ItemBasedCF

December 20, 2016

1 Item-Based Collaborative Filtering

As before, we'll start by importing the MovieLens 100K data set into a pandas DataFrame:

```
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)

        ratings.head()
```

```
Out[1]:    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 we'll pivot this table to construct a nice matrix of users and the movies they rated. NaN indicates missing data, or movies that a given user did not watch:

```
In [2]: userRatings = ratings.pivot_table(index=['user_id'],columns=['title'],values='rating')
        userRatings.head()
```

```
Out[2]: 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]
```

Now the magic happens - pandas has a built-in corr() method that will compute a correlation score for every column pair in the matrix! This gives us a correlation score between every pair of movies (where at least one user rated both movies - otherwise NaN's will show up.) That's amazing!

```
In [3]: corrMatrix = userRatings.corr()
        corrMatrix.head()
```

```
Out[3]: title                     'Til There Was You (1997)  1-900 (1994)  \
        title
        'Til There Was You (1997)                        1.0           NaN
        1-900 (1994)                                     NaN           1.0
        101 Dalmatians (1996)                           -1.0           NaN
        12 Angry Men (1957)                             -0.5           NaN
        187 (1997)                                      -0.5           NaN

        title                     101 Dalmatians (1996)  12 Angry Men (1957)  \
        title
        'Til There Was You (1997)              -1.000000            -0.500000
        1-900 (1994)                                 NaN                  NaN
        101 Dalmatians (1996)                   1.000000            -0.049890
        12 Angry Men (1957)                    -0.049890             1.000000
        187 (1997)                              0.269191             0.666667

        title                     187 (1997)  2 Days in the Valley (1996)  \
        title
        'Til There Was You (1997)  -0.500000                     0.522233
        1-900 (1994)                     NaN                          NaN
        101 Dalmatians (1996)       0.269191                     0.048973
        12 Angry Men (1957)         0.666667                     0.256625
        187 (1997)                  1.000000                     0.596644
```

```
title                      20,000 Leagues Under the Sea (1954)  \
title
'Til There Was You (1997)                                  NaN
1-900 (1994)                                               NaN
101 Dalmatians (1996)                                 0.266928
12 Angry Men (1957)                                   0.274772
187 (1997)                                                 NaN

title                      2001: A Space Odyssey (1968)  \
title
'Til There Was You (1997)                          -0.426401
1-900 (1994)                                       -0.981981
101 Dalmatians (1996)                              -0.043407
12 Angry Men (1957)                                 0.178848
187 (1997)                                         -0.554700

title                      3 Ninjas: High Noon At Mega Mountain (1998)  \
title
'Til There Was You (1997)                                            NaN
1-900 (1994)                                                         NaN
101 Dalmatians (1996)                                                NaN
12 Angry Men (1957)                                                  NaN
187 (1997)                                                           NaN

title                      39 Steps, The (1935)  \
title
'Til There Was You (1997)                   NaN
1-900 (1994)                                NaN
101 Dalmatians (1996)                  0.111111
12 Angry Men (1957)                    0.457176
187 (1997)                             1.000000

title                                      ...                  \
title                                      ...
'Til There Was You (1997)                  ...
1-900 (1994)                               ...
101 Dalmatians (1996)                      ...
12 Angry Men (1957)                        ...
187 (1997)                                 ...

title                      Yankee Zulu (1994)  Year of the Horse (1997)  \
title
'Til There Was You (1997)                 NaN                       NaN
1-900 (1994)                              NaN                       NaN
101 Dalmatians (1996)                     NaN                 -1.000000
12 Angry Men (1957)                       NaN                       NaN
187 (1997)                                NaN                  0.866025
```

```
title                    You So Crazy (1994)  Young Frankenstein (1974)  \
title
'Til There Was You (1997)                  NaN                        NaN
1-900 (1994)                               NaN                  -0.944911
101 Dalmatians (1996)                      NaN                   0.158840
12 Angry Men (1957)                        NaN                   0.096546
187 (1997)                                 NaN                   0.455233

title                    Young Guns (1988)  Young Guns II (1990)  \
title
'Til There Was You (1997)               NaN                   NaN
1-900 (1994)                            NaN                   NaN
101 Dalmatians (1996)              0.119234              0.680414
12 Angry Men (1957)                0.068944             -0.361961
187 (1997)                        -0.500000              0.500000

title                    Young Poisoner's Handbook, The (1995)  \
title
'Til There Was You (1997)                                   NaN
1-900 (1994)                                                NaN
101 Dalmatians (1996)                                  0.000000
12 Angry Men (1957)                                    0.144338
187 (1997)                                             0.475327

title                    Zeus and Roxanne (1997)  unknown  \
title
'Til There Was You (1997)                      NaN      NaN
1-900 (1994)                                   NaN      NaN
101 Dalmatians (1996)                     0.707107      NaN
12 Angry Men (1957)                       1.000000      1.0
187 (1997)                                     NaN      NaN

title                    Á köldum klaka (Cold Fever) (1994)
title
'Til There Was You (1997)                                NaN
1-900 (1994)                                             NaN
101 Dalmatians (1996)                                    NaN
12 Angry Men (1957)                                      NaN
187 (1997)                                               NaN

[5 rows x 1664 columns]
```

However, we want to avoid spurious results that happened from just a handful of users that happened to rate the same pair of movies. In order to restrict our results to movies that lots of people rated together - and also give us more popular results that are more easily recongnizable - we'll use the min_periods argument to throw out results where fewer than 100 users rated a given movie pair:

```
In [4]: corrMatrix = userRatings.corr(method='pearson', min_periods=100)
        corrMatrix.head()
```

```
Out[4]: title                    'Til There Was You (1997)  1-900 (1994)  \
        title
        'Til There Was You (1997)                        NaN           NaN
        1-900 (1994)                                     NaN           NaN
        101 Dalmatians (1996)                            NaN           NaN
        12 Angry Men (1957)                              NaN           NaN
        187 (1997)                                       NaN           NaN

        title                    101 Dalmatians (1996)  12 Angry Men (1957)  \
        title
        'Til There Was You (1997)                    NaN                  NaN
        1-900 (1994)                                 NaN                  NaN
        101 Dalmatians (1996)                        1.0                  NaN
        12 Angry Men (1957)                          NaN                  1.0
        187 (1997)                                   NaN                  NaN

        title                    187 (1997)  2 Days in the Valley (1996)  \
        title
        'Til There Was You (1997)        NaN                          NaN
        1-900 (1994)                     NaN                          NaN
        101 Dalmatians (1996)            NaN                          NaN
        12 Angry Men (1957)              NaN                          NaN
        187 (1997)                       NaN                          NaN

        title                    20,000 Leagues Under the Sea (1954)  \
        title
        'Til There Was You (1997)                                  NaN
        1-900 (1994)                                               NaN
        101 Dalmatians (1996)                                      NaN
        12 Angry Men (1957)                                        NaN
        187 (1997)                                                 NaN

        title                    2001: A Space Odyssey (1968)  \
        title
        'Til There Was You (1997)                           NaN
        1-900 (1994)                                        NaN
        101 Dalmatians (1996)                               NaN
        12 Angry Men (1957)                                 NaN
        187 (1997)                                          NaN

        title                    3 Ninjas: High Noon At Mega Mountain (1998)  \
        title
        'Til There Was You (1997)                                          NaN
        1-900 (1994)                                                       NaN
        101 Dalmatians (1996)                                              NaN
```

```
12 Angry Men (1957)                                          NaN
187 (1997)                                                   NaN

title                     39 Steps, The (1935)  \
title
'Til There Was You (1997)                  NaN
1-900 (1994)                               NaN
101 Dalmatians (1996)                      NaN
12 Angry Men (1957)                        NaN
187 (1997)                                 NaN

title                                     ...                  \
title                                     ...
'Til There Was You (1997)                 ...
1-900 (1994)                              ...
101 Dalmatians (1996)                     ...
12 Angry Men (1957)                       ...
187 (1997)                                ...

title                     Yankee Zulu (1994)  Year of the Horse (1997)  \
title
'Til There Was You (1997)               NaN                       NaN
1-900 (1994)                            NaN                       NaN
101 Dalmatians (1996)                   NaN                       NaN
12 Angry Men (1957)                     NaN                       NaN
187 (1997)                              NaN                       NaN

title                     You So Crazy (1994)  Young Frankenstein (1974)  \
title
'Til There Was You (1997)                NaN                        NaN
1-900 (1994)                             NaN                        NaN
101 Dalmatians (1996)                    NaN                        NaN
12 Angry Men (1957)                      NaN                        NaN
187 (1997)                               NaN                        NaN

title                     Young Guns (1988)  Young Guns II (1990)  \
title
'Til There Was You (1997)              NaN                  NaN
1-900 (1994)                           NaN                  NaN
101 Dalmatians (1996)                  NaN                  NaN
12 Angry Men (1957)                    NaN                  NaN
187 (1997)                             NaN                  NaN

title                     Young Poisoner's Handbook, The (1995)  \
title
'Til There Was You (1997)                                   NaN
1-900 (1994)                                                NaN
101 Dalmatians (1996)                                       NaN
```

```
12 Angry Men (1957)                                          NaN
187 (1997)                                                   NaN

title                      Zeus and Roxanne (1997)  unknown  \
title
'Til There Was You (1997)                      NaN      NaN
1-900 (1994)                                   NaN      NaN
101 Dalmatians (1996)                          NaN      NaN
12 Angry Men (1957)                            NaN      NaN
187 (1997)                                     NaN      NaN

title                      Á köldum klaka (Cold Fever) (1994)
title
'Til There Was You (1997)                                  NaN
1-900 (1994)                                               NaN
101 Dalmatians (1996)                                      NaN
12 Angry Men (1957)                                        NaN
187 (1997)                                                 NaN

