TOPICS IN DATA ANLASYSIS

Clustering the 1m-Movie Lens dataset

Topics in big data analysis
Mini project
Guy Amit: 311170971
Yuval Hadad: 204527253

Mini project- data analysis

Clustering the Movie Lens data set

Table of Contents

Topics in big data analysis
Mini project
Guy Amit: 311170971
Yuval Hadad: 204527253

The Datasets

The Movie Lens database contains 3 tables:

- Movies
- Users
- Ratings

From those tables we have created new datasets as follows:

- Correlations_data set: we wanted to store all the probabilities in a sperate dataset, so we have created the matrix c that is defined as : $c_{i,j} = \begin{cases} p(m_i, m_j) & if\ i \neq j \\ p(m_i) & if\ i = j \end{cases}$
- Correlations_indicator: a matrix of 1's and 0's s.t. $c_{i,j} = \begin{cases} 1\ if\ p(m_i, m_j) \geq p(m_i)p(m_j) \\ 0 \qquad\qquad\qquad\quad otherwise \end{cases}$

  At both of those Datasets the first column and row holds the id's of the movies.

- Movies_name: a vector that contain in i^{th} entry the name of the movie with the id i.
- Bad_movies: a 630×2 matrix of movies id's and their corresponding number of rating. This matrix lets us track movies that have less than 10 ratings.

The algorithms

The Pivot algorithm:

The Pivot algorithm is factor 3 approximation to the problem of correlation clustering.
It is a pretty naïve algorithm, that choose a random vertex from the graph and cluster it together with all of its neighbors. In our implementation we are using the correlation matrix as an adjacency matrix that represent the graph of the problem.

We've looked up the algorithm in Wikipedia and found the pseudo code for it and implemented it in a vectorized way to save computational time. Our implementation runs pretty fast even when trying to cluster all the movies in the dataset.

Main challenge:

Realizing that the algorithm was pretty easy, when we've got the algorithm pseudo code it took a short time to implement.

Scalability:

As said before the algorithm scales very well.

Average cost on subsets of size 100: 733.5534

The Max Clique algorithm-first attempt for improvement:

Among the assignment's specifications there is a theorem that states:

$$cost(C_1) \le cost(C_2) \iff p(m_1, m_2) \ge p(m_1)p(m_2)$$

where $C_1 = \{\{m_1, m_2\}\}$ and $C_2 = \{\{m_1\}, \{m_2\}\}$

We gave an "edge" a + sign between two correlated movies as in the naïve algorithm, but used a graph class instead of matrix.
Looking again at the theorem we induced that the "cheapest" clusters would be those of the form C_k s.t. $\forall m_i, m_j \in C_k$ there is an edge $(i, j) \in E$ in our graph representation of the problem.

Understanding this, we thought that it would be best to try a greedy approach, taking each iteration the biggest clique available in the graph, then remove the clique from the graph and repeat until all the vertices are in clusters.

Trying to think of a way to implement this method, we found a library named Networkx which have many algorithms regarding graphs.

As we predicted, this improvement provided us with lower cost, of about 30-50, from the naïve algorithm (when working only on 100 movies of course).

But later we have found out that this approach is wrong. By looking at the cost function we understood that every cluster, which hold only 1 movie, cost us a lot and the Max Clique left us with many clusters like that (about a half). So, we checked for clusters with maximum size of 2,3,4,5 and found out that the best sizes for clusters are 2 or 3 (the answer wasn't consistent). So, we've implemented that improvement, and cut each clique we have found to maximal size of 3. This improvement results in another 2-3 cost units comparing to the Max Clique.

Main challenge:

When we tried to write this algorithm it was difficult for us to think of a way to implement a method for Max-clique discovering.
Luckily, we found the Networkx library which hold many functions and features for graphs.

Scalability

The algorithm run pretty fast for subsets that contain 100 movies, but when we increased the subsets sizes the algorithm did not scale very well, for subsets of size 1000 the algorithm run pretty slow. for some computers even, a few minutes to run.

Average cost on subsets of size 100: max clique: 692.0016 , max clique≤3 = 689.3597

A Jupyter Notebook file is included with this algorithm

The Linear program algorithm-second attempt for improvement-breakthrough:

On the Max Clique algorithm, we have work in the two first weeks of the semester and got the feelings that we were done, and ready to submit the assignment. Weeks went by and we didn’t try to implement any further improvements, partly because we thought that the theorem mentioned in the assignment, meant that a cluster could not have low cost if there is a movie in it, that is not correlated to all other movies in the cluster. We were wrong, it took us time, we realized that we could probably do better, but didn’t know how.

After a little research we found out that the pivot algorithm is a factor- 3 approximation to the solution of the problem and wondered if we could do better. It is then we’ve come across an online paper that suggested the use of a linear program as a factor-2.06 (**link to explanation**) approximation to the problem. And apparently a lot of other papers suggest similar programs to reach about 2-2.5 approximation.

