data_cleaning

February 28, 2020

0.1 Initial Data Cleaning

In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
```

In [25]:
```
# load data
movies = pd.read_csv('data/movies.txt',delimiter="\t", header=None)
movies.columns = ['ID', 'Title', 'Unknown', 'Action', 'Adventure', 'Animation', 'Chil

data = pd.read_csv('data/data.txt',delimiter="\t", header=None)
data.columns = ['User', 'Movie', 'Rating']

train = np.loadtxt('data/train.txt')
test = np.loadtxt('data/test.txt')
```

In [28]:
```
# find movies with no ratings (we see that none exist)
for i in range(1, len(movies['ID'] + 1)):
    if i not in data['Movie']:
        print(movies[movies['ID']==i]['Title'])
```

In [4]:
```
# find movies with duplicate titles but different IDs and remove
# duplicate rows
titles = []
row = 0
drop = []
dups = {}

for title in movies['Title']:
    if title in titles:
        print(title)
        drop.append(row)
        dups[titles.index(title) + 1] = row + 1

    titles.append(title)
    row += 1

movies.drop(drop, inplace = True)
```

```
Chasing Amy (1997)
Ulee's Gold (1997)
Desperate Measures (1998)
Fly Away Home (1996)
Body Snatchers (1993)
Kull the Conqueror (1997)
Ice Storm, The (1997)
Money Talks (1997)
That Darn Cat! (1997)
Designated Mourner, The (1997)
Deceiver (1997)
Hurricane Streets (1998)
Hugo Pool (1997)
Nightwatch (1997)
Butcher Boy, The (1998)
Chairman of the Board (1998)
Substance of Fire, The (1996)
Sliding Doors (1998)
```

```
In [5]: # reindex movies
        movies = movies.reset_index(drop=True)
        index_map = {}
        count = 0
        for i in movies['ID']:
            index_map[i] = count+1 # keep 1-indexing convention
            movies['ID'][movies['ID'] == i] = count+1
            count += 1
```

```
C:\Users\amanda\Anaconda\lib\site-packages\ipykernel_launcher.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/i
  import sys
```

```
In [6]: # replace duplicate IDs in data, train, test and reindex
        for duplicate in dups:
            data['Movie'].replace(dups[duplicate], duplicate, inplace=True)
            train[:,1][train[:,1] == dups[duplicate]] = duplicate
            test[:,1][test[:,1] == dups[duplicate]] = duplicate

        for i in range(len(data)):
            data['Movie'][i] = index_map[data['Movie'][i]]
        for i in range(len(train)):
            train[:,1][i] = index_map[train[:,1][i]]
        for i in range(len(test)):
            test[:,1][i] = index_map[test[:,1][i]]
```

In [7]:
```
# find number of ratings and average rating for each movie
movies['num_ratings'] = [0]*len(movies)
movies['tot_rating'] = [0]*len(movies)
movies['avg_rating'] = [0.0]*len(movies)

# get number of ratings
for i in range(len(data)):
    movie = data['Movie'][i] - 1 # IDs are 1-indexed
    movies.loc[movie, 'num_ratings'] += 1
    movies.loc[movie, 'tot_rating'] += data['Rating'][i]

for i in range(len(movies)):
    if movies['num_ratings'][i] != 0:
        movies['avg_rating'][i]= movies['tot_rating'][i] / movies['num_ratings'][i]
```

```
C:\Users\amanda\Anaconda\lib\site-packages\ipykernel_launcher.py:14: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/i
```

In [23]:
```
# save cleaned movie and data files
movies.to_csv('data/movies.csv', index=False)
data.to_csv('data/data.csv', index=False)
data.to_csv('data/data2.txt', header=False, index=False, sep='\t')
```

In [24]:
```
# save cleaned train and test files
np.savetxt('data/train2.txt', train)
np.savetxt('data/test2.txt', test)
```