PRACTICUM PROBLEMS

2.1 Problem 1

Load the Movielens 100k dataset (ml-100k.zip) into Python using Pandas dataframes. Build a user profile on unscaled data for both users 200 and 15, and calculate the cosine similarity and distance between the user's preferences and the item/movie 95. Which user would a recommender system suggest this movie to?

Answer:

The recommender system will suggest movie 95 (Aladin) to user 200 based on the values of cosine distance obtained which are 0.405 and 0266 for users 200 and 15 respectively.

jupyter sam assignment4-1 Last Checkpoint: 3 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Help Python [Root]

Code CellToolbar

In [1]:
```
import numpy as np
import pandas as pd
from sklearn import metrics
```

In [2]:
```
user_cols = ['user_id',
             'age',
             'gender',
             'occupation',
             'zip_code']
users = pd.read_csv(r'E:\Fall16\DM\HW\hw4\ml-100k\u.user',
                    sep='|',
                    names=user_cols)
```

In [7]:
```
rating_cols = ['user_id',
               'movie_id',
               'rating',
               'timestamp']
ratings = pd.read_csv(r'E:\Fall16\DM\HW\hw4\ml-100k\u.data',
                      sep='\t',
                      names=rating_cols)

r=pd.DataFrame(pd.read_table(r'E:\Fall16\DM\HW\hw4\ml-100k\u.data',sep = '\t',names=rating_cols ))
```

In [8]:
```
item_cols = ['movie id',
             'movie title',
```

jupyter sam assignment4-1 Last Checkpoint: 3 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Help Python [Root]

Code CellToolbar

In [8]:
```
item_cols = ['movie id',
             'movie title',
             'release date',
             'video release date',
             'IMDb URL',
             'Unknown',
             'Action',
             'Adventure',
             'Animation',
             'Childrens',
             'Comedy',
             'Crime',
             'Documentary',
             'Drama',
             'Fantasy',
             'FilmNoir',
             'Horror',
             'Musical',
             'Mystery',
             'Romance',
             'SciFi',
             'Thriller',
             'War',
             'Western']
items = pd.read_csv(r'E:\Fall16\DM\HW\hw4\ml-100k\u.item',
                    sep='|',
                    names=item_cols,
                    encoding='latin-1')
i=pd.DataFrame(pd.read_table(r'E:\Fall16\DM\HW\hw4\ml-100k\u.item',sep = '|',names=item_cols,encoding='latin-1' ))
```

Ask me anything 8:46 PM 11/30/2016

jupyter sam assignment4-1 Last Checkpoint: 3 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Help Python [Root]

Code CellToolbar

In [9]:
```
user200 = r.loc[r['user_id'].isin([200])]
v=user200['movie_id']
v200 =pd.DataFrame(data=v)
#for row in  i,v1
```

In [22]:
```
v2= i.loc[i['movie id'].isin(v200['movie_id'])]
```

In [23]:
```
v3 = v2.iloc[:,5:24]
mean3 = v3.mean(axis=0)
```

In [24]:
```
user2 = r.loc[r['user_id'].isin([15])]
v4=user2['movie_id']
v4 =pd.DataFrame(data=v4)
v4
v5= i.loc[i['movie id'].isin(v4['movie_id'])]
v5
v5 = v5.iloc[:,5:24]
v5.head()
mean5 = v5.mean(axis=0)
```

In [25]:
```
item2 = items[94:95]
feat2 = item2.iloc[:,5:24]
feat2.head()
mean2 = feat2.mean()
```

In [26]:
```
mean200 = mean3.reshape(1,-1)
```

jupyter sam assignment4-1 Last Checkpoint: 3 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Help | Python [Root]

Code | CellToolbar

In [25]:
```
item2 = items[94:95]
feat2 = item2.iloc[:,5:24]
feat2.head()
mean2 = feat2.mean()
```

In [26]:
```
mean200 = mean3.reshape(1,-1)
mean200
```

Out[26]:
```
array([[ 0.        ,  0.36574074,  0.26388889,  0.07407407,  0.19444444,
         0.24074074,  0.0462963 ,  0.00462963,  0.28240741,  0.0462963 ,
         0.00925926,  0.0462963 ,  0.0787037 ,  0.01851852,  0.16666667,
         0.19907407,  0.23148148,  0.0787037 ,  0.01851852]])
```

In [27]:
```
mean15 = mean5.reshape(1,-1)
mean15
```

Out[27]:
```
array([[ 0.        ,  0.21153846,  0.11538462,  0.01923077,  0.06730769,
         0.26923077,  0.05769231,  0.        ,  0.46153846,  0.02884615,
         0.01923077,  0.01923077,  0.01923077,  0.06730769,  0.24038462,
         0.11538462,  0.24038462,  0.09615385,  0.        ]])
```

In [28]:
```
mean95 = mean2.reshape(1,-1)
mean95
```

Out[28]:
```
array([[ 0.,  0.,  0.,  1.,  1.,  1.,  0.,  0.,  0.,  0.,  0.,  0.,  1.,
         0.,  0.,  0.,  0.,  0.,  0.]])
```

In [29]:
```
result = metrics.pairwise.cosine_similarity(mean200,mean95)
result
```

localhost:8888/notebooks/sam%20assignment4-1.ipynb

jupyter sam assignment4-1 Last Checkpoint: 3 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Help | Python [Root]

Code | CellToolbar

In [27]:
```
mean15 = mean5.reshape(1,-1)
mean15
```

Out[27]:
```
array([[ 0.        ,  0.21153846,  0.11538462,  0.01923077,  0.06730769,
         0.26923077,  0.05769231,  0.        ,  0.46153846,  0.02884615,
         0.01923077,  0.01923077,  0.01923077,  0.06730769,  0.24038462,
         0.11538462,  0.24038462,  0.09615385,  0.        ]])
```

In [28]:
```
mean95 = mean2.reshape(1,-1)
mean95
```

Out[28]:
```
array([[ 0.,  0.,  0.,  1.,  1.,  1.,  0.,  0.,  0.,  0.,  0.,  0.,  1.,
         0.,  0.,  0.,  0.,  0.,  0.]])
```

In [29]:
```
result = metrics.pairwise.cosine_similarity(mean200,mean95)
result
```

Out[29]:
```
array([[ 0.40572802]])
```

In [30]:
```
result2 = metrics.pairwise.cosine_similarity(mean15,mean95)
result2
```