We have tried to implement the paper’s suggested solution, but the implementation was pretty complex, due to the fact that the linear program has a huge amount of constraints and a very complex objective, that we could not implement without adding more variables. We have tried several Python libraries for Linear programing until we have found the library Plup, that lets the user write very complicated programs with dictionary representation, and **Not** matrix representation as all other libraries.

Our algorithm solves the linear program as presented in the paper and uses the random pivoting suggestion method for building the clusters; e.g. the algorithm chooses a vertex i randomly and for every other vertex j, j will be in the same cluster with i with probability of $1 - x_{i,j}$.

The Pulp library let us implement the paper’s linear program, but to our disappointment this algorithm did not do very well on our testing sets, It gave us Costs of about 730 on subsets of size 100, which is pretty bad.

After some thinking about the problem, we realized that it will be smart to try to model the objective function of the linear program as similar as possible to the cost function. Due to the fact that the cost function takes into account the sizes of the clusters – which is not possible to implement in a linear program- we have decided to minimize the following function:

$$\widehat{cost}(G) = \sum_{(i,j)\in E} x_{i,j} \cdot \log\left(\frac{1}{p(m_i,m_j)}\right) + \lambda \cdot \sum_{(i,j)\notin E} (1 - x_{i,j}) \cdot \log\left(\frac{1}{p(m_i,m_j)}\right)$$

Where:

- $p(m_i, m_j)$ is a probability that someone will rate both m_i and m_j
- $x_{i,j} \in [0,1]$ is a variable of the linear program that represent the probability that the movies i and j will be in the same cluster.
- $\log\left(\frac{1}{p(m_i,m_j)}\right)$ is the cost of the movies i and j according to original cost function, but without dividing by the cluster’s size.
- λ regulazation parameter: we want to ‘punish’ the cost for making clusters out of uncorrelated movies.

We noticed that this cost function is somewhat similar to the expected value of the original cost function; thus, minimizing it make sense.

We have used the constraint from the original program and run the program, the results were much suited to the assignment requirements. We have achieved low costs, between 450-600. This difference is due to the fact that part of the algorithm is random, and we are minimizing the expected value of the cost and not the cost itself.

Interesting results:

After running the algorithm several times, we have examined the results of the linear program and noticed a very interesting phenomenal:

- $x_{i,j}$ for almost every $(i,j) \in E$ was 0.5
- $x_{i,j}$ was 1 for $(i,j) \notin E$

Those observations will lead us to the next algorithm which we have named "The random pivot algorithm"

Main challenge:

As stated, before it was very complicated to find a way to implement the linear program, due to its size. Luckily, we have found the Plup library, that let us do it quite easily.

Scalability

Due to the use of a huge linear program the algorithm is slow, it will take it about 20-30 seconds to run on standard computer with subset's sizes of 100, because part of the algorithm is random. It is smart to run that part several times to achieve the best result. In our implementation the random part runs 10 times and return the best result, but the linear program runs only once. On subsets of size bigger than 100 the algorithm will not run, because the number of variables and constraints needed is very large.$(O(n^3)\ constraints)$.

Average cost on subsets of size 100: 514.30061

The Random pivot algorithm - third attempt for improvement - final algorithm - heuristic approach to the linear program algorithm:

After implementing the second algorithm we were satisfied with the cost results but not satisfied with the running time. By looking of the resulting $x_{i,j}$ of the linear program we have decided to improve the standard pivot algorithm in the following way: the algorithm will choose a random vertex i and for every other vertex j if $(i,j) \in E$ the algorithm will cluster i, j together with probability of 0.5. This decision makes sense because the results of the linear program. Recall that for $(i,j) \in E$ the linear program will result $x_{i,j} = 0.5$ for almost $(i,j) \in E$, and the random pivoting algorithm that comes after the linear program also chooses a random pivot vertex i, and then for every other vertex j , i, j will be in the same cluster with probably $1 - x_{i,j}$ which is 0.5 for $(i,j) \in E$ exactly, as we chosen.

The above algorithm gives good results on our testing sets and run with high speed due to the highly vectorized implementation of the pivoting algorithm, and the vectorized improvements. But sometimes the algorithm gives standard results like the Max Clique algorithm or the pivot algorithm. To tackle this problem, we have decided to run this algorithm 10 times and take the best clustering. This approach is possible because the algorithm is very fast and running it even 100 will take less than a second.

<u>Main challenge:</u>

The algorithm was almost complete when we decided to improve it, recall that it is the regular pivot algorithm, the additions were pretty simple to add, and as mentioned before we continued with the vectorized implementation, so the additions merely added running time.

<u>Scalability</u>

This algorithm is fitted to work with large data sets and can be run on the entire movie data set easily.

<u>Average cost on subsets of size 100:</u> <u>628.3674</u>

<u>A Jupyter Notebook file is included with this algorithm</u>

Cost table on 20 movie subset of size 100

subset #	pivot	max clique	max clique <=2	max clique <=3	random pivot	LP
1	735.3169	697.9060	689.7953	691.6570	653.2412	479.372
2	766.3478	722.9981	718.5549	720.1576	642.2483	601.9602
3	727.7172	685.1987	691.1612	687.5604	611.5929	539.6767
4	775.7222	735.6137	725.5017	728.9858	646.0587	558.0304
5	770.2257	727.4813	727.2303	732.5187	622.5139	541.5064
6	758.5775	713.6869	707.5359	697.0648	668.3732	526.501
7	698.2732	649.9028	648.6551	646.1596	562.9663	447.3303
8	763.9009	715.0987	706.1776	706.2211	653.1165	543.8779
9	722.6606	682.9608	681.4874	679.7038	622.5966	466.6678
10	679.5997	645.2727	646.4236	641.9415	640.3303	554.4635
11	798.8127	747.7042	748.3270	746.7665	654.6001	460.7671
12	704.6530	663.9033	654.8571	659.2691	637.1879	474.9066
13	690.7833	663.9358	659.2699	661.1846	620.4045	536.9274
14	715.3158	672.6604	670.0146	674.4171	605.0285	495.5405
15	749.6032	715.0640	716.6590	713.6069	643.8264	629.0419
16	748.2059	703.3214	693.6193	699.7040	591.8716	361.5472
17	699.8522	656.9912	661.9198	661.1862	636.1283	624.1923
18	748.1191	697.3941	702.5069	704.9288	658.2759	472.4392
19	744.4012	706.2195	707.4848	708.1236	566.4546	457.0223
20	672.9789	636.7184	626.6751	626.0363	630.533	514.2414
Avg:	733.5534	692.0016	689.1928	689.3597	628.3674	514.30061

The table clearly shows that the Linear program algorithm is the best-cost wise. But for large data sets we will prefer the Random pivot algorithm.

example for improved cost on subset:

we have taken one of our subsets (randomsubset16) and those are the results we got:

- Using the pivot algorithm, we have got the following results:

17 Sense and Sensibility (1995), 461 Go Fish (1994), 556 War Room, The (1993), 1187 Passion Fish (1992), 1188 Strictly Ballroom (1992), 1844 Live Flesh (1997), 2175 Dj Vu (1997), 2203 Shadow of a Doubt (1943), 2671 Notting Hill (1999), 3108 Fisher King, The (1991), 3109 River, The (1984)

91 , 544 Striking Distance (1993), 1085 Old Man and the Sea, The (1958), 1506 , 1587 Conan the Barbarian (1982), 1737 , 1766 , 1813 , 2210 Sabotage (1936), 2535 Earthquake (1974), 3137 Sea Wolves, The (1980), 3139 Tarzan the Fearless (1933), 3174 Man on the Moon (1999), 3634 Seven Days in May (1964), 3738 Sugarland Express, The (1974)

200 Tie That Binds, The (1995), 504 No Escape (1994), 2327 Tales from the Darkside: The Movie (1990), 2328 Vampires (1998), 3572 Carnosaur (1993), 3576 Hidden, The (1987), 3917 Hellraiser (1987)

107 Muppet Treasure Island (1996), 945 Top Hat (1935), 3028 Taming of the Shrew, The (1967), 3375 Destination Moon (1950), 3379 On the Beach (1959), 3549 Guys and Dolls (1955)

239 Goofy Movie, A (1995), 747 Stupids, The (1996), 1431 Beverly Hills Ninja (1997), 1485 Liar Liar (1997), 1772 Blues Brothers 2000 (1998), 2016 Apple Dumpling Gang Rides Again, The (1979)

19 Ace Ventura: When Nature Calls (1995), 101 Bottle Rocket (1996), 205 Unstrung Heroes (1995), 214 Before the Rain (Pred dozhdot) (1994), 1854 Kissing a Fool (1998), 2329 American History X (1998), 2890 Three Kings (1999)

2858 American Beauty (1999)

283 New Jersey Drive (1995), 1361 Paradise Lost: The Child Murders at Robin Hood Hills (1996), 1889 Insomnia (1997), 2627 Endurance (1998)

1103 Rebel Without a Cause (1955), 2398 Miracle on 34th Street (1947), 3095 Grapes of Wrath, The (1940), 3457 Waking the Dead (1999)

1179 Grifters, The (1990), 2677 Buena Vista Social Club (1999)

3301 Whole Nine Yards, The (2000)

2386 Jerry Springer: Ringmaster (1998), 3203 Dead Calm (1989)

345 Adventures of Priscilla, Queen of the Desert, The (1994), 347 Bitter Moon (1992), 2388 Steam: The Turkish Bath (Hamam) (1997), 2417 Heartburn (1986)

2284 Bandit Queen (1994)

1394 Raising Arizona (1987)

1132 Manon of the Spring (Manon des sources) (1986), 2964 Julien Donkey-Boy (1999)

255 Jerky Boys, The (1994), 3385 Volunteers (1985)

2211 Secret Agent (1936)

1909 X-Files: Fight the Future, The (1998)

377 Speed (1994)

688.9826934577599

- Using the improved algorithm we have got the following results:

2388 Steam: The Turkish Bath (Hamam) (1997), 2964 Julien Donkey-Boy (1999)

377 Speed (1994), 2671 Notting Hill (1999)

1909 X-Files: Fight the Future, The (1998)

1587 Conan the Barbarian (1982)

239 Goofy Movie, A (1995), 2016 Apple Dumpling Gang Rides Again, The (1979)

3137 Sea Wolves, The (1980), 3203 Dead Calm (1989)

3576 Hidden, The (1987), 3917 Hellraiser (1987)

2203 Shadow of a Doubt (1943)

2210 Sabotage (1936), 3095 Grapes of Wrath, The (1940), 3375 Destination Moon (1950)

1394 Raising Arizona (1987), 2386 Jerry Springer: Ringmaster (1998)

1431 Beverly Hills Ninja (1997), 2328 Vampires (1998)

3174 Man on the Moon (1999)

3108 Fisher King, The (1991)

2858 American Beauty (1999)

945 Top Hat (1935)

3738 Sugarland Express, The (1974)

107 Muppet Treasure Island (1996), 1772 Blues Brothers 2000 (1998), 3301 Whole Nine Yards, The (2000)

1132 Manon of the Spring (Manon des sources) (1986), 1188 Strictly Ballroom (1992)

2890 Three Kings (1999), 3457 Waking the Dead (1999)

200 Tie That Binds, The (1995), 205 Unstrung Heroes (1995), 255 Jerky Boys, The (1994), 283 New Jersey Drive (1995), 461 Go Fish (1994), 504 No Escape (1994), 544 Striking Distance (1993), 1085 Old Man and the Sea, The (1958), 1737 , 1766 , 1854 Kissing a Fool (1998), 2327 Tales from the Darkside: The Movie (1990), 3109 River, The (1984), 3139 Tarzan the Fearless (1933), 3379 On the Beach (1959), 3385 Volunteers (1985), 3572 Carnosaur (1993)

17 Sense and Sensibility (1995), 214 Before the Rain (Pred dozhdot) (1994), 345 Adventures of Priscilla, Queen of the Desert, The (1994), 347 Bitter Moon (1992), 556 War Room, The (1993), 1103 Rebel Without a Cause (1955), 1179 Grifters, The (1990), 1187 Passion Fish (1992), 1506 , 2329 American History X (1998), 2417 Heartburn (1986), 2677 Buena Vista Social Club (1999)

91 , 101 Bottle Rocket (1996), 1361 Paradise Lost: The Child Murders at Robin Hood Hills (1996), 1813 , 1844 Live Flesh (1997), 1889 Insomnia (1997), 2175 Dj Vu (1997), 2211 Secret Agent (1936), 2284 Bandit Queen (1994), 2627 Endurance (1998)

19 Ace Ventura: When Nature Calls (1995), 747 Stupids, The (1996), 1485 Liar Liar (1997)

2398 Miracle on 34th Street (1947), 2535 Earthquake (1974), 3028 Taming of the Shrew, The (1967), 3549 Guys and Dolls (1955), 3634 Seven Days in May (1964)

234.16284039340135

As we can see the cost was improved by more then 450.

interesting subsets

1. Subset1**:**

this subset includes only movies from the comedy category, this subsest is interesting because we think people who like to watch comedy movies would rate a lot of comedy movies, and give them good ratings, resulting in high correlation between those movies.

- Using the pivot algorithm we have got the following results:

```
5 Father of the Bride Part II (1995), 19 Ace Ventura: When Nature Calls (1995), 38 It Takes Two (1995), 63 Don't Be a Menace to South Central While Drinking Your Juice, 65 Bio-Dome (1996), 88 Black Sheep (1996), 101 Bottle Rocket (1996), 102 Mr. Wrong (1996), 104 Happy Gilmore (1996), 119 Steal Big, Steal Little (1995), 125 Flirting With Disaster (1996), 135 Down Periscope (1996), 141 Birdcage, The (1996), 144 Brothers McMullen, The (1995), 156 Blue in the Face (1995), 171 Jeffrey (1995), 174 Jury Duty (1995), 176 Living in Oblivion (1995), 178 Love & Human Remains (1993), 180 Mallrats (1995), 186 Nine Months (1995), 187 Party Girl (1995), 189 Reckless (1995), 203 To Wong Foo, Thanks for Everything! Julie Newmar (1995), 216 Billy Madison (1995), 228 Destiny Turns on the Radio (1995), 231 Dumb & Dumber (1994), 234 Exit to Eden (1994), 243 Gordy (1995), 248 Houseguest (1994), 255 Jerky Boys, The (1994), 267 Major Payne (1994), 274 Man of the House (1995), 275 Mixed Nuts (1994), 278 Miami Rhapsody (1995), 305 Ready to Wear (Pret-A-Porter) (1994), 312 Stuart Saves His Family (1995), 324 Sum of Us, The (1994), 325 National Lampoon's Senior Trip (1995), 333 Tommy Boy (1995), 344 Ace Ventura: Pet Detective (1994), 370 Naked Gun 33 1/3: The Final Insult (1994), 410 Addams Family Values (1993), 411 You So Crazy (1994), 413 Airheads (1994), 414 Air Up There, The (1994), 419 Beverly Hillbillies, The (1993), 429 Cabin Boy (1994), 433 Clean Slate (1994), 437 Cops and Robbersons (1994), 441 Dazed and Confused (1993), 445 Fatal Instinct (1993), 453 For Love or Money (1993), 460 Getting Even with Dad (1994), 467 Live Nude Girls (1995), 478 Jimmy Hollywood (1994), 486 Life with Mikey (1993), 489 Made in America (1993), 500 Mrs. Doubtfire (1993), 505 North (1994), 514 Ref, The (1994), 518 Road to Wellville, The (1994), 520 Robin Hood: Men in Tights (1993), 542 Son in Law (1993), 564 Chasers (1994), 585 Brady Bunch Movie, The (1995), 603 Bye Bye, Love (1995), 612 Pallbearer, The (1996), 619 Ed (1996), 626 Thin Line Between Love and Hate, A (1996), 637 Sgt. Bilko (1996), 656 Eddie (1996), 663 Kids in the Hall: Brain Candy (1996)
352 Crooklyn (1994), 470 House Party 3 (1994)
52 Mighty Aphrodite (1995), 96 In the Bleak Midwinter (1995), 298 Pushing Hands (1992), 348 Bullets Over Broadway (1994), 449 Fear of a Black Hat (1993), 633 Denise Calls Up (1995), 639 Girl 6 (1996)
212 Bushwhacked (1995)
69 Friday (1995), 223 Clerks (1994)
750.2192588981404
```

As expected we got few clusters with the first one a hugh cluster.

- Using the improved algorithm we have got the following results:

```
449 Fear of a Black Hat (1993)
223 Clerks (1994), 663 Kids in the Hall: Brain Candy (1996)
333 Tommy Boy (1995), 429 Cabin Boy (1994), 505 North (1994)
101 Bottle Rocket (1996), 144 Brothers McMullen, The (1995), 180 Mallrats (1995), 419 Beverly Hillbillies, The (1993), 467 Live Nude Girls (1995)
104 Happy Gilmore (1996), 186 Nine Months (1995), 312 Stuart Saves His Family (1995)
69 Friday (1995), 125 Flirting With Disaster (1996), 141 Birdcage, The (1996), 156 Blue in the Face (1995), 176 Living in Oblivion (1995), 243 Gordy (1995), 298 Pushing Hands (1992), 344 Ace Ventura: Pet Detective (1994), 441 Dazed and Confused (1993), 518 Road to Wellville, The (1994)
19 Ace Ventura: When Nature Calls (1995), 96 In the Bleak Midwinter (1995), 171 Jeffrey (1995), 178 Love & Human Remains (1993), 189 Reckless (1995), 305 Ready to Wear (Pret-A-Porter) (1994), 324 Sum of Us, The (1994), 348 Bullets Over Broadway (1994), 352 Crooklyn (1994), 410 Addams Family Values (1993), 633 Denise Calls Up (1995), 639 Girl 6 (1996)
52 Mighty Aphrodite (1995), 63 Don't Be a Menace to South Central While Drinking Your Juice, 187 Party Girl (1995), 216 Billy Madison (1995), 411 You So Crazy (1994), 445 Fatal Instinct (1993), 612 Pallbearer, The (1996)
5 Father of the Bride Part II (1995), 135 Down Periscope (1996), 203 To Wong Foo, Thanks for Everything! Julie Newmar (1995), 231 Dumb & Dumber (1994), 234 Exit to Eden (1994), 248 Houseguest (1994), 274 Man of the House (1995), 275 Mixed Nuts (1994), 433 Clean Slate (1994), 460 Getting Even with Dad (1994), 470 House Party 3 (1994), 603 Bye Bye, Love (1995)
88 Black Sheep (1996), 102 Mr. Wrong (1996), 119 Steal Big, Steal Little (1995), 278 Miami Rhapsody (1995), 413 Airheads (1994), 437 Cops and Robbersons (1994), 478 Jimmy Hollywood (1994), 514 Ref, The (1994), 520 Robin Hood: Men in Tights (1993), 564 Chasers (1994)
38 It Takes Two (1995), 65 Bio-Dome (1996), 174 Jury Duty (1995), 212 Bushwhacked (1995), 228 Destiny Turns on the Radio (1995), 255 Jerky Boys, The (1994), 267 Major Payne (1994), 325 National Lampoon's Senior Trip (1995), 370 Naked Gun 33 1/3: The Final Insult (1994), 414 Air Up There, The (1994), 453 For Love or Money (1993), 486 Life with Mikey (1993), 489 Made in America (1993), 500 Mrs. Doubtfire (1993), 542 Son in Law (1993), 585 Brady Bunch Movie, The (1995), 619 Ed (1996), 626 Thin Line Between Love and Hate, A (1996), 637 Sgt. Bilko (1996), 656 Eddie (1996)
366.1265861150464
```

Now we got many clusters and much better cost.

2. Subset2:

this subset includes only movies which user with id 8 have rated, this subset is interesting because we assume that if one user liked those movies there should be some correlation between many of them.

- Using the pivot algorithm we have got the following results:

```
17 Sense and Sensibility (1995), 25 Leaving Las Vegas (1995), 36 Dead Man Walking (1995), 58 Postino,
Il (The Postman) (1994), 105 Bridges of Madison County, The (1995), 150 Apollo 13 (1995), 282 Nell
(1994), 337 What's Eating Gilbert Grape (1993), 506 Orlando (1993), 508 Philadelphia (1993), 650 Moll
Flanders (1996), 1120 People vs. Larry Flynt, The (1996), 1393 Jerry Maguire (1996), 1660 Eve's Bayou
(1997), 1678 Joy Luck Club, The (1993), 1682 Truman Show, The (1998), 1693 Amistad (1997), 1704
Good Will Hunting (1997), 1810 Primary Colors (1998), 2268 Few Good Men, A (1992), 2314 Beloved
(1998), 2320 Apt Pupil (1998), 2329 American History X (1998), 2336 Elizabeth (1998)
14 Nixon (1995), 16 Casino (1995), 161 Crimson Tide (1995), 230 Dolores Claiborne (1994), 269 My
Crazy Life (Mi vida loca) (1993), 454 Firm, The (1993), 510 Poetic Justice (1993), 1411 Hamlet (1996),
1466 Donnie Brasco (1997), 1589 Cop Land (1997), 1621 Soul Food (1997), 1639 Chasing Amy (1997),
1701 Deconstructing Harry (1997), 1735 Great Expectations (1998), 2278 Ronin (1998), 2427 Thin Red
Line, The (1998)
163 Desperado (1995), 288 Natural Born Killers (1994), 349 Clear and Present Danger (1994), 377
Speed (1994), 393 Street Fighter (1994), 733 Rock, The (1996), 1488 Devil's Own, The (1997), 1573
Face/Off (1997), 1801 Man in the Iron Mask, The (1998), 2006 Mask of Zorro, The (1998)
4 Waiting to Exhale (1995), 265 Like Water for Chocolate (Como agua para chocolate) (1992), 345
Adventures of Priscilla, Queen of the Desert, The (1994), 465 Heaven & Earth (1993), 476 Inkwell, The
(1994), 562 Welcome to the Dollhouse (1995), 1059 William Shakespeare's Romeo and Juliet (1996),
1277 Cyrano de Bergerac (1990), 1357 Shine (1996), 2297 What Dreams May Come (1998), 2324 Life Is
Beautiful (La Vita  bella) (1997)
110 Braveheart (1995), 266 Legends of the Fall (1994), 480 Jurassic Park (1993), 524 Rudy (1993), 538
Six Degrees of Separation (1993), 589 Terminator 2: Judgment Day (1991), 908 North by Northwest
(1959), 1580 Men in Black (1997), 1673 Boogie Nights (1997), 2023 Godfather: Part III, The (1990),
2291 Edward Scissorhands (1990)
253 Interview with the Vampire (1994)
111 Taxi Driver (1976), 296 Pulp Fiction (1994), 527 Schindler's List (1993), 555 True Romance (1993),
608 Fargo (1996), 1213 GoodFellas (1990), 1730 Kundun (1997)
1210 Star Wars: Episode VI - Return of the Jedi (1983)
73 Misrables, Les (1995)
741 Ghost in the Shell (Kokaku kidotai) (1995), 1274 Akira (1988), 1653 Gattaca (1997)
42 Dead Presidents (1995), 1027 Robin Hood: Prince of Thieves (1991), 1476 Private Parts (1997), 1711
Midnight in the Garden of Good and Evil (1997), 1836 Last Days of Disco, The (1998), 1840 He Got
Game (1998), 1916 Buffalo 66 (1998)
1 Toy Story (1995), 39 Clueless (1995), 1265 Groundhog Day (1993), 2396 Shakespeare in Love (1998)
151 Rob Roy (1995)
2028 Saving Private Ryan (1998)
1721 Titanic (1997)
24 Powder (1995)
681.4013755860226
```

As expected we got 5 big clusters which implies on high correlation between many movies.

- Using the improved algorithm we have got the following results:

```
105 Bridges of Madison County, The (1995), 265 Like Water for Chocolate (Como agua para chocolate)
(1992), 1059 William Shakespeare's Romeo and Juliet (1996), 1357 Shine (1996)
2336 Elizabeth (1998)
151 Rob Roy (1995), 377 Speed (1994)
253 Interview with the Vampire (1994), 454 Firm, The (1993), 1639 Chasing Amy (1997), 1678 Joy Luck Club,
The (1993), 1810 Primary Colors (1998), 2297 What Dreams May Come (1998)
1660 Eve's Bayou (1997)
908 North by Northwest (1959)
150 Apollo 13 (1995)
296 Pulp Fiction (1994), 608 Fargo (1996), 2329 American History X (1998)
39 Clueless (1995), 345 Adventures of Priscilla, Queen of the Desert, The (1994), 1027 Robin Hood: Prince of
Thieves (1991), 2396 Shakespeare in Love (1998)
111 Taxi Driver (1976), 1701 Deconstructing Harry (1997), 1840 He Got Game (1998)
349 Clear and Present Danger (1994)
110 Braveheart (1995), 266 Legends of the Fall (1994), 1673 Boogie Nights (1997), 2028 Saving Private Ryan
(1998)
16 Casino (1995), 161 Crimson Tide (1995), 1682 Truman Show, The (1998), 2268 Few Good Men, A (1992),
2427 Thin Red Line, The (1998)
480 Jurassic Park (1993), 524 Rudy (1993), 589 Terminator 2: Judgment Day (1991), 1580 Men in Black
(1997), 2291 Edward Scissorhands (1990)
25 Leaving Las Vegas (1995), 58 Postino, Il (The Postman) (1994), 230 Dolores Claiborne (1994), 337 What's
Eating Gilbert Grape (1993), 506 Orlando (1993), 527 Schindler's List (1993), 538 Six Degrees of Separation
(1993), 562 Welcome to the Dollhouse (1995), 650 Moll Flanders (1996), 1213 GoodFellas (1990), 1277
Cyrano de Bergerac (1990), 1693 Amistad (1997), 1721 Titanic (1997), 1730 Kundun (1997)
1210 Star Wars: Episode VI - Return of the Jedi (1983)
741 Ghost in the Shell (Kokaku kidotai) (1995), 1274 Akira (1988)
1 Toy Story (1995), 1265 Groundhog Day (1993), 2324 Life Is Beautiful (La Vita  bella) (1997)
17 Sense and Sensibility (1995), 36 Dead Man Walking (1995), 73 Misrables, Les (1995), 269 My Crazy Life
(Mi vida loca) (1993), 508 Philadelphia (1993), 1393 Jerry Maguire (1996), 1411 Hamlet (1996), 1704 Good
Will Hunting (1997)
733 Rock, The (1996), 2320 Apt Pupil (1998)
14 Nixon (1995), 163 Desperado (1995), 1735 Great Expectations (1998), 1916 Buffalo 66 (1998), 2006 Mask
of Zorro, The (1998)
24 Powder (1995), 393 Street Fighter (1994)
4 Waiting to Exhale (1995), 42 Dead Presidents (1995), 282 Nell (1994), 288 Natural Born Killers (1994), 465
Heaven & Earth (1993), 476 Inkwell, The (1994), 510 Poetic Justice (1993), 555 True Romance (1993), 1120
People vs. Larry Flynt, The (1996), 1466 Donnie Brasco (1997), 1476 Private Parts (1997), 1488 Devil's Own,
The (1997), 1573 Face/Off (1997), 1589 Cop Land (1997), 1621 Soul Food (1997), 1711 Midnight in the
Garden of Good and Evil (1997), 1801 Man in the Iron Mask, The (1998), 1836 Last Days of Disco, The
(1998), 2023 Godfather: Part III, The (1990), 2278 Ronin (1998), 2314 Beloved (1998)
1653 Gattaca (1997)
513.9984343392448
```

Which have only 2 big clusters.

3. Subset3:

This subset includes movies from both romance and action genres.
this subset is interesting because we expect there to be mainly 2 big clusters (one for each genre) as explained for subset 1. Also we expect there to be much more clusters then for subset 1 because the movies are from 2 different genres

- Using the pivot algorithm we have got the following results:

9 Sudden Death (1995), 20 Money Train (1995), 71 Fair Game (1995), 145 Bad Boys (1995), 204 Under Siege 2: Dark Territory (1995), 227 Drop Zone (1994), 251 Hunted, The (1995), 315 Specialist, The (1994), 384 Bad Company (1995), 393 Street Fighter (1994), 394 Coldblooded (1995), 459 Getaway, The (1994), 479 Judgment Night (1993), 533 Shadow, The (1994), 544 Striking Distance (1993), 548 Terminal Velocity (1994), 667 Bloodsport 2 (1995), 694 Substitute, The (1996), 886 Bulletproof (1996), 1170 Best of the Best 3: No Turning Back (1995), 1181 Shooter, The (1995), 1385 Under Siege (1992), 1429 Jackie Chan's First Strike (1996), 1432 Metro (1997), 1497 Double Team (1997), 1520 Commandments (1997), 1599 Steel (1997), 2196 Knock Off (1998), 2292 Overnight Delivery (1996), 2340 Meet Joe Black (1998), 2347 Pope of Greenwich Village, The (1984), 2376 View to a Kill, A (1985), 2403 First Blood (1982), 2535 Earthquake (1974), 2625 Black Mask (Hak hap) (1996), 2737 Assassination (1987), 2756 Wanted: Dead or Alive (1987), 2965 Omega Code, The (1999), 2989 For Your Eyes Only (1981), 2990 Licence to Kill (1989), 2991 Live and Let Die (1973), 3197 Presidio, The (1988), 3274 Single White Female (1992), 3283 Minnie and Moskowitz (1971), 3442 Band of the Hand (1986), 3444 Bloodsport (1988)

638 Jack and Sarah (1995), 803 Walking and Talking (1996), 1458 Touch (1997), 1574 Fall (1997), 1656 Swept from the Sea (1997), 1666 Hugo Pool (1997), 1749 Leading Man, The (1996), 1755 Shooting Fish (1997), 1835 City of Angels (1998), 2157 Chambermaid on the Titanic, The (1998), 2998 Dreaming of Joseph Lees (1998)

28 Persuasion (1995), 932 Affair to Remember, An (1957), 1477 Love Jones (1997), 1493 Love and Other Catastrophes (1996), 1502 Kissed (1996), 1669 Tango Lesson, The (1997), 1684 Mrs. Dalloway (1997), 2497 Message in a Bottle (1999), 2708 Autumn Tale, An (Conte d'automne) (1998), 2721 Trick (1999), 3414 Love Is a Many-Splendored Thing (1955)

2534 Avalanche (1978), 2947 Goldfinger (1964), 2993 Thunderball (1965), 3310 Kid, The (1921), 3624 Shanghai Noon (2000)

1475 Kama Sutra: A Tale of Love (1996), 2949 Dr. No (1962), 3384 Taking of Pelham One Two Three, The (1974)

2948 From Russia with Love (1963)

876 Police Story 4: Project S (Chao ji ji hua) (1993)

3206 Against All Odds (1984)

707.2266355640986

As expected we got few big clusters (we expected 2 but we got 1 huge and another 2 big clusters).

- Using the improved algorithm we have got the following results:

204 Under Siege 2: Dark Territory (1995), 251 Hunted, The (1995), 2949 Dr. No (1962)

667 Bloodsport 2 (1995), 2534 Avalanche (1978)

1170 Best of the Best 3: No Turning Back (1995)

1666 Hugo Pool (1997), 1835 City of Angels (1998), 2292 Overnight Delivery (1996), 2340 Meet Joe Black (1998)

2998 Dreaming of Joseph Lees (1998)

876 Police Story 4: Project S (Chao ji ji hua) (1993)

2948 From Russia with Love (1963)

9 Sudden Death (1995), 2347 Pope of Greenwich Village, The (1984), 2376 View to a Kill, A (1985), 2737 Assassination (1987), 2990 Licence to Kill (1989)

932 Affair to Remember, An (1957), 3206 Against All Odds (1984)

315 Specialist, The (1994), 544 Striking Distance (1993), 2196 Knock Off (1998), 2403 First Blood (1982), 2625 Black Mask (Hak hap) (1996), 2756 Wanted: Dead or Alive (1987)

227 Drop Zone (1994), 394 Coldblooded (1995), 533 Shadow, The (1994), 1520 Commandments (1997), 2708 Autumn Tale, An (Conte d'automne) (1998), 2989 For Your Eyes Only (1981), 2993 Thunderball (1965), 3442 Band of the Hand (1986), 3444 Bloodsport (1988)

638 Jack and Sarah (1995)

3414 Love Is a Many-Splendored Thing (1955)

20 Money Train (1995), 145 Bad Boys (1995), 384 Bad Company (1995), 393 Street Fighter (1994), 459 Getaway, The (1994), 479 Judgment Night (1993), 548 Terminal Velocity (1994), 694 Substitute, The (1996), 886 Bulletproof (1996), 1181 Shooter, The (1995), 1385 Under Siege (1992), 1429 Jackie Chan's First Strike (1996), 1497 Double Team (1997), 1599 Steel (1997), 2535 Earthquake (1974), 3197 Presidio, The (1988)

1458 Touch (1997), 2157 Chambermaid on the Titanic, The (1998), 3274 Single White Female (1992)

1432 Metro (1997), 3310 Kid, The (1921), 3624 Shanghai Noon (2000)

3283 Minnie and Moskowitz (1971)

28 Persuasion (1995), 803 Walking and Talking (1996), 1477 Love Jones (1997), 1493 Love and Other Catastrophes (1996), 1574 Fall (1997), 1755 Shooting Fish (1997), 2721 Trick (1999), 2965 Omega Code, The (1999)

71 Fair Game (1995), 1475 Kama Sutra: A Tale of Love (1996), 1502 Kissed (1996), 1656 Swept from the Sea (1997), 1669 Tango Lesson, The (1997), 1684 Mrs. Dalloway (1997), 1749 Leading Man, The (1996), 2497 Message in a Bottle (1999), 2947 Goldfinger (1964), 2991 Live and Let Die (1973), 3384 Taking of Pelham One Two Three, The (1974)

526.543579394175

Cost table on the 3 subsets each is of size 100

	Pivot algorithm	LP algorithm
Subset1	750.2192	366.1265
Subset2	681.4013	513.9984
Subset3	707.2266	526.5435