[5 rows x 1664 columns]
```

Now let's produce some movie recommendations for user ID 0, who I manually added to the data set as a test case. This guy really likes Star Wars and The Empire Strikes Back, but hated Gone with the Wind. I'll extract his ratings from the userRatings DataFrame, and use dropna() to get rid of missing data (leaving me only with a Series of the movies I actually rated:)

```
In [5]: myRatings = userRatings.loc[0].dropna()
        myRatings

Out[5]: title
        Empire Strikes Back, The (1980)    5.0
        Gone with the Wind (1939)          1.0
        Star Wars (1977)                   5.0
        Name: 0, dtype: float64
```

Now, let's go through each movie I rated one at a time, and build up a list of possible recommendations based on the movies similar to the ones I rated.

So for each movie I rated, I'll retrieve the list of similar movies from our correlation matrix. I'll then scale those correlation scores by how well I rated the movie they are similar to, so movies similar to ones I liked count more than movies similar to ones I hated:

```
In [6]: simCandidates = pd.Series()
        for i in range(0, len(myRatings.index)):
            print "Adding sims for " + myRatings.index[i] + "..."
            # Retrieve similar movies to this one that I rated
            sims = corrMatrix[myRatings.index[i]].dropna()
            # Now scale its similarity by how well I rated this movie
            sims = sims.map(lambda x: x * myRatings[i])
            # Add the score to the list of similarity candidates
```

```
            simCandidates = simCandidates.append(sims)

        #Glance at our results so far:
        print "sorting..."
        simCandidates.sort_values(inplace = True, ascending = False)
        print simCandidates.head(10)
```

```
Adding sims for Empire Strikes Back, The (1980)...
Adding sims for Gone with the Wind (1939)...
Adding sims for Star Wars (1977)...
sorting...
title
Empire Strikes Back, The (1980)                       5.000000
Star Wars (1977)                                      5.000000
Empire Strikes Back, The (1980)                       3.741763
Star Wars (1977)                                      3.741763
Return of the Jedi (1983)                             3.606146
Return of the Jedi (1983)                             3.362779
Raiders of the Lost Ark (1981)                        2.693297
Raiders of the Lost Ark (1981)                        2.680586
Austin Powers: International Man of Mystery (1997)    1.887164
Sting, The (1973)                                     1.837692
dtype: float64
```

This is starting to look like something useful! Note that some of the same movies came up more than once, because they were similar to more than one movie I rated. We'll use groupby() to add together the scores from movies that show up more than once, so they'll count more:

```
In [7]: simCandidates = simCandidates.groupby(simCandidates.index).sum()

In [8]: simCandidates.sort_values(inplace = True, ascending = False)
        simCandidates.head(10)
```

```
Out[8]: title
        Empire Strikes Back, The (1980)              8.877450
        Star Wars (1977)                             8.870971
        Return of the Jedi (1983)                    7.178172
        Raiders of the Lost Ark (1981)               5.519700
        Indiana Jones and the Last Crusade (1989)    3.488028
        Bridge on the River Kwai, The (1957)         3.366616
        Back to the Future (1985)                    3.357941
        Sting, The (1973)                            3.329843
        Cinderella (1950)                            3.245412
        Field of Dreams (1989)                       3.222311
        dtype: float64
```

The last thing we have to do is filter out movies I've already rated, as recommending a movie I've already watched isn't helpful:

```
In [9]: filteredSims = simCandidates.drop(myRatings.index)
        filteredSims.head(10)

Out[9]: title
        Return of the Jedi (1983)                    7.178172
        Raiders of the Lost Ark (1981)               5.519700
        Indiana Jones and the Last Crusade (1989)    3.488028
        Bridge on the River Kwai, The (1957)         3.366616
        Back to the Future (1985)                    3.357941
        Sting, The (1973)                            3.329843
        Cinderella (1950)                            3.245412
        Field of Dreams (1989)                       3.222311
        Wizard of Oz, The (1939)                     3.200268
        Dumbo (1941)                                 2.981645
        dtype: float64
```

There we have it!

1.1 Exercise

Can you improve on these results? Perhaps a different method or min_periods value on the correlation computation would produce more interesting results.

Also, it looks like some movies similar to Gone with the Wind - which I hated - made it through to the final list of recommendations. Perhaps movies similar to ones the user rated poorly should actually be penalized, instead of just scaled down?

There are also probably some outliers in the user rating data set - some users may have rated a huge amount of movies and have a disporportionate effect on the results. Go back to earlier lectures to learn how to identify these outliers, and see if removing them improves things.

For an even bigger project: we're evaluating the result qualitatively here, but we could actually apply train/test and measure our ability to predict user ratings for movies they've already watched. Whether that's actually a measure of a "good" recommendation is debatable, though!

```
In [ ]:
```