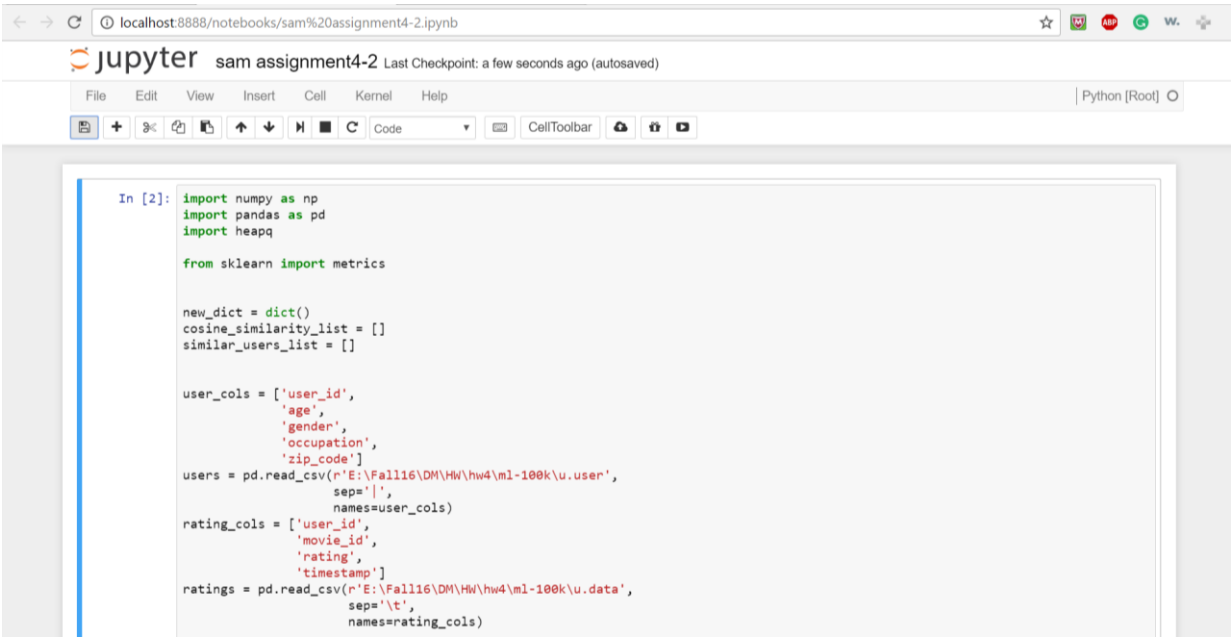
Out[30]:
```
array([[ 0.26612637]])
```

2.2 Problem 2

Load the Movielens 100k dataset (ml-100k.zip) into Python using Pandas dataframes. Convert the ratings data into a utility matrix representation, and find the 10 most similar users for user 1 based on cosine similarity of the user ratings data. Based on the average of of the ratings for item 508 from the similar users, what is the expected rating for this item for user 1?

Answer:

The expected rating for item 508 by user 1 will be 0.2689 ~ 0.27



localhost:8888/notebooks/sam%20assignment4-2.ipynb

jupyter sam assignment4-2 Last Checkpoint: a few seconds ago (autosaved)

File Edit View Insert Cell Kernel Help | Python [Root]

Code CellToolbar

```
ratings = pd.read_csv(r'E:\Fall16\DM\HW\hw4\ml-100k\u.data',
                      sep='\t',
                      names=rating_cols)

item_cols = ['movie id',
             'movie title',
             'release date',
             'video release date',
             'IMDb URL',
             'Unknown',
             'Action',
             'Adventure',
             'Animation',
             'Childrens',
             'Comedy',
             'Crime',
             'Documentary',
             'Drama',
             'Fantasy',
             'FilmNoir',
             'Horror',
             'Musical',
             'Mystery',
             'Romance',
             'SciFi',
             'Thriller',
             'War',
             'Western']
items = pd.read_csv(r'E:\Fall16\DM\HW\hw4\ml-100k\u.item',
                    sep='|',
```

Home | sam assignment4-2 | sam assignment4-1

localhost:8888/notebooks/sam%20assignment4-2.ipynb

jupyter sam assignment4-2 Last Checkpoint: a few seconds ago (autosaved)

File Edit View Insert Cell Kernel Help | Python [Root]

Code CellToolbar

```
items = pd.read_csv(r'E:\Fall16\DM\HW\hw4\ml-100k\u.item',
                    sep='|',
                    names=item_cols,
                    encoding='latin-1')

utility = ratings.pivot(index='user_id',
                        columns='movie_id',
                        values='rating')

user_means = utility.mean(axis=1)

utility_centered = utility - user_means
utility_centered = utility_centered.where((pd.notnull(utility_centered)), 0)

print("Utility")

print(utility_centered)
```

```
935     -0.610294  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
936      0.389706  0.000000  1.203704  0.000000  0.000000  1.364929  0.034739
937      0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
938      0.389706  0.000000  0.000000  0.000000  0.000000  0.000000  0.034739
939      0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
940      0.000000  0.000000  0.000000 -2.333333  0.000000  0.000000  0.034739
941      1.389706  0.000000  0.000000  0.000000  0.000000  0.000000  0.034739
942      0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
943      0.000000  1.290323  0.000000  0.000000  0.000000  0.000000  0.000000

           8         9         10    ...   1673  1674  1675  1676  1677  \
```

jupyter sam assignment4-2 Last Checkpoint: a minute ago (autosaved)

