	...	...
9995	9996	reynella, south australia, australia	29.0
9996	9997	willisburg, kentucky, usa	56.0
9997	9998	warren, michigan, usa	NaN
9998	9999	beaverton, oregon, usa	NaN
9999	10000	jacksonville, florida, usa	38.0

10000 rows × 3 columns

```
data_books
```

	isbn	book_title	book_author	year_of_publication	publisher
0	195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press
1	2005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada
2	60973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial
3	374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux
4	393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton & Company
...	...	...	...	...	...
9995	140283404	Beloved (Penguin Great Books of the 20th Century)	Toni Morrison	2000	Penguin Books
9996	380730774	Read This and Tell Me What It Says : Stories (...	A. Manette Ansay	1998	William Morrow
9997	862418879	The Star Rover	Jack London	2000	Canongate Books
9998	340414645X	Die Keltennadel.	Patrick Dunne	2001	LÃ?Â¼bbe
9999	3442730988	Tod in der Datscha.	Anna Malyschewa	2003	btb

10000 rows × 5 columns

```
data_books_ratings
```

	user_id	isbn	rating
0	276725	034545104X	0
1	276726	155061224	5
2	276727	446520802	0
3	276729	052165615X	3
4	276729	521795028	6
...	...	...	...
9995	243	425164403	0
9996	243	440224764	0
9997	243	440225701	0
9998	243	440226430	0
9999	243	440234743	0

10000 rows × 3 columns

```
data_books.describe()
```

	isbn	book_title	book_author	year_of_publication	publisher
count	10000	10000	10000	10000	10000
unique	10000	9553	5754	63	1702
top	385495641	The Golden Compass (His Dark Materials, Book 1)	Stephen King	2002	Ballantine Books
freq	1	5	68	919	300

```
data_books.isnull().sum()
```

```
isbn                 0
book_title           0
book_author          0
year_of_publication  0
publisher            0
dtype: int64
```

```
data_books.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 5 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   isbn                 10000 non-null  object
 1   book_title           10000 non-null  object
 2   book_author          10000 non-null  object
 3   year_of_publication  10000 non-null  object
 4   publisher            10000 non-null  object
dtypes: object(5)
memory usage: 390.8+ KB
```

```
data_books.head()
```

	isbn	book_title	book_author	year_of_publication	publisher
0	195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press
1	2005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada
2	60973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial
3	374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux
4	393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton & Company

2. Clean up NaN values

```
data_books.isnull().sum()
```

```
isbn                 0
book_title           0
book_author          0
year_of_publication  0
publisher            0
dtype: int64
```

There aren't any null values associated with this data set.

3. Read the data where ratings are given by the users.

```
data_books_ratings
```

	user_id	isbn	rating
0	276725	034545104X	0
1	276726	155061224	5
2	276727	446520802	0
3	276729	052165615X	3
4	276729	521795028	6
...	...	...	...
9995	243	425164403	0
9996	243	440224764	0
9997	243	440225701	0
9998	243	440226430	0
9999	243	440234743	0

10000 rows × 3 columns

4. Take a quick look at the number of unique users and books.

```
data_book.nunique()
```

```
rating              11
user_id_index      465
book_id_index     1366
dtype: int64
```

5. Perform data conversion for consistency.

Here we have three csv files, and we need user ID and isbn for the model, so we will merge the Bx – Books csv and the Bx -book-ratings.

```
data_book = pd.merge(data_books_ratings,data_books,on='isbn')
```

```
data_book
```

	user_id	isbn	rating	book_title	book_author	year_of_publication	publisher
0	276725	034545104X	0	Flesh Tones: A Novel	M. J. Rose	2002	Ballantine Books
1	276744	038550120X	7	A Painted House	JOHN GRISHAM	2001	Doubleday
2	278418	038550120X	0	A Painted House	JOHN GRISHAM	2001	Doubleday
3	276746	425115801	0	Lightning	Dean R. Koontz	1996	Berkley Publishing Group
4	277427	425115801	0	Lightning	Dean R. Koontz	1996	Berkley Publishing Group
...	...	...	...	...	...	...	...
1824	243	380807866	0	The Elusive Flame	Kathleen E. Woodiwiss	1999	Avon
1825	243	385316895	6	Legacy of Silence	Belva Plain	1998	Bantam Dell Pub Group
1826	243	385509456	0	The Curious Incident of the Dog in the Night-T...	MARK HADDON	2003	Doubleday
1827	243	385720106	7	A Map of the World	Jane Hamilton	1999	Anchor Books/Doubleday
1828	243	425164403	0	Only Love (Magical Love)	Erich Segal	1998	Berkley Publishing Group

1829 rows × 7 columns

6. Split your data into two sets (Training and testing)

```
In [280]: from sklearn.model_selection import train_test_split

In [281]: data_book.columns

Out[281]: Index(['user_id', 'isbn', 'rating', 'book_title', 'book_author',
                 'year_of_publication', 'publisher'],
                dtype='object')

In [282]: n_users=data_book.user_id.unique().shape[0]

In [283]: n_books=data_book.isbn.unique().shape[0]

In [284]: train_data,test_data=train_test_split(data_book,test_size=0.25)

In [285]: train_data_matrix=np.zeros((n_users,n_books))

In [286]: train_data_matrix

Out[286]: array([[0., 0., 0., ..., 0., 0., 0.],
                 [0., 0., 0., ..., 0., 0., 0.],
                 [0., 0., 0., ..., 0., 0., 0.],
                 ...,
                 [0., 0., 0., ..., 0., 0., 0.],
                 [0., 0., 0., ..., 0., 0., 0.],
                 [0., 0., 0., ..., 0., 0., 0.]])
```

```
train_data
```

	user_id	isbn	rating	book_title	book_author	year_of_publication	publisher
448	277378	446322180	0	Name of the Rose-Nla	Umberto Eco	1984	Warner Books
1604	278633	515131083	10	Plantation: A Lowcountry Tale	Dorothea Benton Frank	2001	Jove Books
179	276964	886773741	7	Tailchaser's Song	Tad Williams	1994	Daw Books
1108	278144	399139087	0	Second Nature	Alice Hoffman	1994	Putnam Pub Group
872	277759	553211404	7	Jane Eyre (Bantam Classics)	Charlotte Bronte	1983	Bantam
...	...	...	...	...	...	...	...
1084	278137	440940001	8	Island of the Blue Dolphins (Laurel Leaf Books)	Scott O'Dell	1978	Laure Leaf
1466	278418	590224735	0	Kristy's Great Idea (The Baby-Sitter's Club #1)	Ann M. Martin	1995	Scholastic
184	277378	689817851	0	Go Ask Alice	Anonymous	1998	Simon Pulse
691	277478	425161242	0	Chromosome 6	Robin Cook	2000	Berkley Publishing Group
751	278418	64400557	0	Charlotte's Web (Trophy Newbery)	E. B. White	1974	HarperTrophy

7. Calculate the Similarity.

```
from sklearn.metrics import pairwise_distances
```

```
user_similarity=pairwise_distances(train_data_matrix,metric='cosine')
```

```
books_similarity=pairwise_distances(train_data_matrix.T,metric='cosine')
```

```
user_similarity
```

```
array([[0., 1., 1., ..., 1., 1., 1.],
       [1., 0., 1., ..., 1., 1., 1.],
       [1., 1., 0., ..., 1., 1., 1.],
       ...,
       [1., 1., 1., ..., 0., 1., 1.],
       [1., 1., 1., ..., 1., 0., 1.],
       [1., 1., 1., ..., 1., 1., 0.]])
```

```
books_similarity
```

```
array([[0., 1., 1., ..., 1., 1., 1.],
       [1., 0., 1., ..., 1., 1., 1.],
       [1., 1., 0., ..., 1., 1., 1.],
       ...,
       [1., 1., 1., ..., 0., 1., 1.],
       [1., 1., 1., ..., 1., 0., 1.],
       [1., 1., 1., ..., 1., 1., 0.]])
```

8. Use the evaluation metrics to make predictions.

```
def predict(ratings, similarity, type='user'):
    if type == 'user':
        mean_user_rating = ratings.mean(axis=1)
        #You use np.newaxis so that mean_user_rating has same format as ratings
        ratings_diff = (ratings - mean_user_rating[:, np.newaxis])
        pred = mean_user_rating[:, np.newaxis] + similarity.dot(ratings_diff) / np.array([np.abs(similarity).sum(axis=1)]).T
    elif type == 'item':
        pred = ratings.dot(similarity) / np.array([np.abs(similarity).sum(axis=1)])
    return pred
```

```
books_prediction = predict(train_data_matrix, books_similarity, type='item')
user_prediction = predict(train_data_matrix, user_similarity, type='user')
```

```
user_prediction
```

```
array([[ 0.01476897,  0.03202627,  0.01476897, ...,  0.01476897,
         0.01476897,  0.02986911],
       [-0.00651285,  0.01072853, -0.00651285, ..., -0.00651285,
        -0.00651285,  0.00857336],
       [ 0.14976785,  0.16575402,  0.14976785, ...,  0.14976785,
         0.14976785,  0.16487463],
       ...,
       [-0.00651285,  0.01072853, -0.00651285, ..., -0.00651285,
        -0.00651285,  0.00857336],
       [-0.00651285,  0.01072853, -0.00651285, ..., -0.00651285,
        -0.00651285,  0.00857336],
       [-0.00137735,  0.01586403, -0.00137735, ..., -0.00137735,
        -0.00137735, -0.00137735]])
```

```
books_prediction
```

```
array([[0.02124542, 0.02135077, 0.02124542, ..., 0.02124542, 0.02124542,
        0.02124542],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.15604396, 0.15241558, 0.15604396, ..., 0.15604396, 0.15604396,
        0.15604396],
       ...,
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.00512821, 0.00515363, 0.00512821, ..., 0.00512821, 0.00512821,
        0.        ]])
```