Mini_Project_IMDb

January 30, 2019

Mini Project
IMDb Data Set https://grouplens.org/datasets/movielens/ ml-20m.zip

In [1]:
```
import pandas as pd
import csv
```

In [2]:
```
# Contents of Data Set Zip

!ls ./movielens
```

```
Icon?             genome-scores.csv links.csv         ratings.csv
README.txt        genome-tags.csv   movies.csv        tags.csv
```

In [3]:
```
#movies = pd.read_csv('./movielens/movies.csv', sep=',')
#print(type(movies))
#movies.head(5)
```

In [4]:
```
movies = csv.reader(open('./movielens/movies.csv','r'))

years = []
year_1995 = []

for line in movies:
    line = ''.join(line)
    #print(line)

    if '(1900)' in line:
        year_1995.append(line)
'''
for line in movies:
    if '1995' in line:
        year_1995.append(line)
'''
```

Out[4]:
```
"\nfor line in movies:\n    if '1995' in line:\n        year_1995.append(line)\n"
```

In [5]:
```
print(year_1995)
```

```
['117909The Kiss (1900)Romance']
```

In [6]:
```
print(len(year_1995))
```

```
1
```

In [14]:
```
movies = csv.reader(open('./movielens/movies.csv','r'))

        years = []



        for line in movies:
            line = ''.join(line)
            #print(line)

            for i in range(1800, 2020):
                year_parentheses = '(' + str(i) + ')'
                year_list_name = 'years_' + str(i)

                if year_parentheses in line:
                    years.append(line)
                    #print(i)



        print(len(years))
        #print(years)
```

```
27256
```

In [8]:
```
# Discrepency in 27,256 row being read versus 27,279 in the original is probably due t

        with open('./movielens/movies.csv',"r") as f:
            reader = csv.reader(f,delimiter = ",")
            data = list(reader)
            row_count = len(data)
            print(row_count)
```

```
27279
```

In [9]:
```
# Need to figure out a way to read each line into appropriate dictionary key value


        movies = csv.reader(open('./movielens/movies.csv','r'))
```

```
        years = []



        for line in movies:
            line = ''.join(line)  # converts lines to String
            #print(line)

            for i in range(1800, 2020):
                year_parentheses = '(' + str(i) + ')'
                year_list_name = 'years_' + str(i)

                if year_parentheses in line:
                    years.append(line)
                    #print(i)




        x = range(1800,2020)
        dct = {}
        for i in x:
            dct['year_%s' % i] = []

        print(dct)
```

```
{'year_1800': [], 'year_1801': [], 'year_1802': [], 'year_1803': [], 'year_1804': [], 'year_180
```

```
In [10]: movies = csv.reader(open('./movielens/movies.csv','r'))
```

```
         dct = {}

         for line in movies:
             line = ''.join(line)  # converts lines to String

             #i = 1800

             for i in range(1800, 2020):
                 year_parentheses = '(' + str(i) + ')'

                 if year_parentheses in line:
                     dct['year_%s' % i] = [line]
                 i = i + 1

         print(len(dct))
```

118

As a dictionary

```
In [11]: movies = csv.reader(open('./movielens/movies.csv','r'))

         years = {}

         for line in movies:
             line = ''.join(line)

             year = line[line.find("(") + 1 : line.find(")") ]

             if year.isdigit():
                 #print(year + "|" + line)     # List all the years from each line
                 years[year] = [line]



         print(years)
```

```
{'1995': ['131128Flodder 3 (1995)Comedy'], '1994': ['131142Voll Normaaal (1994)Comedy'], '1996
```

```
In [13]: from collections import Counter
         import matplotlib.pyplot as plt
         import numpy as np
         %matplotlib inline


         movies = csv.reader(open('./movielens/movies.csv','r'))



         years = []

         for line in movies:
             line = ''.join(line)

             year = line[line.find("(") + 1 : line.find(")") ]

             if year.isdigit():
                 #print(year + "|" + line)     # List all the years from each line
                 years.append(int(year))

         print(len(years)) # 22,059 is the number of lines with years included
```

```
C = Counter(years)
year_list_of_lists = [[k,]*v for k,v in C.items()]
print(len(year_list_of_lists)) # 119 unique years
#print(year_list_of_lists)



unique_years = [item[0] for item in year_list_of_lists]
#print(len(unique_years))


sorted_years = sorted(unique_years)
#print(sorted_years)


x_axis = []
y_axis = []
graph_pairs = []
r = 0
while r < 119:
    year_num = year_list_of_lists[0 + r][0]
    entries_num = len(year_list_of_lists[0 + r])

    x_axis.append(year_num)
    y_axis.append(entries_num)

    plt.bar(x_axis, y_axis, label='Movie Entries by Year') # Create Bar Graph
    ax = plt.gca()
    ax.set_xlim([1890, 2020])

    ### Change the size of the plot and save it to folder as PNG
    fig = plt.gcf()
    fig.set_size_inches(18.5, 10.5)
    fig.savefig('test2019.png', dpi=100)



    combo = str(year_num) + "|" + str(entries_num)
    #print(combo)

    graph_pairs.append(year_num | entries_num)

    r = r + 1

    plt.xlabel('Year')
    plt.ylabel('Number of Films')
```

```
        plt.title('Movie Entries by Year')

    #print(graph)

    graph_pairs.sort()
    print(graph_pairs[0:-3])
```

```
22059
119
[7, 69, 501, 1891, 1893, 1894, 1895, 1898, 1901, 1901, 1902, 1903, 1905, 1910, 1911, 1913, 1915
```