File Edit View Insert Cell Kernel Help Python [Root]

Code CellToolbar

```
29       0.00000  0.000000  0.000000  ...    0.0   0.0   0.0   0.0   0.0
30       0.00000  0.000000  0.000000  ...    0.0   0.0   0.0   0.0   0.0
...          ...       ...       ...  ...    ...   ...   ...   ...   ...
914      0.00000  0.000000  0.000000  ...    0.0   0.0   0.0   0.0   0.0
915      0.00000  0.000000  0.000000  ...    0.0   0.0   0.0   0.0   0.0
916      0.00000  0.727273  0.000000  ...    0.0   0.0   0.0   0.0   0.0
917      0.00000  0.727273  0.000000  ...    0.0   0.0   0.0   0.0   0.0
918      0.00000  0.000000  0.000000  ...    0.0   0.0   0.0   0.0   0.0
```

In [3]:

```
user_1 = utility_centered[0:1]

for index, row in utility_centered.iterrows():
    np_row = np.array(row)
    np_row.resize(1, 1682)
    if index != 1:
        list = metrics.pairwise.cosine_similarity(user_1, np_row)
        cosine_similarity_list.append(list)
        new_dict.__setitem__(index, list)

similar_users_top_ten = heapq.nlargest(10, new_dict, key=new_dict.get)
user_columns = ['user_id' , '508']

data =0
for i in similar_users_top_ten:

        print("Item: ", i); print("Cosine Similarity ", new_dict.get(i));
        similar_users_list.append(utility_centered.loc[i, 508])
```

```
Item:  738
Cosine Similarity  [[ 0.29148679]]
```

jupyter sam assignment4-2 Last Checkpoint: a minute ago (autosaved)

File Edit View Insert Cell Kernel Help Python [Root]

Code CellToolbar

```
data =0
for i in similar_users_top_ten:

        print("Item: ", i); print("Cosine Similarity ", new_dict.get(i));
        similar_users_list.append(utility_centered.loc[i, 508])
```

```
Item:  738
Cosine Similarity  [[ 0.29148679]]
Item:  592
Cosine Similarity  [[ 0.27840172]]
Item:  276
Cosine Similarity  [[ 0.26815054]]
Item:  267
Cosine Similarity  [[ 0.26476147]]
Item:  643
Cosine Similarity  [[ 0.2640026]]
Item:  757
Cosine Similarity  [[ 0.26236785]]
Item:  457
Cosine Similarity  [[ 0.26233704]]
Item:  606
Cosine Similarity  [[ 0.26084701]]
Item:  916
Cosine Similarity  [[ 0.25562438]]
Item:  44
Cosine Similarity  [[ 0.2529544]]
```

In [4]:

```
np_array_similar_users = np.array(similar_users_list)

print("Expected rating of user 1 for item 508:")
print(np_array_similar_users.mean())
```

jupyter sam assignment4-2 Last Checkpoint: a minute ago (autosaved)

File Edit View Insert Cell Kernel Help Python [Root]

Code CellToolbar

```
Item:  643
Cosine Similarity  [[ 0.2640026]]
Item:  757
Cosine Similarity  [[ 0.26236785]]
Item:  457
Cosine Similarity  [[ 0.26233704]]
Item:  606
Cosine Similarity  [[ 0.26084701]]
Item:  916
Cosine Similarity  [[ 0.25562438]]
Item:  44
Cosine Similarity  [[ 0.2529544]]
```

In [4]:

```
np_array_similar_users = np.array(similar_users_list)

print("Expected rating of user 1 for item 508:")
print(np_array_similar_users.mean())
```

```
Expected rating of user 1 for item 508:
0.268965517241
```

In [ ]: