SimilarMovies

December 20, 2016

1 Finding Similar Movies

We'll start by loading up the MovieLens dataset. Using Pandas, we can very quickly load the rows of the u.data and u.item files that we care about, and merge them together so we can work with movie names instead of ID's. (In a real production job, you'd stick with ID's and worry about the names at the display layer to make things more efficient. But this lets us understand what's going on better for now.)

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

        r_cols = ['user_id', 'movie_id', 'rating']
        ratings = pd.read_csv('e:/sundog-consult/udemy/datascience/ml-100k/u.data', sep='\t', na

        m_cols = ['movie_id', 'title']
        movies = pd.read_csv('e:/sundog-consult/udemy/datascience/ml-100k/u.item', sep='|', name

        ratings = pd.merge(movies, ratings)

In [2]: ratings.head()

Out[2]:    movie_id             title  user_id  rating
        0         1  Toy Story (1995)      308       4
        1         1  Toy Story (1995)      287       5
        2         1  Toy Story (1995)      148       4
        3         1  Toy Story (1995)      280       4
        4         1  Toy Story (1995)       66       3
```

Now the amazing pivot_table function on a DataFrame will construct a user / movie rating matrix. Note how NaN indicates missing data - movies that specific users didn't rate.

```
In [3]: movieRatings = ratings.pivot_table(index=['user_id'],columns=['title'],values='rating')
        movieRatings.head()

Out[3]: title    'Til There Was You (1997)  1-900 (1994)  101 Dalmatians (1996)  \
        user_id
        0                              NaN           NaN                    NaN
        1                              NaN           NaN                    2.0
        2                              NaN           NaN                    NaN
```

```
3                              NaN           NaN                     NaN
4                              NaN           NaN                     NaN

title    12 Angry Men (1957)  187 (1997)  2 Days in the Valley (1996)  \
user_id
0                        NaN         NaN                          NaN
1                        5.0         NaN                          NaN
2                        NaN         NaN                          NaN
3                        NaN         2.0                          NaN
4                        NaN         NaN                          NaN

title    20,000 Leagues Under the Sea (1954)  2001: A Space Odyssey (1968)  \
user_id
0                                        NaN                           NaN
1                                        3.0                           4.0
2                                        NaN                           NaN
3                                        NaN                           NaN
4                                        NaN                           NaN

title    3 Ninjas: High Noon At Mega Mountain (1998)  39 Steps, The (1935)  \
user_id
0                                              NaN                   NaN
1                                              NaN                   NaN
2                                              1.0                   NaN
3                                              NaN                   NaN
4                                              NaN                   NaN

title                    ...                  Yankee Zulu (1994)  \
user_id                  ...
0                        ...                                 NaN
1                        ...                                 NaN
2                        ...                                 NaN
3                        ...                                 NaN
4                        ...                                 NaN

title    Year of the Horse (1997)  You So Crazy (1994)  \
user_id
0                             NaN                  NaN
1                             NaN                  NaN
2                             NaN                  NaN
3                             NaN                  NaN
4                             NaN                  NaN

title    Young Frankenstein (1974)  Young Guns (1988)  Young Guns II (1990)  \
user_id
0                              NaN                NaN                   NaN
1                              5.0                3.0                   NaN
2                              NaN                NaN                   NaN
```

```
3                                NaN                 NaN                     NaN
4                                NaN                 NaN                     NaN

title    Young Poisoner's Handbook, The (1995)  Zeus and Roxanne (1997)  \
user_id
0                                          NaN                      NaN
1                                          NaN                      NaN
2                                          NaN                      NaN
3                                          NaN                      NaN
4                                          NaN                      NaN

title    unknown  Á köldum klaka (Cold Fever) (1994)
user_id
0            NaN                                 NaN
1            4.0                                 NaN
2            NaN                                 NaN
3            NaN                                 NaN
4            NaN                                 NaN

[5 rows x 1664 columns]
```

Let's extract a Series of users who rated Star Wars:

```
In [4]: starWarsRatings = movieRatings['Star Wars (1977)']
        starWarsRatings.head()

Out[4]: user_id
        0    5.0
        1    5.0
        2    5.0
        3    NaN
        4    5.0
        Name: Star Wars (1977), dtype: float64
```

Pandas' corrwith function makes it really easy to compute the pairwise correlation of Star Wars' vector of user rating with every other movie! After that, we'll drop any results that have no data, and construct a new DataFrame of movies and their correlation score (similarity) to Star Wars:

```
In [5]: similarMovies = movieRatings.corrwith(starWarsRatings)
        similarMovies = similarMovies.dropna()
        df = pd.DataFrame(similarMovies)
        df.head(10)

C:\Users\Frank\AppData\Local\Enthought\Canopy\User\lib\site-packages\numpy\lib\function_base.py:
  warnings.warn("Degrees of freedom <= 0 for slice", RuntimeWarning)


Out[5]:                                          0
        title
```

```
'Til There Was You (1997)            0.872872
1-900 (1994)                        -0.645497
101 Dalmatians (1996)                0.211132
12 Angry Men (1957)                  0.184289
187 (1997)                           0.027398
2 Days in the Valley (1996)          0.066654
20,000 Leagues Under the Sea (1954)  0.289768
2001: A Space Odyssey (1968)         0.230884
39 Steps, The (1935)                 0.106453
8 1/2 (1963)                        -0.142977
```

(That warning is safe to ignore.) Let's sort the results by similarity score, and we should have the movies most similar to Star Wars! Except... we don't. These results make no sense at all! This is why it's important to know your data - clearly we missed something important.

```
In [6]: similarMovies.sort_values(ascending=False)
```

```
Out[6]: title
        Star Wars (1977)                                                          1.0
        Man of the Year (1995)                                                    1.0
        Full Speed (1996)                                                         1.0
        Mondo (1996)                                                              1.0
        Line King: Al Hirschfeld, The (1996)                                      1.0
        Outlaw, The (1943)                                                        1.0
        Hurricane Streets (1998)                                                  1.0
        Hollow Reed (1996)                                                        1.0
        Scarlet Letter, The (1926)                                                1.0
        Safe Passage (1994)                                                       1.0
        Good Man in Africa, A (1994)                                              1.0
        Golden Earrings (1947)                                                    1.0
        Old Lady Who Walked in the Sea, The (Vieille qui marchait dans la mer, La) (1991)    1.0
        No Escape (1994)                                                          1.0
        Ed's Next Move (1996)                                                     1.0
        Stripes (1981)                                                            1.0
        Cosi (1996)                                                               1.0
        Commandments (1997)                                                       1.0
        Twisted (1996)                                                            1.0
        Beans of Egypt, Maine, The (1994)                                         1.0
        Last Time I Saw Paris, The (1954)                                         1.0
        Maya Lin: A Strong Clear Vision (1994)                                    1.0
        Designated Mourner, The (1997)                                            0.9
        Albino Alligator (1996)                                                   0.9
        Angel Baby (1995)                                                         0.9
        Prisoner of the Mountains (Kavkazsky Plennik) (1996)                      0.9
        Love in the Afternoon (1957)                                              0.9
        'Til There Was You (1997)                                                 0.8
        A Chef in Love (1996)                                                     0.8
        Quiet Room, The (1996)                                                    0.8
```

```
Collectionneuse, La (1967)                                    -1.0
Bewegte Mann, Der (1994)                                      -1.0
Lamerica (1994)                                               -1.0
Frankie Starlight (1995)                                      -1.0
To Have, or Not (1995)                                        -1.0
Legal Deceit (1997)                                           -1.0
Slingshot, The (1993)                                         -1.0
Swept from the Sea (1997)                                     -1.0
For Ever Mozart (1996)                                        -1.0
Love and Death on Long Island (1997)                          -1.0
Glass Shield, The (1994)                                      -1.0
Squeeze (1996)                                                -1.0
Crossfire (1947)                                              -1.0
Neon Bible, The (1995)                                        -1.0
American Dream (1990)                                         -1.0
Theodore Rex (1995)                                           -1.0
Horse Whisperer, The (1998)                                   -1.0
Lover's Knot (1996)                                           -1.0
S.F.W. (1994)                                                 -1.0
Fille seule, La (A Single Girl) (1995)                        -1.0
Sliding Doors (1998)                                          -1.0
Nightwatch (1997)                                             -1.0
Show, The (1995)                                              -1.0
Nil By Mouth (1997)                                           -1.0
Fall (1997)                                                   -1.0
I Like It Like That (1994)                                    -1.0
Sudden Manhattan (1996)                                       -1.0
Salut cousin! (1996)                                          -1.0
Tough and Deadly (1995)                                       -1.0
Dream Man (1995)                                              -1.0
dtype: float64
```

Our results are probably getting messed up by movies that have only been viewed by a handful of people who also happened to like Star Wars. So we need to get rid of movies that were only watched by a few people that are producing spurious results. Let's construct a new DataFrame that counts up how many ratings exist for each movie, and also the average rating while we're at it - that could also come in handy later.

In [7]:
```
import numpy as np
movieStats = ratings.groupby('title').agg({'rating': [np.size, np.mean]})
movieStats.head()
```

Out[7]:

	rating	
	size	mean
title		
'Til There Was You (1997)	9	2.333333
1-900 (1994)	5	2.600000

```
101 Dalmatians (1996)        109  2.908257
12 Angry Men (1957)          125  4.344000
187 (1997)                    41  3.024390
```

Let's get rid of any movies rated by fewer than 100 people, and check the top-rated ones that are left:

```
In [8]: popularMovies = movieStats['rating']['size'] >= 100
        movieStats[popularMovies].sort_values([('rating', 'mean')], ascending=False)[:15]
```

Out[8]:

	rating	
	size	mean
title		
Close Shave, A (1995)	112	4.491071
Schindler's List (1993)	298	4.466443
Wrong Trousers, The (1993)	118	4.466102
Casablanca (1942)	243	4.456790
Shawshank Redemption, The (1994)	283	4.445230
Rear Window (1954)	209	4.387560
Usual Suspects, The (1995)	267	4.385768
Star Wars (1977)	584	4.359589
12 Angry Men (1957)	125	4.344000
Citizen Kane (1941)	198	4.292929
To Kill a Mockingbird (1962)	219	4.292237
One Flew Over the Cuckoo's Nest (1975)	264	4.291667
Silence of the Lambs, The (1991)	390	4.289744
North by Northwest (1959)	179	4.284916
Godfather, The (1972)	413	4.283293

100 might still be too low, but these results look pretty good as far as "well rated movies that people have heard of." Let's join this data with our original set of similar movies to Star Wars:

```
In [9]: df = movieStats[popularMovies].join(pd.DataFrame(similarMovies, columns=['similarity']))

In [10]: df.head()
```

Out[10]:

title	(rating, size)	(rating, mean)	similarity
101 Dalmatians (1996)	109	2.908257	0.211132
12 Angry Men (1957)	125	4.344000	0.184289
2001: A Space Odyssey (1968)	259	3.969112	0.230884
Absolute Power (1997)	127	3.370079	0.085440
Abyss, The (1989)	151	3.589404	0.203709

And, sort these new results by similarity score. That's more like it!

```
In [11]: df.sort_values(['similarity'], ascending=False)[:15]
```

Out[11]:

title	(rating, size)	\
Star Wars (1977)	584	
Empire Strikes Back, The (1980)	368	
Return of the Jedi (1983)	507	
Raiders of the Lost Ark (1981)	420	
Austin Powers: International Man of Mystery (1997)	130	
Sting, The (1973)	241	
Indiana Jones and the Last Crusade (1989)	331	
Pinocchio (1940)	101	
Frighteners, The (1996)	115	
L.A. Confidential (1997)	297	
Wag the Dog (1997)	137	
Dumbo (1941)	123	
Bridge on the River Kwai, The (1957)	165	
Philadelphia Story, The (1940)	104	
Miracle on 34th Street (1994)	101	

title	(rating, mean)	similarity
Star Wars (1977)	4.359589	1.000000
Empire Strikes Back, The (1980)	4.206522	0.748353
Return of the Jedi (1983)	4.007890	0.672556
Raiders of the Lost Ark (1981)	4.252381	0.536117
Austin Powers: International Man of Mystery (1997)	3.246154	0.377433
Sting, The (1973)	4.058091	0.367538
Indiana Jones and the Last Crusade (1989)	3.930514	0.350107
Pinocchio (1940)	3.673267	0.347868
Frighteners, The (1996)	3.234783	0.332729
L.A. Confidential (1997)	4.161616	0.319065
Wag the Dog (1997)	3.510949	0.318645
Dumbo (1941)	3.495935	0.317656
Bridge on the River Kwai, The (1957)	4.175758	0.316580
Philadelphia Story, The (1940)	4.115385	0.314272
Miracle on 34th Street (1994)	3.722772	0.310921

Ideally we'd also filter out the movie we started from - of course Star Wars is 100% similar to itself. But otherwise these results aren't bad.

1.1 Activity

100 was an arbitrarily chosen cutoff. Try different values - what effect does it have on the end results?

In [ ]: