MovieLensCluestering

November 1, 2016

Projeto da disciplina de Data Mining
PESC - Programa de Engenharia de Sistemas e Computação
COPPE / UFRJ
Autor: Rafael Lopes Conde dos Reis
E-mail: condereis@cos.ufrj.br
GitHub: https://github.com/condereis/data-mining

1 Resumo

O trabalho consiste em analisar a aplicação da técnica K-means para clusterizar os filmes da base do MovieLens. Deve-se observar os clusters gerados para diferentes valores de k (15-30), assim como antes e depois de executar redução de dimensionalidade, com PCA (10-15 dimensões).

2 Pacotes Utilizados

```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        from fancyimpute import MICE
        from sklearn.cluster import KMeans
        from sklearn.decomposition import PCA

        %matplotlib inline
```

```
Using Theano backend.
```

3 Pré-processamento

A base fornecida pelo MovieLens contém 100.004 avaliações, feitas por 671 usuários, referentes a 9.125 filmes. Ela possui 4 colunas (userId, movieId, rating e timestamp) e 100.004 linhas, como pode ser visto abaixo. O pré-processamento irá gerar um dataframe *filmes x usuários* e irá avaliar o desempenho de diferentes formas de substituir os valores NaNs (filmes que não foram avaliados por um dado usuário).

```
In [2]: raw_data = pd.read_csv('../data/movielens/ratings.csv')
        raw_data.tail()
```

```
Out[2]:         userId  movieId  rating   timestamp
        99999      671     6268     2.5  1065579370
        100000     671     6269     4.0  1065149201
        100001     671     6365     4.0  1070940363
        100002     671     6385     2.5  1070979663
        100003     671     6565     3.5  1074784724
```

O site também disponibiliza uma base associando os movieIds aos títulos e gêneros dos respectivos filmes, exibida abaixo.

```
In [3]: movie_titles = pd.read_csv('../data/movielens/movies.csv')
        movie_titles.tail()
```

```
Out[3]:       movieId                                              title  \
        9120   162672                                Mohenjo Daro (2016)
        9121   163056                              Shin Godzilla (2016)
        9122   163949  The Beatles: Eight Days a Week - The Touring Y...
        9123   164977                           The Gay Desperado (1936)
        9124   164979                              Women of '69, Unboxed

                                    genres
        9120          Adventure|Drama|Romance
        9121  Action|Adventure|Fantasy|Sci-Fi
        9122                     Documentary
        9123                          Comedy
        9124                     Documentary
```

Gerando o dataframe *filmes x usuários* e concatenando a informação dos filmes chegamos ao dataframe abaixo:

```
In [4]: # Dataframe com colunas identificadas pelos ids dos filmes
        ratings_df = pd.DataFrame(
            columns=raw_data.movieId.sort_values().unique())

        for index, row in raw_data.iterrows():
            ratings_df.set_value(row.userId, row.movieId, row.rating)
```

A dimensão do dataframe se mantem a mesma (desconsiderando o fato de que foi transposto) após tentar remover linhas ou colunas que sejam apenas NaN, ou seja, não há nenhum filme que não tenha sido avaliado pelo menos por um usuário, nem nenhum usuário que não tenha avaliado ao menos um filme.

```
In [5]: print ratings_df.shape
        ratings_df = ratings_df.dropna(axis=0, how='all').dropna(
            axis=1, how='all').T
        print ratings_df.shape
```

```
(671, 9066)
(9066, 671)
```

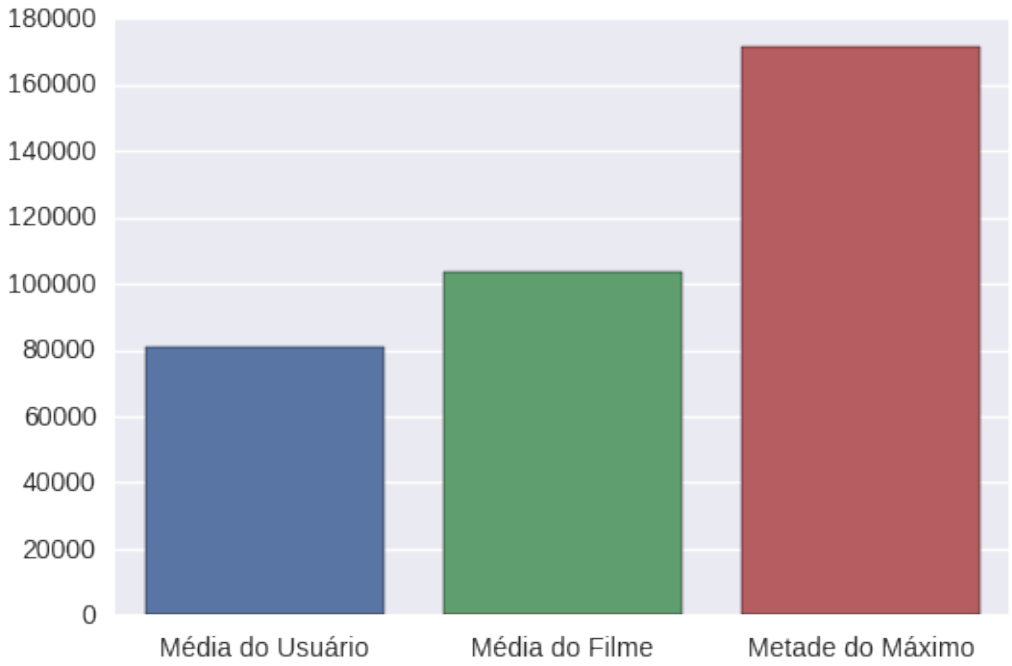
Foram testadas 4 formas de substituir os valores faltantes:

- Substituir pela média das notas do filme
- Substituir pela média das notas do usuário
- Substituir pela metade da nota máxima (2.5)

A cada dataset foi aplicado k-means com k=15 e observada a soma da distância dos pontos aos centroides dos clusters. Para tentar se aproximar do mínimo global cada teste foi repetido 10 vezes e o melhor resultado considerado. Como pode-se ver no gráfico abaixo

Serão treinados 15 modelos diferentes, variando o número de clusters entre 15 e 30. Para cada modelo foram feitas 10 inicializações (n_init=10) e escolhida a melhor, para tentar encontrar um modelo mais próximo do mínimo global, uma vez que k-means apenas garante convergência para um mínimo local. A inicialização usada foi a k-means++, que inicia os centroides afastados uns dos outros, o que tende a garantir melhores resultados que uma inicialização aleatória [1].

```
In [6]: score=[]
        method=[]
        kmeans = KMeans(n_clusters=15, init='k-means++', n_init=10,
                        n_jobs=-1)

        # Substitui NaNs pela média das notas do usuário
        fill_by_user = ratings_df.fillna(ratings_df.mean())
        score.append(-kmeans.fit(fill_by_user).score(fill_by_user))
        method.append(u'Média do Usuário')

        # Substitui NaNs pela média das notas do filme
        fill_by_movie = ratings_df.T.fillna(ratings_df.T.mean()).T
        score.append(-kmeans.fit(fill_by_movie).score(fill_by_movie))
        method.append(u'Média do Filme')

        # Substitui NaNs por 2.5
        fill_by_mean = ratings_df.fillna(2.5)
        score.append(-kmeans.fit(fill_by_mean).score(fill_by_mean))
        method.append(u'Metade do Máximo')
        sns.barplot(method, score);
```



Os demais estudos aqui realizados usaram os a média do usuário para substituir os NaNs, por ter permitido uma melhor clusterização que os demais.

```
In [7]: # Salva o dataframe em arquivo para facilitar o acesso
        fill_by_user.to_csv('../data/movielens/input.csv',
                            index=False)
        del fill_by_user
        del fill_by_movie
        del fill_by_mean
        del ratings_df
```

4 Clusterização sem Redução de Dimensionalidade

O primeiro passo é aplicas k-means a matriz de dados completa, sem nenhuma forma de redução de dimensionalidade.

```
In [8]: X = pd.read_csv('../data/movielens/input.csv')
```

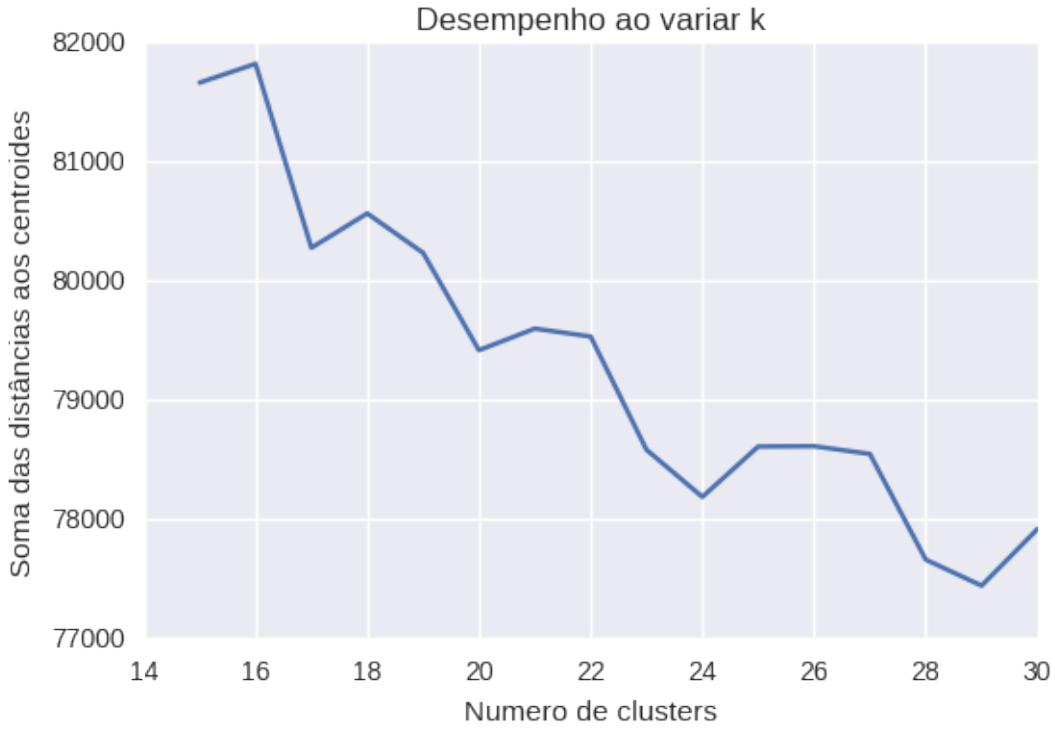
Foram treinados 15 modelos diferentes, variando o número de clusters entre 15 e 30. Para cada modelo foram feitas 10 inicializações (n_init=10) e escolhida a melhor, para tentar encontrar um modelo mais próximo do mínimo global, uma vez que k-means apenas garante convergência para um mínimo local. A inicialização usada foi a k-means++, que inicia os centroides afastados uns dos outros, o que tende a garantir melhores resultados que uma inicialização aleatória [1].

```
In [29]: score_list = []
         for k in range(15,31):
```

```
        kmeans = KMeans(n_clusters=k, init='k-means++', n_init=10,
                        n_jobs=-1).fit(X)
        score_list.append(-kmeans.score(X))
```

```
In [30]: plt.plot(range(15,31), score_list)
         plt.title('Desempenho ao variar k')
         plt.xlabel('Numero de clusters')
         plt.ylabel(u'Soma das distâncias aos centroides')
```

```
Out[30]: <matplotlib.text.Text at 0x7fbe2baf0f50>
```

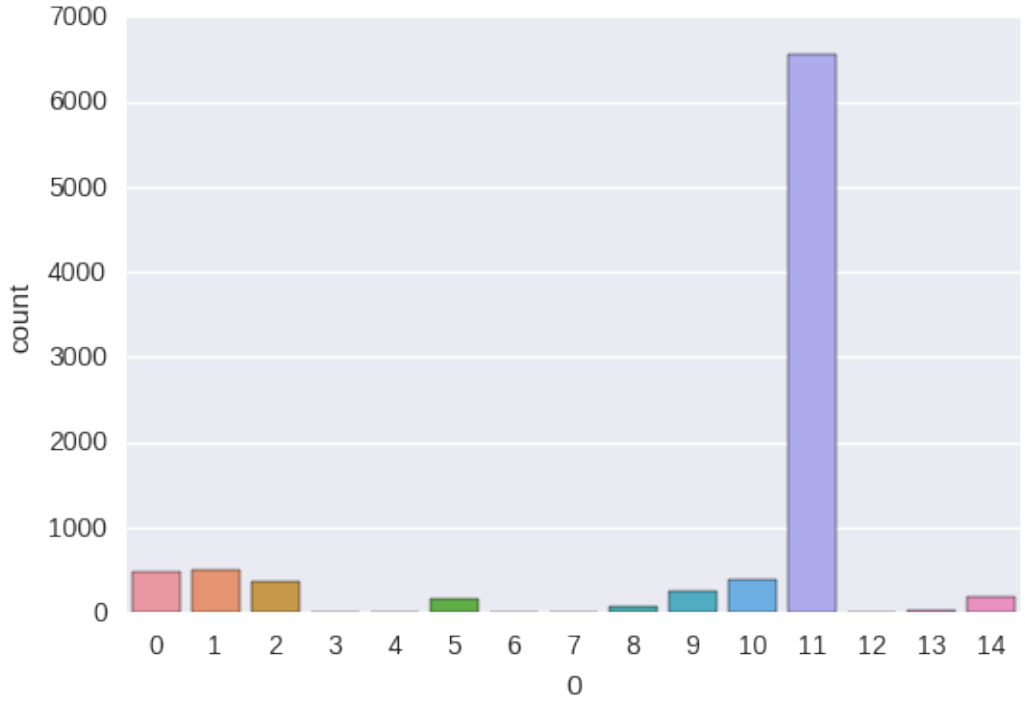


Como esperado a soma das distâncias para o centros dos clusters diminui com o aumento do numero de clusters. Houve um salto grande da soma ao mudar k de 22 para 23, o que pode ser um indicativo de um bom valor para k.Portanto analisados em detalhe os modelos com 15, 23 e 30 clusteres. Para cada um foi plotado a distribuição de eventos por cluster assim como os filmes que ficaram no mesmo cluster de *Star Wars: Episode IV - A New Hope (1977)*.

4.0.1 15 clusters

```
In [11]: kmeans = KMeans(n_clusters=15, init='k-means++', n_init=20,
                         n_jobs=-1).fit(X)
         predict = pd.Series(kmeans.predict(X), index=X.index)
         predict.head()
         out = pd.concat([predict, movie_titles], axis=1, join='inner')
         sns.countplot(out[0])
```

```
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbe2c3ec190>
```



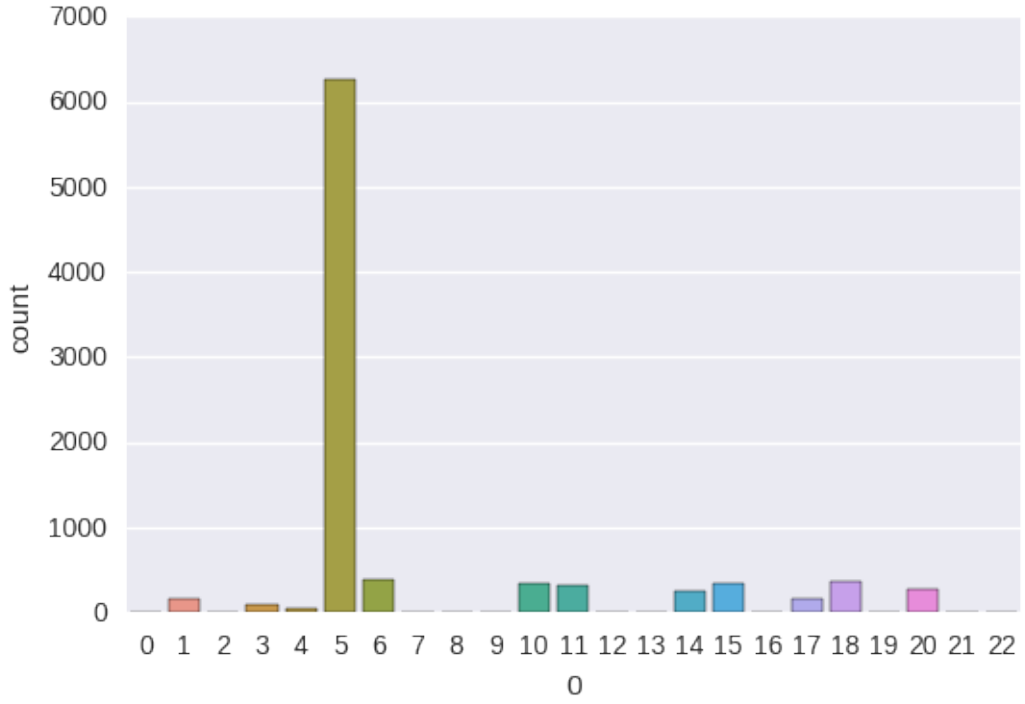

```
In [31]: cluster_starwars = out[out['movieId']==260][0].tolist()[0]
         out[out[0]==cluster_starwars]
```

```
Out[31]:       0  movieId                                              title  \
         232  20      260          Star Wars: Episode IV - A New Hope (1977)
         953  20     1196  Star Wars: Episode V - The Empire Strikes Back...
         955  20     1198  Raiders of the Lost Ark (Indiana Jones and the...
         966  20     1210  Star Wars: Episode VI - Return of the Jedi (1983)

                              genres
         232  Action|Adventure|Sci-Fi
         953  Action|Adventure|Sci-Fi
         955         Action|Adventure
         966  Action|Adventure|Sci-Fi
```

4.0.2 23 Clusters

```
In [13]: kmeans = KMeans(n_clusters=23, init='k-means++', n_init=20,
                         n_jobs=-1).fit(X)
         predict = pd.Series(kmeans.predict(X), index=X.index)
         predict.head()
         out = pd.concat([predict, movie_titles], axis=1, join='inner')
         sns.countplot(out[0])
```

Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbe2c93ec50>



In [14]:
```
cluster_starwars = out[out['movieId']==260][0].tolist()[0]
out[out[0]==cluster_starwars]
```

Out[14]:

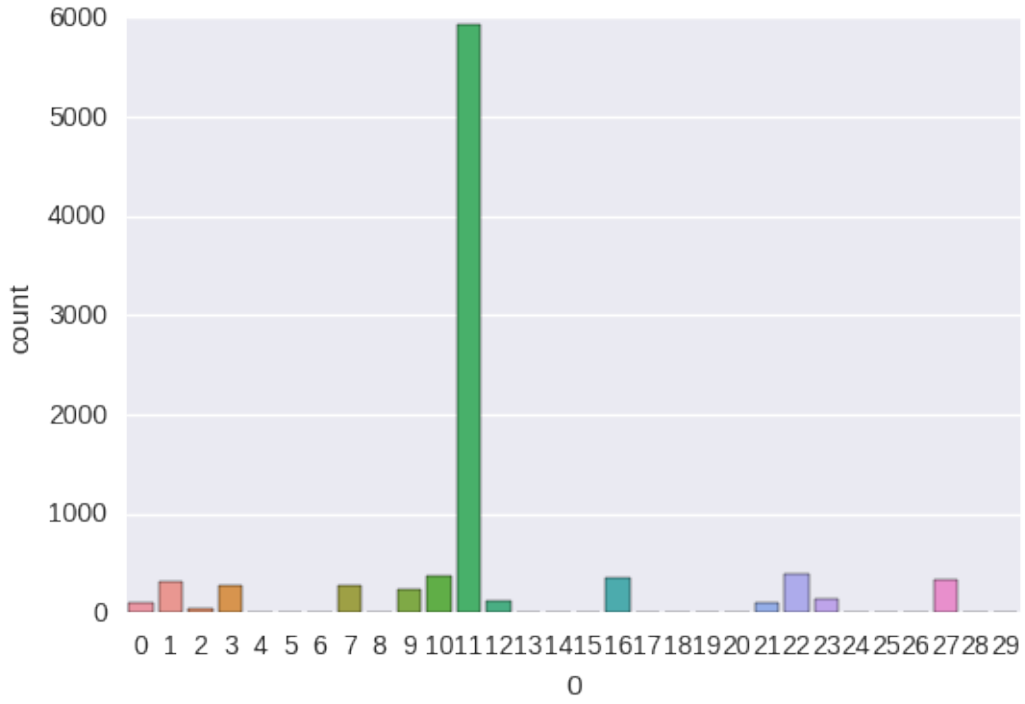
	0	movieId	title	\
232	13	260	Star Wars: Episode IV - A New Hope (1977)	
953	13	1196	Star Wars: Episode V - The Empire Strikes Back...	
955	13	1198	Raiders of the Lost Ark (Indiana Jones and the...	
966	13	1210	Star Wars: Episode VI - Return of the Jedi (1983)	

	genres
232	Action|Adventure|Sci-Fi
953	Action|Adventure|Sci-Fi
955	Action|Adventure
966	Action|Adventure|Sci-Fi

4.0.3 30 Clusters

In [15]:
```
kmeans = KMeans(n_clusters=30, init='k-means++', n_init=20,
                n_jobs=-1).fit(X)
predict = pd.Series(kmeans.predict(X), index=X.index)
predict.head()
out = pd.concat([predict, movie_titles], axis=1, join='inner')
sns.countplot(out[0])
```

```
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbe2c2d8e50>
```




```
In [16]: cluster_starwars = out[out['movieId']==260][0].tolist()[0]
         out[out[0]==cluster_starwars]
```

```
Out[16]:       0  movieId                                              title  \
         232   4      260          Star Wars: Episode IV - A New Hope (1977)
         953   4     1196  Star Wars: Episode V - The Empire Strikes Back...
         955   4     1198  Raiders of the Lost Ark (Indiana Jones and the...
         966   4     1210  Star Wars: Episode VI - Return of the Jedi (1983)

                              genres
         232  Action|Adventure|Sci-Fi
         953  Action|Adventure|Sci-Fi
         955         Action|Adventure
         966  Action|Adventure|Sci-Fi
```

É possível observar que tanto com 23 quanto com 30 clusters existem vários clusters com muito poucos filmes. Isso pode ser positivo, pois facilita encontrar padrões, mas pode ser negativo criando clusters muito especialistas. O cluster de *Star Wars* é um exemplo de ambos. Uma vez que todos os modelos conseguiram criar um cluster contendo os filmes de *Star Wars*, foi escolhido usar k=23 para as análises de clusterização, por ter apresentado um salto na eficiência e por ser apresentar uma quantidade intermediária de clusters com poucos filmes.

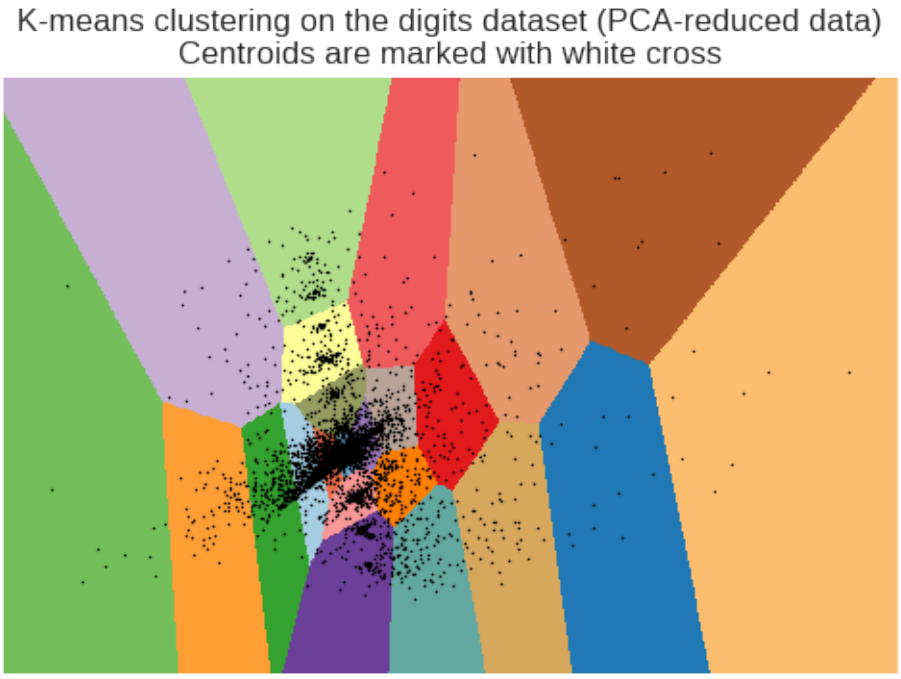
5 Aplicando PCA aos Dados

Para termos uma visualização gráfica dos dados clusterizados foi aplicada PCA com 2 dimensões. É possível observar um comportamento peculiar na distribuição. Os dados parecem se organizar em uma reta no plano gerado pela PCA. Podemos ver também que existem aproximadamente 10 agrupamentos, o que pode indicar que isso representa a nota média de cada filme. Como agrupar pela média não traz informações relevantes ao problema, é possível concluir que aplicar a PCA com dimensões próximas a 1, não irá contribuir para chegar a bons clusters.

In [17]:
```
reduced_data = PCA(n_components=2).fit_transform(X)
kmeans = KMeans(init='k-means++', n_clusters=23, n_init=10)
kmeans.fit(reduced_data)
print reduced_data.shape
# Step size of the mesh. Decrease to increase the quality of the VQ.
h = .02     # point in the mesh [x_min, x_max]x[y_min, y_max].

# Plot the decision boundary. For that, we will assign a color to each
x_min, x_max = reduced_data[:, 0].min() - 1, reduced_data[:, 0].max() + 1
y_min, y_max = reduced_data[:, 1].min() - 1, reduced_data[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_max, h

# Obtain labels for each point in mesh. Use last trained model.
Z = kmeans.predict(np.c_[xx.ravel(), yy.ravel()])

# Put the result into a color plot
Z = Z.reshape(xx.shape)
plt.figure(1)
plt.clf()
plt.imshow(Z, interpolation='nearest',
           extent=(xx.min(), xx.max(), yy.min(), yy.max()),
           cmap=plt.cm.Paired,
           aspect='auto', origin='lower')

plt.plot(reduced_data[:, 0], reduced_data[:, 1], 'k.', markersize=2)
# Plot the centroids as a white X

plt.title('K-means clustering on the digits dataset (PCA-reduced data)\n'
          'Centroids are marked with white cross')
plt.xlim(x_min, x_max)
plt.ylim(y_min, y_max)
plt.xticks(())
plt.yticks(())
```

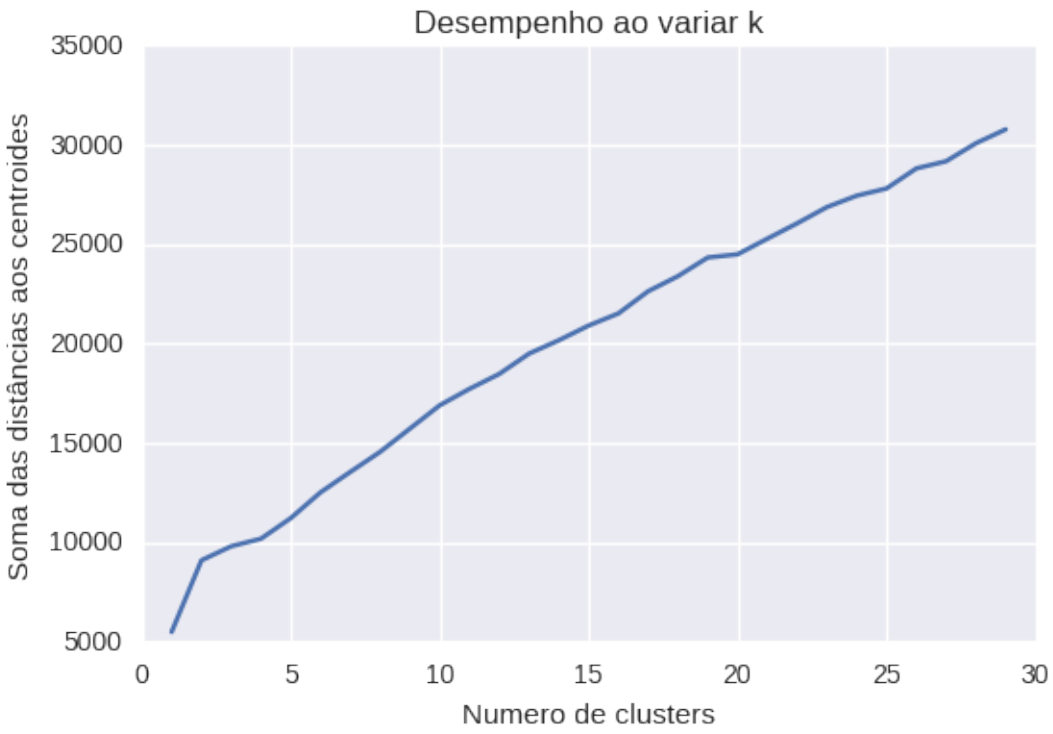
```
(9066, 2)
```

Out[17]:
```
([], <a list of 0 Text yticklabel objects>)
```



Variando o número de dimensões de 1 a 30 podemos observar que o desempenho do kmeans aumenta conforme o número de dimensões diminui. Porém como observamos anteriormente, diminuir muito o número de dimensões pode nos levar a agrupar pela média do filme.

```
In [18]: score_list = []
         for n_comp in range(1,30):
             pca = PCA(n_components=n_comp, svd_solver='arpack')
             matrix_reduc = pca.fit_transform(X.transpose())
             kmeans = KMeans(n_clusters=23, init='k-means++', n_init=20,
                             n_jobs=-1).fit(matrix_reduc)
             score_list.append(-kmeans.score(matrix_reduc))

In [19]: plt.plot(range(1,30), score_list)
         plt.title('Desempenho ao variar k')
         plt.xlabel('Numero de clusters')
         plt.ylabel(u'Soma das distâncias aos centroides')

Out[19]: <matplotlib.text.Text at 0x7fbe2c33c7d0>
```



5.0.1 13 Dimensões

```
In [20]: pca = PCA(n_components=13, svd_solver='arpack')
         matrix_reduc = pca.fit_transform(X)
         kmeans = KMeans(n_clusters=23, init='k-means++', n_init=20,
                         n_jobs=-1).fit(matrix_reduc)
         predict = pd.Series(kmeans.predict(matrix_reduc), index=X.index)
         out = pd.concat([predict, movie_titles], axis=1, join='inner')
         cluster_starwars = out[out['movieId']==260][0].tolist()[0]
         out[out[0]==cluster_starwars]
```

```
Out[20]:        0  movieId                                              title  \
         232   17      260          Star Wars: Episode IV - A New Hope (1977)
         953   17     1196  Star Wars: Episode V - The Empire Strikes Back...
         955   17     1198  Raiders of the Lost Ark (Indiana Jones and the...
         966   17     1210  Star Wars: Episode VI - Return of the Jedi (1983)
         2062  17     2571                                  Matrix, The (1999)
         3869  17     4990                Jimmy Neutron: Boy Genius (2001)
         4391  17     5944                       Star Trek: Nemesis (2002)
         5017  17     7137                               Cooler, The (2003)

                                      genres
         232              Action|Adventure|Sci-Fi
```

```
953                Action|Adventure|Sci-Fi
955                       Action|Adventure
966                Action|Adventure|Sci-Fi
2062                Action|Sci-Fi|Thriller
3869  Adventure|Animation|Children|Comedy
4391          Action|Drama|Sci-Fi|Thriller
5017                  Comedy|Drama|Romance
```

5.0.2 20 Dimensões

```
In [21]: pca = PCA(n_components=20, svd_solver='arpack')
         matrix_reduc = pca.fit_transform(X)
         kmeans = KMeans(n_clusters=23, init='k-means++', n_init=20,
                         n_jobs=-1).fit(matrix_reduc)
         predict = pd.Series(kmeans.predict(matrix_reduc), index=X.index)
         out = pd.concat([predict, movie_titles], axis=1, join='inner')
         cluster_starwars = out[out['movieId']==260][0].tolist()[0]
         out[out[0]==cluster_starwars]
```

```
Out[21]:       0  movieId                                              title  \
         232   6      260          Star Wars: Episode IV - A New Hope (1977)
         953   6     1196  Star Wars: Episode V - The Empire Strikes Back...
         955   6     1198  Raiders of the Lost Ark (Indiana Jones and the...
         966   6     1210  Star Wars: Episode VI - Return of the Jedi (1983)
         2062  6     2571                                  Matrix, The (1999)
         3869  6     4990                Jimny Neutron: Boy Genius (2001)
         4391  6     5944                       Star Trek: Nemesis (2002)
         5017  6     7137                               Cooler, The (2003)

                                      genres
         232               Action|Adventure|Sci-Fi
         953               Action|Adventure|Sci-Fi
         955                      Action|Adventure
         966               Action|Adventure|Sci-Fi
         2062               Action|Sci-Fi|Thriller
         3869  Adventure|Animation|Children|Comedy
         4391         Action|Drama|Sci-Fi|Thriller
         5017                 Comedy|Drama|Romance
```

5.0.3 30 Dimensões

```
In [32]: pca = PCA(n_components=30, svd_solver='arpack')
         matrix_reduc = pca.fit_transform(X)
         kmeans = KMeans(n_clusters=23, init='k-means++', n_init=20,
                         n_jobs=-1).fit(matrix_reduc)
         predict = pd.Series(kmeans.predict(matrix_reduc), index=X.index)
         out = pd.concat([predict, movie_titles], axis=1, join='inner')
         cluster_starwars = out[out['movieId']==260][0].tolist()[0]
         out[out[0]==cluster_starwars]
```

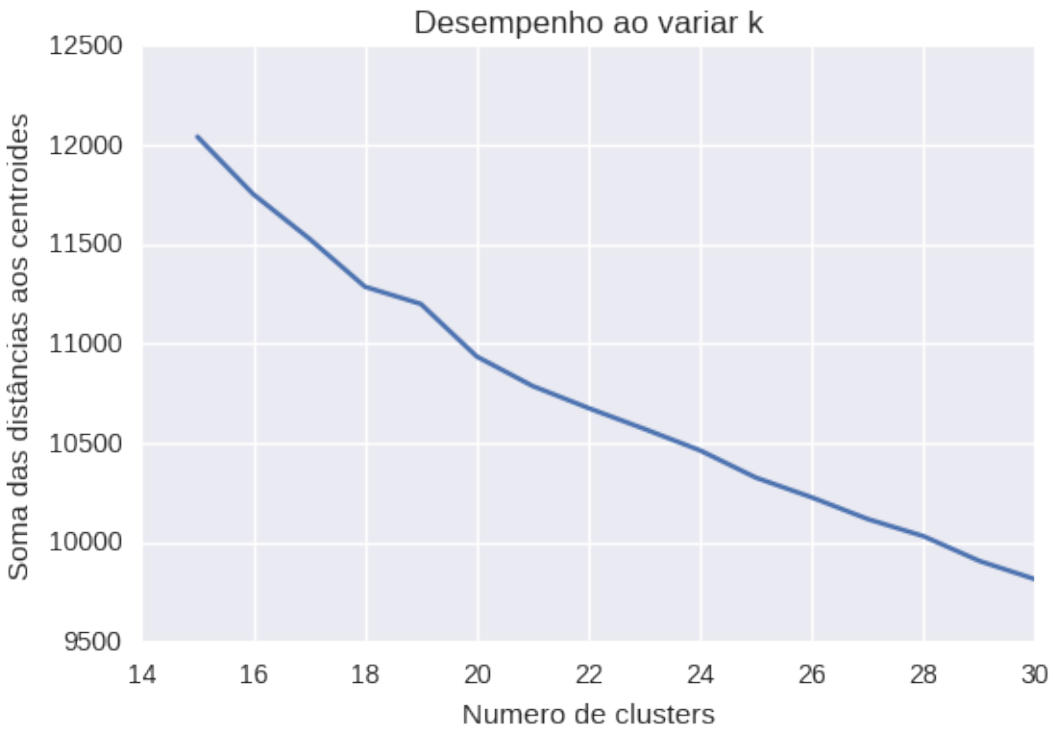
```
Out[32]:        0  movieId                                              title  \
          232   16      260          Star Wars: Episode IV - A New Hope (1977)
          953   16     1196  Star Wars: Episode V - The Empire Strikes Back...
          955   16     1198  Raiders of the Lost Ark (Indiana Jones and the...
          966   16     1210  Star Wars: Episode VI - Return of the Jedi (1983)
          2062  16     2571                                  Matrix, The (1999)
          3869  16     4990                Jimmy Neutron: Boy Genius (2001)
          4391  16     5944                       Star Trek: Nemesis (2002)
          5017  16     7137                               Cooler, The (2003)

                                          genres
          232                Action|Adventure|Sci-Fi
          953                Action|Adventure|Sci-Fi
          955                       Action|Adventure
          966                Action|Adventure|Sci-Fi
          2062                Action|Sci-Fi|Thriller
          3869  Adventure|Animation|Children|Comedy
          4391         Action|Drama|Sci-Fi|Thriller
          5017                 Comedy|Drama|Romance
```

A variação no número de dimensões não apresentou mudança significativa nos clusters observados. Foi escolhido então usar 13 dimensões uma vez que: um número menor de dimensões apresentou melhor eficiência, o kmeans não funciona bem com muitas dimensões no geral e também menos dimensões facilita o processamento.

6 Clusterização com Redução de Dimensionalidade

Variando novamente o número de clusters, após a redução de dimensionalidade, é possível notar que a variação do desempenho foi mais suave. Como esperado, a soma das distâncias aos centros continua diminuindo conforme k aumenta.

```
In [23]: score_list = []
         pca = PCA(n_components=13, svd_solver='arpack')
         matrix_reduc = pca.fit_transform(X)
         for k in range(15,31):
             kmeans = KMeans(n_clusters=k, init='k-means++', n_init=20,
                             n_jobs=-1).fit(matrix_reduc)
             score_list.append(kmeans.score(matrix_reduc))

In [24]: plt.plot(range(15,31), [-x for x in score_list])
         plt.title('Desempenho ao variar k')
         plt.xlabel('Numero de clusters')
         plt.ylabel(u'Soma das distâncias aos centroides')

Out[24]: <matplotlib.text.Text at 0x7fbe2bfd68d0>
```

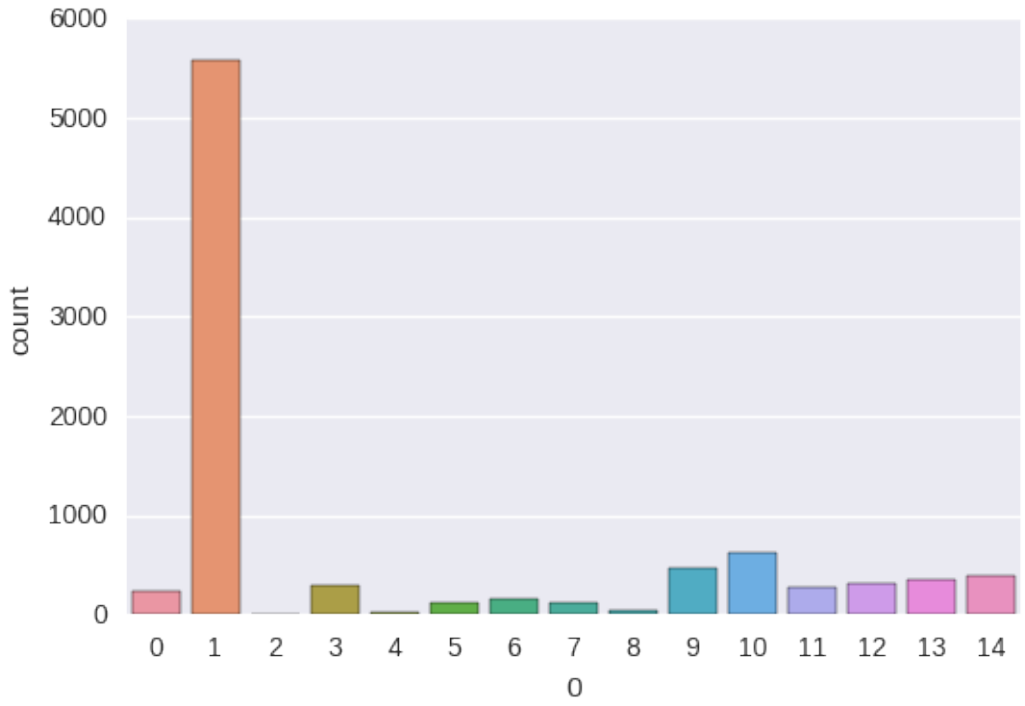


6.1 Distribuição por Cluster

É possível notar que com redução de dimensionalidade o cluster com maior representatividade passou de ter em torno de 6000 filmes para ter por volta de 5000 filmes. Podemos concluir que de fato reduzir a dimensão ajudou na capacidade do algoritmo de separar os filmes.

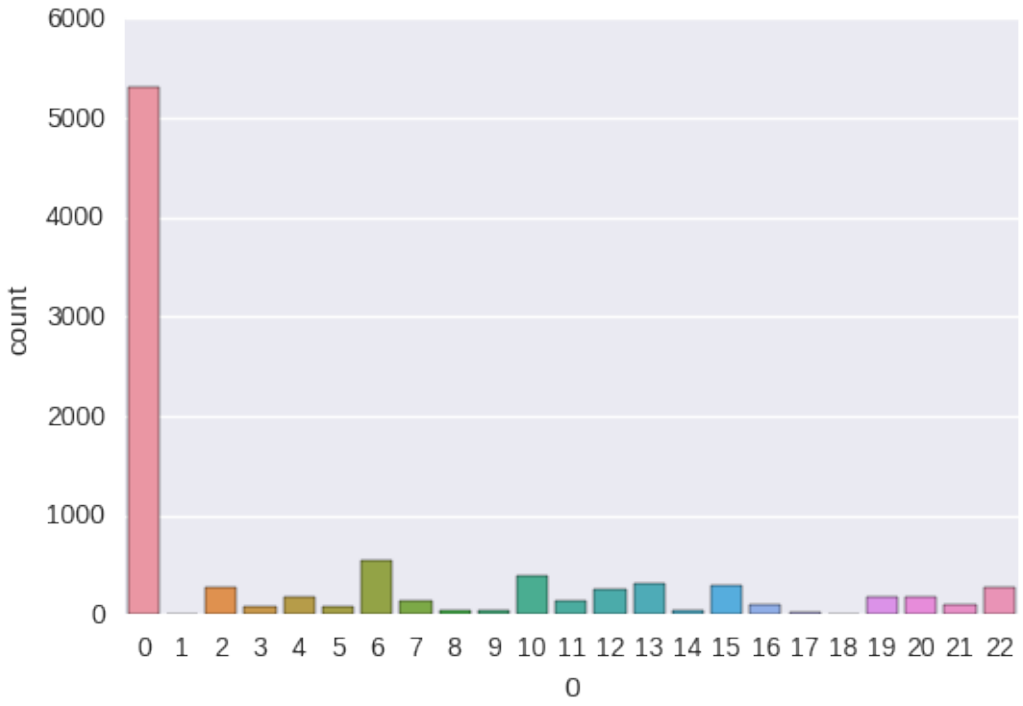
6.1.1 k=15

```
In [52]: kmeans = KMeans(n_clusters=15, init='k-means++', n_init=20,
                         n_jobs=-1).fit(matrix_reduc)
         predict = pd.Series(kmeans.predict(matrix_reduc), index=X.index)
         predict.head()
         out_15 = pd.concat([predict, movie_titles], axis=1, join='inner')
         sns.countplot(out_15[0])

Out[52]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbe255160d0>
```



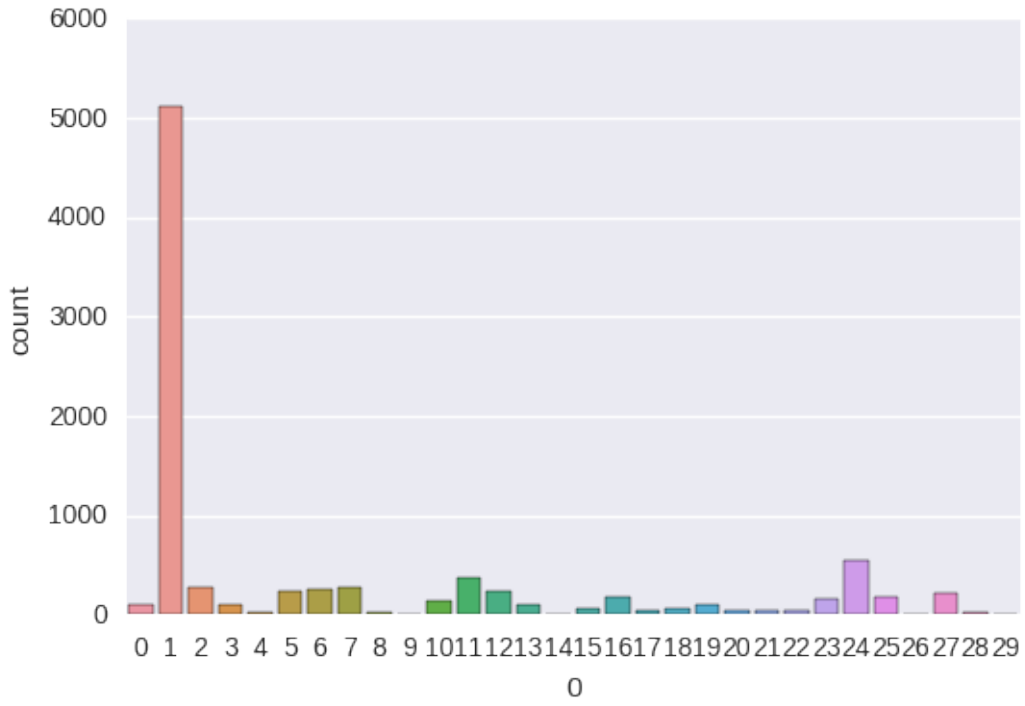
6.1.2 k=23

```
In [26]: kmeans = KMeans(n_clusters=23, init='k-means++', n_init=20,
                         n_jobs=-1).fit(matrix_reduc)
         predict = pd.Series(kmeans.predict(matrix_reduc), index=X.index)
         predict.head()
         out_23 = pd.concat([predict, movie_titles], axis=1, join='inner')
         sns.countplot(out_23[0])

Out[26]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbe2bdd0810>
```



6.1.3 k=30

```
In [27]: kmeans = KMeans(n_clusters=30, init='k-means++', n_init=20,
                         n_jobs=-1).fit(matrix_reduc)
         predict = pd.Series(kmeans.predict(matrix_reduc), index=X.index)
         predict.head()
         out_30 = pd.concat([predict, movie_titles], axis=1, join='inner')
         sns.countplot(out_30[0])

Out[27]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbe2bbf1b10>
```



6.2 Análise dos Clusters

Abaixo estão alguns clusters intereçantes encontrados usando k iqual a 15, 23 e 30.

6.2.1 k=15

Cluster Star Wars - Sci-Fi

```
In [69]: cluster_starwars = out_15[out_15['movieId']==260][0].tolist()[0]
         out_15[out_15[0]==cluster_starwars]
```

```
Out[69]:       0  movieId                                              title  \
         232   2      260          Star Wars: Episode IV - A New Hope (1977)
         953   2     1196  Star Wars: Episode V - The Empire Strikes Back...
         955   2     1198  Raiders of the Lost Ark (Indiana Jones and the...
         966   2     1210  Star Wars: Episode VI - Return of the Jedi (1983)
         2062  2     2571                                Matrix, The (1999)
         3869  2     4990                Jimmy Neutron: Boy Genius (2001)
         4391  2     5944                       Star Trek: Nemesis (2002)
         5017  2     7137                                Cooler, The (2003)

                                      genres
         232              Action|Adventure|Sci-Fi
         953              Action|Adventure|Sci-Fi
```

```
955                      Action|Adventure
966                Action|Adventure|Sci-Fi
2062                Action|Sci-Fi|Thriller
3869  Adventure|Animation|Children|Comedy
4391          Action|Drama|Sci-Fi|Thriller
5017                  Comedy|Drama|Romance
```

Gênero Crime

```
In [55]: out_15[out_15[0]==4]
```

```
Out[55]:       0  movieId                                              title  \
45     4       47                          Seven (a.k.a. Se7en) (1995)
48     4       50                           Usual Suspects, The (1995)
101    4      111                                   Taxi Driver (1976)
263    4      293  Léon: The Professional (a.k.a. The Professiona...
266    4      296                                   Pulp Fiction (1994)
284    4      318                       Shawshank Redemption, The (1994)
472    4      527                                Schindler's List (1993)
525    4      593                       Silence of the Lambs, The (1991)
535    4      608                                           Fargo (1996)
642    4      778                                    Trainspotting (1996)
695    4      858                                  Godfather, The (1972)
880    4     1089                                  Reservoir Dogs (1992)
951    4     1193               One Flew Over the Cuckoo's Nest (1975)
969    4     1213                                       Goodfellas (1990)
977    4     1221                     Godfather: Part II, The (1974)
2288   4     2858                                American Beauty (1999)
2374   4     2959                                     Fight Club (1999)
3367   4     4226                                        Memento (2000)

                              genres
45                 Mystery|Thriller
48           Crime|Mystery|Thriller
101            Crime|Drama|Thriller
263   Action|Crime|Drama|Thriller
266   Comedy|Crime|Drama|Thriller
284                 Crime|Drama
472                   Drama|War
525          Crime|Horror|Thriller
535   Comedy|Crime|Drama|Thriller
642             Comedy|Crime|Drama
695                 Crime|Drama
880          Crime|Mystery|Thriller
951                       Drama
969                 Crime|Drama
977                 Crime|Drama
2288              Drama|Romance
```

```
2374  Action|Crime|Drama|Thriller
3367              Mystery|Thriller
```

Gênero Comédia

```
In [57]: out_15[out_15[0]==7]
```

```
Out[57]:       0  movieId                                              title  \
2      7        3                            Grumpier Old Men (1995)
4      7        5                  Father of the Bride Part II (1995)
18     7       19                Ace Ventura: When Nature Calls (1995)
44     7       46                 How to Make an American Quilt (1995)
60     7       64                                  Two if by Sea (1996)
61     7       65                                       Bio-Dome (1996)
65     7       70                             From Dusk Till Dawn (1996)
75     7       81      Things to Do in Denver When You're Dead (1995)
94     7      102                                       Mr. Wrong (1996)
96     7      104                                   Happy Gilmore (1996)
136    7      157                                 Canadian Bacon (1995)
152    7      173                                    Judge Dredd (1995)
164    7      186                                    Nine Months (1995)
203    7      231              Dumb & Dumber (Dumb and Dumber) (1994)
248    7      276                                     Milk Money (1994)
283    7      317                             Santa Clause, The (1994)
309    7      344                  Ace Ventura: Pet Detective (1994)
320    7      355                             Flintstones, The (1994)
334    7      370          Naked Gun 33 1/3: The Final Insult (1994)
338    7      374                                   Richie Rich (1994)
363    7      413                                     Airheads (1994)
364    7      414                            Air Up There, The (1994)
369    7      419                     Beverly Hillbillies, The (1993)
370    7      420                        Beverly Hills Cop III (1994)
379    7      429                                     Cabin Boy (1994)
381    7      432  City Slickers II: The Legend of Curly's Gold (...
384    7      435                                    Coneheads (1993)
414    7      466                          Hot Shots! Part Deux (1993)
418    7      470                                 House Party 3 (1994)
451    7      505                                          North (1994)
...   ..      ...                                                  ...
1899   7     2389                                       Psycho (1998)
1926   7     2416                                Back to School (1986)
1955   7     2447                                 Varsity Blues (1999)
1974   7     2468                           Jumpin' Jack Flash (1986)
1975   7     2469                         Peggy Sue Got Married (1986)
1983   7     2478                                ¡Three Amigos! (1986)
1999   7     2497                           Message in a Bottle (1999)
2000   7     2498                           My Favorite Martian (1999)
2007   7     2506                             Other Sister, The (1999)
```

2050	7	2558	Forces of Nature (1999)
2129	7	2657	Rocky Horror Picture Show, The (1975)
2159	7	2699	Arachnophobia (1990)
2174	7	2717	Ghostbusters II (1989)
2192	7	2735	Golden Child, The (1986)
2236	7	2792	Airplane II: The Sequel (1982)
2238	7	2794	European Vacation (aka National Lampoon's Euro...
2240	7	2796	Funny Farm (1988)
2249	7	2805	Mickey Blue Eyes (1999)
2369	7	2950	Blue Lagoon, The (1980)
2375	7	2961	Story of Us, The (1999)
2445	7	3042	Meatballs III (1987)
2446	7	3043	Meatballs 4 (1992)
2530	7	3146	Deuce Bigalow: Male Gigolo (1999)
2578	7	3206	Against All Odds (1984)
2594	7	3235	Where the Buffalo Roam (1980)
2604	7	3248	Sister Act 2: Back in the Habit (1993)
2610	7	3254	Wayne's World 2 (1993)
2614	7	3258	Death Becomes Her (1992)
2646	7	3301	Whole Nine Yards, The (2000)
2819	7	3528	Prince of Tides, The (1991)

	genres
2	Comedy|Romance
4	Comedy
18	Comedy
44	Drama|Romance
60	Comedy|Romance
61	Comedy
65	Action|Comedy|Horror|Thriller
75	Crime|Drama|Romance
94	Comedy
96	Comedy
136	Comedy|War
152	Action|Crime|Sci-Fi
164	Comedy|Romance
203	Adventure|Comedy
248	Comedy|Romance
283	Comedy|Drama|Fantasy
309	Comedy
320	Children|Comedy|Fantasy
334	Action|Comedy
338	Children|Comedy
363	Comedy
364	Comedy
369	Comedy
370	Action|Comedy|Crime|Thriller
379	Comedy

```
381                    Adventure|Comedy|Western
384                              Comedy|Sci-Fi
414                         Action|Comedy|War
418                                     Comedy
451                                     Comedy
...                                        ...
1899                     Crime|Horror|Thriller
1926                                    Comedy
1955                              Comedy|Drama
1974           Action|Comedy|Romance|Thriller
1975                              Comedy|Drama
1983                            Comedy|Western
1999                                   Romance
2000                             Comedy|Sci-Fi
2007                      Comedy|Drama|Romance
2050                             Comedy|Romance
2129              Comedy|Horror|Musical|Sci-Fi
2159                             Comedy|Horror
2174                      Comedy|Fantasy|Sci-Fi
2192  Action|Adventure|Comedy|Fantasy|Mystery
2236                                    Comedy
2238                  Adventure|Comedy|Romance
2240                                    Comedy
2249                             Comedy|Romance
2369                   Adventure|Drama|Romance
2375                              Comedy|Drama
2445                                    Comedy
2446                                    Comedy
2530                                    Comedy
2578                                   Romance
2594                                    Comedy
2604                                    Comedy
2610                                    Comedy
2614                             Comedy|Fantasy
2646                               Comedy|Crime
2819                              Drama|Romance

[132 rows x 4 columns]
```

6.2.2 k=23

Usando k=23 Foi possível encontrar um cluster bem representativo para o gênero drama, além dos clusters de comédia, crime e sci-fi já encontrada anteriormente. O cluster crime passou a conter menos filmes. Um deles foi *One Flew Over the Cuckoo's Nest (1975)* que passou corretamente para o cluster drama.

Cluster Star Wars - Sci-Fi

```
In [58]: cluster_starwars = out_23[out_23['movieId']==260][0].tolist()[0]
         out_23[out_23[0]==cluster_starwars]
```

```
Out[58]:       0  movieId                                               title  \
         232   1      260          Star Wars: Episode IV - A New Hope (1977)
         953   1     1196  Star Wars: Episode V - The Empire Strikes Back...
         955   1     1198  Raiders of the Lost Ark (Indiana Jones and the...
         966   1     1210  Star Wars: Episode VI - Return of the Jedi (1983)
         2062  1     2571                                  Matrix, The (1999)
         3869  1     4990                Jimmy Neutron: Boy Genius (2001)
         4391  1     5944                       Star Trek: Nemesis (2002)
         5017  1     7137                                Cooler, The (2003)

                                      genres
         232               Action|Adventure|Sci-Fi
         953               Action|Adventure|Sci-Fi
         955                      Action|Adventure
         966               Action|Adventure|Sci-Fi
         2062               Action|Sci-Fi|Thriller
         3869  Adventure|Animation|Children|Comedy
         4391         Action|Drama|Sci-Fi|Thriller
         5017                Comedy|Drama|Romance
```

Gênero Crime

```
In [60]: out_23[out_23[0]==18]
```

```
Out[60]:       0  movieId                                title  \
         48    18       50         Usual Suspects, The (1995)
         266   18      296                Pulp Fiction (1994)
         284   18      318  Shawshank Redemption, The (1994)
         472   18      527            Schindler's List (1993)
         525   18      593  Silence of the Lambs, The (1991)
         535   18      608                       Fargo (1996)
         695   18      858              Godfather, The (1972)
         969   18     1213                  Goodfellas (1990)
         977   18     1221     Godfather: Part II, The (1974)
         2288  18     2858             American Beauty (1999)

                                 genres
         48          Crime|Mystery|Thriller
         266   Comedy|Crime|Drama|Thriller
         284                 Crime|Drama
         472                   Drama|War
         525         Crime|Horror|Thriller
         535   Comedy|Crime|Drama|Thriller
         695                 Crime|Drama
         969                 Crime|Drama
```

```
977                    Crime|Drama
2288                 Drama|Romance
```

Gênero Drama

```
In [63]: out_23[out_23[0]==14]
```

```
Out[63]:        0  movieId                                              title  \
101    14      111                                   Taxi Driver (1976)
218    14      246                                   Hoop Dreams (1994)
626    14      750  Dr. Strangelove or: How I Learned to Stop Worr...
720    14      899                             Singin' in the Rain (1952)
724    14      903                                        Vertigo (1958)
725    14      904                                   Rear Window (1954)
729    14      908                            North by Northwest (1959)
733    14      912                                    Casablanca (1942)
734    14      913                            Maltese Falcon, The (1941)
740    14      919                              Wizard of Oz, The (1939)
744    14      923                                   Citizen Kane (1941)
745    14      924                          2001: A Space Odyssey (1968)
773    14      953                          It's a Wonderful Life (1946)
858    14     1060                                       Swingers (1996)
912    14     1136                  Monty Python and the Holy Grail (1975)
951    14     1193                  One Flew Over the Cuckoo's Nest (1975)
960    14     1203                                   12 Angry Men (1957)
963    14     1207                          To Kill a Mockingbird (1962)
964    14     1208                                Apocalypse Now (1979)
975    14     1219                                        Psycho (1960)
981    14     1225                                      Amadeus (1984)
984    14     1228                                   Raging Bull (1980)
985    14     1230                                    Annie Hall (1977)
989    14     1234                                    Sting, The (1973)
998    14     1244                                     Manhattan (1979)
1001   14     1247                                 Graduate, The (1967)
1004   14     1250                   Bridge on the River Kwai, The (1957)
1006   14     1252                                     Chinatown (1974)
1012   14     1258                                  Shining, The (1980)
1019   14     1265                                Groundhog Day (1993)
1021   14     1267                      Manchurian Candidate, The (1962)
1042   14     1288                               This Is Spinal Tap (1984)
1057   14     1304               Butch Cassidy and the Sundance Kid (1969)
1060   14     1307                           When Harry Met Sally... (1989)
1125   14     1387                                            Jaws (1975)
1288   14     1617                               L.A. Confidential (1997)
1581   14     2019             Seven Samurai (Shichinin no samurai) (1954)
1624   14     2064                                   Roger & Me (1989)
1811   14     2289                                  Player, The (1992)
2235   14     2791                                    Airplane! (1980)
```

2688	14	3362	Dog Day Afternoon (1975)
2780	14	3481	High Fidelity (2000)
2798	14	3504	Network (1976)

	genres
101	Crime|Drama|Thriller
218	Documentary
626	Comedy|War
720	Comedy|Musical|Romance
724	Drama|Mystery|Romance|Thriller
725	Mystery|Thriller
729	Action|Adventure|Mystery|Romance|Thriller
733	Drama|Romance
734	Film-Noir|Mystery
740	Adventure|Children|Fantasy|Musical
744	Drama|Mystery
745	Adventure|Drama|Sci-Fi
773	Children|Drama|Fantasy|Romance
858	Comedy|Drama
912	Adventure|Comedy|Fantasy
951	Drama
960	Drama
963	Drama
964	Action|Drama|War
975	Crime|Horror
981	Drama
984	Drama
985	Comedy|Romance
989	Comedy|Crime
998	Comedy|Drama|Romance
1001	Comedy|Drama|Romance
1004	Adventure|Drama|War
1006	Crime|Film-Noir|Mystery|Thriller
1012	Horror
1019	Comedy|Fantasy|Romance
1021	Crime|Thriller|War
1042	Comedy
1057	Action|Western
1060	Comedy|Romance
1125	Action|Horror
1288	Crime|Film-Noir|Mystery|Thriller
1581	Action|Adventure|Drama
1624	Documentary
1811	Comedy|Crime|Drama
2235	Comedy
2688	Crime|Drama
2780	Comedy|Drama|Romance
2798	Comedy|Drama

```
In [67]: cluster_starwars = out_23[out_23['movieId']==231][0].tolist()[0]
         out_23[out_23[0]==16]
```

Out[67]:

	0	movieId	title
2	16	3	Grumpier Old Men (1995)
4	16	5	Father of the Bride Part II (1995)
60	16	64	Two if by Sea (1996)
61	16	65	Bio-Dome (1996)
65	16	70	From Dusk Till Dawn (1996)
75	16	81	Things to Do in Denver When You're Dead (1995)
94	16	102	Mr. Wrong (1996)
96	16	104	Happy Gilmore (1996)
136	16	157	Canadian Bacon (1995)
203	16	231	Dumb & Dumber (Dumb and Dumber) (1994)
248	16	276	Milk Money (1994)
320	16	355	Flintstones, The (1994)
334	16	370	Naked Gun 33 1/3: The Final Insult (1994)
338	16	374	Richie Rich (1994)
363	16	413	Airheads (1994)
364	16	414	Air Up There, The (1994)
369	16	419	Beverly Hillbillies, The (1993)
379	16	429	Cabin Boy (1994)
381	16	432	City Slickers II: The Legend of Curly's Gold (...
414	16	466	Hot Shots! Part Deux (1993)
418	16	470	House Party 3 (1994)
451	16	505	North (1994)
462	16	516	Renaissance Man (1994)
464	16	518	Road to Wellville, The (1994)
466	16	520	Robin Hood: Men in Tights (1993)
489	16	546	Super Mario Bros. (1993)
495	16	552	Three Musketeers, The (1993)
569	16	663	Kids in the Hall: Brain Candy (1996)
601	16	710	Celtic Pride (1996)
606	16	719	Multiplicity (1996)
...	..	...	...
1893	16	2383	Police Academy 6: City Under Siege (1989)
1895	16	2385	Home Fries (1998)
1899	16	2389	Psycho (1998)
1902	16	2392	Jack Frost (1998)
1955	16	2447	Varsity Blues (1999)
1974	16	2468	Jumpin' Jack Flash (1986)
1975	16	2469	Peggy Sue Got Married (1986)
1979	16	2473	Soul Man (1986)
1983	16	2478	¡Three Amigos! (1986)
1999	16	2497	Message in a Bottle (1999)
2000	16	2498	My Favorite Martian (1999)
2007	16	2506	Other Sister, The (1999)
2050	16	2558	Forces of Nature (1999)

2129	16	2657	Rocky Horror Picture Show, The (1975)
2159	16	2699	Arachnophobia (1990)
2174	16	2717	Ghostbusters II (1989)
2192	16	2735	Golden Child, The (1986)
2236	16	2792	Airplane II: The Sequel (1982)
2238	16	2794	European Vacation (aka National Lampoon's Euro...
2240	16	2796	Funny Farm (1988)
2369	16	2950	Blue Lagoon, The (1980)
2375	16	2961	Story of Us, The (1999)
2445	16	3042	Meatballs III (1987)
2446	16	3043	Meatballs 4 (1992)
2530	16	3146	Deuce Bigalow: Male Gigolo (1999)
2578	16	3206	Against All Odds (1984)
2594	16	3235	Where the Buffalo Roam (1980)
2604	16	3248	Sister Act 2: Back in the Habit (1993)
2610	16	3254	Wayne's World 2 (1993)
2646	16	3301	Whole Nine Yards, The (2000)

	genres
2	Comedy|Romance
4	Comedy
60	Comedy|Romance
61	Comedy
65	Action|Comedy|Horror|Thriller
75	Crime|Drama|Romance
94	Comedy
96	Comedy
136	Comedy|War
203	Adventure|Comedy
248	Comedy|Romance
320	Children|Comedy|Fantasy
334	Action|Comedy
338	Children|Comedy
363	Comedy
364	Comedy
369	Comedy
379	Comedy
381	Adventure|Comedy|Western
414	Action|Comedy|War
418	Comedy
451	Comedy
462	Comedy|Drama
464	Comedy
466	Comedy
489	Action|Adventure|Children|Comedy|Fantasy|Sci-Fi
495	Action|Adventure|Comedy|Romance
569	Comedy
601	Comedy

```
606                                                Comedy
...                                                   ...
1893                                       Comedy|Crime
1895                                     Comedy|Romance
1899                             Crime|Horror|Thriller
1902                            Children|Comedy|Drama
1955                                       Comedy|Drama
1974                 Action|Comedy|Romance|Thriller
1975                                       Comedy|Drama
1979                                              Comedy
1983                                     Comedy|Western
1999                                            Romance
2000                                     Comedy|Sci-Fi
2007                              Comedy|Drama|Romance
2050                                     Comedy|Romance
2129                   Comedy|Horror|Musical|Sci-Fi
2159                                     Comedy|Horror
2174                            Comedy|Fantasy|Sci-Fi
2192          Action|Adventure|Comedy|Fantasy|Mystery
2236                                              Comedy
2238                         Adventure|Comedy|Romance
2240                                              Comedy
2369                          Adventure|Drama|Romance
2375                                       Comedy|Drama
2445                                              Comedy
2446                                              Comedy
2530                                              Comedy
2578                                             Romance
2594                                              Comedy
2604                                              Comedy
2610                                              Comedy
2646                                       Comedy|Crime

[112 rows x 4 columns]
```

6.2.3 k=30

Os clusters de sci-fi, crime e comédia se mantiveram. O cluster drama passou a ser algo mais para filmes clássicos (que possuem nota alta no IMDB ou de diretores conceituados). Foi possível também encontrar um cluster de filmes dos anos 90 bastante populares que foi chamado de "Sessão da Tarde".

Cluster Star Wars - Sci-Fi

```
In [83]: cluster_starwars = out_30[out_30['movieId']==260][0].tolist()[0]
         out_30[out_30[0]==cluster_starwars]
```

```
Out[83]:        0  movieId                                              title  \
         232   14      260          Star Wars: Episode IV - A New Hope (1977)
```

```
953    14     1196  Star Wars: Episode V - The Empire Strikes Back...
955    14     1198  Raiders of the Lost Ark (Indiana Jones and the...
966    14     1210  Star Wars: Episode VI - Return of the Jedi (1983)
2062   14     2571                                Matrix, The (1999)
3869   14     4990               Jimmy Neutron: Boy Genius (2001)
4391   14     5944                      Star Trek: Nemesis (2002)
5017   14     7137                              Cooler, The (2003)

                                    genres
232               Action|Adventure|Sci-Fi
953               Action|Adventure|Sci-Fi
955                      Action|Adventure
966               Action|Adventure|Sci-Fi
2062               Action|Sci-Fi|Thriller
3869  Adventure|Animation|Children|Comedy
4391         Action|Drama|Sci-Fi|Thriller
5017                 Comedy|Drama|Romance
```

Gênero Crime

```
In [79]: pulp_fiction = out_30[out_30['movieId']==296][0].tolist()[0]
         out_30[out_30[0]==pulp_fiction]
```

```
Out[79]:       0  movieId                             title  \
         48    9       50        Usual Suspects, The (1995)
         266   9      296               Pulp Fiction (1994)
         284   9      318  Shawshank Redemption, The (1994)
         472   9      527           Schindler's List (1993)
         525   9      593  Silence of the Lambs, The (1991)
         535   9      608                      Fargo (1996)
         695   9      858            Godfather, The (1972)
         2288  9     2858           American Beauty (1999)

                              genres
         48         Crime|Mystery|Thriller
         266   Comedy|Crime|Drama|Thriller
         284                Crime|Drama
         472                  Drama|War
         525         Crime|Horror|Thriller
         535   Comedy|Crime|Drama|Thriller
         695                Crime|Drama
         2288              Drama|Romance
```

Gênero Comédia

```
In [81]: dumb_and_dumber = out_30[out_30['movieId']==231][0].tolist()[0]
         out_30[out_30[0]==dumb_and_dumber]
```

Out[81]:

	0	movieId	title
2	19	3	Grumpier Old Men (1995)
4	19	5	Father of the Bride Part II (1995)
60	19	64	Two if by Sea (1996)
61	19	65	Bio-Dome (1996)
65	19	70	From Dusk Till Dawn (1996)
75	19	81	Things to Do in Denver When You're Dead (1995)
94	19	102	Mr. Wrong (1996)
136	19	157	Canadian Bacon (1995)
203	19	231	Dumb & Dumber (Dumb and Dumber) (1994)
248	19	276	Milk Money (1994)
320	19	355	Flintstones, The (1994)
334	19	370	Naked Gun 33 1/3: The Final Insult (1994)
338	19	374	Richie Rich (1994)
363	19	413	Airheads (1994)
364	19	414	Air Up There, The (1994)
369	19	419	Beverly Hillbillies, The (1993)
379	19	429	Cabin Boy (1994)
381	19	432	City Slickers II: The Legend of Curly's Gold (...
414	19	466	Hot Shots! Part Deux (1993)
418	19	470	House Party 3 (1994)
451	19	505	North (1994)
462	19	516	Renaissance Man (1994)
464	19	518	Road to Wellville, The (1994)
466	19	520	Robin Hood: Men in Tights (1993)
489	19	546	Super Mario Bros. (1993)
495	19	552	Three Musketeers, The (1993)
569	19	663	Kids in the Hall: Brain Candy (1996)
601	19	710	Celtic Pride (1996)
606	19	719	Multiplicity (1996)
611	19	725	Great White Hype, The (1996)
...	..	...	...
1891	19	2381	Police Academy 4: Citizens on Patrol (1987)
1892	19	2382	Police Academy 5: Assignment: Miami Beach (1988)
1893	19	2383	Police Academy 6: City Under Siege (1989)
1895	19	2385	Home Fries (1998)
1902	19	2392	Jack Frost (1998)
1955	19	2447	Varsity Blues (1999)
1974	19	2468	Jumpin' Jack Flash (1986)
1975	19	2469	Peggy Sue Got Married (1986)
1979	19	2473	Soul Man (1986)
1983	19	2478	¡Three Amigos! (1986)
1999	19	2497	Message in a Bottle (1999)
2000	19	2498	My Favorite Martian (1999)
2007	19	2506	Other Sister, The (1999)
2050	19	2558	Forces of Nature (1999)
2129	19	2657	Rocky Horror Picture Show, The (1975)
2159	19	2699	Arachnophobia (1990)

2174	19	2717	Ghostbusters II (1989)
2192	19	2735	Golden Child, The (1986)
2236	19	2792	Airplane II: The Sequel (1982)
2238	19	2794	European Vacation (aka National Lampoon's Euro...
2240	19	2796	Funny Farm (1988)
2369	19	2950	Blue Lagoon, The (1980)
2375	19	2961	Story of Us, The (1999)
2445	19	3042	Meatballs III (1987)
2446	19	3043	Meatballs 4 (1992)
2530	19	3146	Deuce Bigalow: Male Gigolo (1999)
2578	19	3206	Against All Odds (1984)
2594	19	3235	Where the Buffalo Roam (1980)
2610	19	3254	Wayne's World 2 (1993)
2646	19	3301	Whole Nine Yards, The (2000)

	genres
2	Comedy|Romance
4	Comedy
60	Comedy|Romance
61	Comedy
65	Action|Comedy|Horror|Thriller
75	Crime|Drama|Romance
94	Comedy
136	Comedy|War
203	Adventure|Comedy
248	Comedy|Romance
320	Children|Comedy|Fantasy
334	Action|Comedy
338	Children|Comedy
363	Comedy
364	Comedy
369	Comedy
379	Comedy
381	Adventure|Comedy|Western
414	Action|Comedy|War
418	Comedy
451	Comedy
462	Comedy|Drama
464	Comedy
466	Comedy
489	Action|Adventure|Children|Comedy|Fantasy|Sci-Fi
495	Action|Adventure|Comedy|Romance
569	Comedy
601	Comedy
606	Comedy
611	Comedy
...	...
1891	Comedy|Crime

```
1892                                    Comedy|Crime
1893                                    Comedy|Crime
1895                                  Comedy|Romance
1902                              Children|Comedy|Drama
1955                                    Comedy|Drama
1974                  Action|Comedy|Romance|Thriller
1975                                    Comedy|Drama
1979                                          Comedy
1983                                  Comedy|Western
1999                                         Romance
2000                                  Comedy|Sci-Fi
2007                               Comedy|Drama|Romance
2050                                  Comedy|Romance
2129                     Comedy|Horror|Musical|Sci-Fi
2159                                   Comedy|Horror
2174                            Comedy|Fantasy|Sci-Fi
2192          Action|Adventure|Comedy|Fantasy|Mystery
2236                                          Comedy
2238                        Adventure|Comedy|Romance
2240                                          Comedy
2369                         Adventure|Drama|Romance
2375                                    Comedy|Drama
2445                                          Comedy
2446                                          Comedy
2530                                          Comedy
2578                                         Romance
2594                                          Comedy
2610                                          Comedy
2646                                    Comedy|Crime

[109 rows x 4 columns]
```

Clássicos

```
In [87]: out_30[out_30[0]==4]
```

```
Out[87]:       0  movieId                                              title  \
101    4      111                                  Taxi Driver (1976)
485    4      541                                 Blade Runner (1982)
626    4      750  Dr. Strangelove or: How I Learned to Stop Worr...
642    4      778                                Trainspotting (1996)
724    4      903                                     Vertigo (1958)
725    4      904                                  Rear Window (1954)
729    4      908                           North by Northwest (1959)
733    4      912                                   Casablanca (1942)
744    4      923                                 Citizen Kane (1941)
745    4      924                       2001: A Space Odyssey (1968)
880    4     1089                                Reservoir Dogs (1992)
```

912	4	1136	Monty Python and the Holy Grail	(1975)
951	4	1193	One Flew Over the Cuckoo's Nest	(1975)
962	4	1206	Clockwork Orange, A	(1971)
964	4	1208	Apocalypse Now	(1979)
969	4	1213	Goodfellas	(1990)
975	4	1219	Psycho	(1960)
977	4	1221	Godfather: Part II, The	(1974)
978	4	1222	Full Metal Jacket	(1987)
1001	4	1247	Graduate, The	(1967)
1006	4	1252	Chinatown	(1974)
1012	4	1258	Shining, The	(1980)
1288	4	1617	L.A. Confidential	(1997)
1367	4	1732	Big Lebowski, The	(1998)
1581	4	2019	Seven Samurai (Shichinin no samurai)	(1954)
2407	4	2997	Being John Malkovich	(1999)
2780	4	3481	High Fidelity	(2000)

	genres
101	Crime|Drama|Thriller
485	Action|Sci-Fi|Thriller
626	Comedy|War
642	Comedy|Crime|Drama
724	Drama|Mystery|Romance|Thriller
725	Mystery|Thriller
729	Action|Adventure|Mystery|Romance|Thriller
733	Drama|Romance
744	Drama|Mystery
745	Adventure|Drama|Sci-Fi
880	Crime|Mystery|Thriller
912	Adventure|Comedy|Fantasy
951	Drama
962	Crime|Drama|Sci-Fi|Thriller
964	Action|Drama|War
969	Crime|Drama
975	Crime|Horror
977	Crime|Drama
978	Drama|War
1001	Comedy|Drama|Romance
1006	Crime|Film-Noir|Mystery|Thriller
1012	Horror
1288	Crime|Film-Noir|Mystery|Thriller
1367	Comedy|Crime
1581	Action|Adventure|Drama
2407	Comedy|Drama|Fantasy
2780	Comedy|Drama|Romance

Filmes da "Sessão da Tarde"

```
In [86]: out_30[out_30[0]==29]
```

```
Out[86]:        0  movieId                                              title  \
          18    29       19                Ace Ventura: When Nature Calls (1995)
          132   29      153                                   Batman Forever (1995)
          184   29      208                                       Waterworld (1995)
          309   29      344                   Ace Ventura: Pet Detective (1994)
          331   29      367                                    Mask, The (1994)
          519   29      586                                   Home Alone (1990)
          644   29      780                 Independence Day (a.k.a. ID4) (1996)
          1243  29     1562                                   Batman & Robin (1997)
          1480  29     1917                                       Armageddon (1998)
          1615  29     2054                     Honey, I Shrunk the Kids (1989)
          2103  29     2628  Star Wars: Episode I - The Phantom Menace (1999)
          2161  29     2701                                   Wild Wild West (1999)
          2893  29     3623                       Mission: Impossible II (2000)

                                                 genres
          18                                     Comedy
          132              Action|Adventure|Comedy|Crime
          184                    Action|Adventure|Sci-Fi
          309                                    Comedy
          331                Action|Comedy|Crime|Fantasy
          519                            Children|Comedy
          644            Action|Adventure|Sci-Fi|Thriller
          1243          Action|Adventure|Fantasy|Thriller
          1480             Action|Romance|Sci-Fi|Thriller
          1615  Adventure|Children|Comedy|Fantasy|Sci-Fi
          2103                   Action|Adventure|Sci-Fi
          2161               Action|Comedy|Sci-Fi|Western
          2893                  Action|Adventure|Thriller
```

7 Conclusão

O modelo usando 30 clusters com redução da dimensão para 15 foi o que apresentou uma maior quantidade de cluster fe fácil identificação. Foram um total de 5 clusters identificados: Sci-fi, crime, comédia, clássicos, Sessão da Tarde. Uma análise mais aprofundada nos clusters com maior ocorrência de filmes pode vir a mostrar padrões mais sutis que possam não ter sido observadas.

Como pode ser observado pela imagem dos clusters em 2 dimensões a informação média das notas tem uma grande influência. Isso provavelmente é causado pelos dados faltantes estarem sendo substituídos apenas pela média do usuário. O uso de alguma técnica que leve em consideração as relações dos dados da base para gerar a nota faltante poder apresentar um melhor resultado.

Por fim, o uso de um modelo que possa criar clusters não convexos e que consiga lidar bem com altas dimensões pode via a apresentar um resultado melhor para este problema.

8 Referências

[1] - k-means++: The advantages of careful seeding, Arthur, David, and Sergei Vassilvitskii, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics (2007)