Spark Lab Session

Initialize the environment

In [5]:

```
import findspark
findspark.init()

import pyspark
import random

sc = pyspark.SparkContext(appName="Pi")
```

Compute the list L of integers, with L = { 0 ... 499 }

In [6]:

```
ints = sc.parallelize(range(3000))
ints.take(20)
```

Out[6]:

```
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]
```

Compute the list C = { x^3 | x ∈ L } and then sum of elements in C

In [7]:

```
cubes = ints.map(lambda x: x*x*x)
```

In [8]:

```
cubes.reduce(lambda x, y: x+y)
cubes.sum()
```

Out[8]:

```
20236502250000
```

What is the repartition for the last digits of integers in C? i.e. how many end with a 0? with a 1? etc.

In [9]:

```
# Première étape : calculer la liste contenant
# les derniers chiffres
lastDigits = cubes.map(lambda x: x%10)

# Deuxième étape : compter combien de fois chaque
# items apparait
countLastDigits = lastDigits.map(lambda x:(x,1)) \
                            .reduceByKey(lambda x,y: x+y)
```

In [10]:

```
cubes.map(lambda x:(x % 10,1)).reduceByKey(lambda x, y: x+y).collect()
```

Out[10]:

```
[(0, 300),
 (8, 300),
 (1, 300),
 (9, 300),
 (2, 300),
 (3, 300),
 (4, 300),
 (5, 300),
 (6, 300),
 (7, 300)]
```

In [14]:

```
#other method

def myFold(acc, element):
    new_acc = acc.copy()
    new_acc[element]+=1
    return new_acc

def myReduce(acc1, acc2):
    new_acc = []
    for i in range(10):
        new_acc.append(acc1[i]+acc2[i])
    return new_acc

val_init = [0]*10


myReduce(myFold(val_init,1),myFold(val_init,2))

myFold(myFold(val_init,1),2)
lastDigits = cubes.map(lambda x:x%10)

lastDigits.aggregate(val_init,myFold,myReduce)
```

Out[14]:

```
[300, 300, 300, 300, 300, 300, 300, 300, 300, 300]
```

What is the repartition of digits for the integers in C?

In [12]:

```
cubes.flatMap(lambda x: [ (e,1) for e in str(x)]).reduceByKey(lambda x, y: x+y).collect
()

#cubes.flatMap(lambda x: [int(e) for e in str(x)]).aggregate(val_init,myFold,myReduce)
```

Out[12]:

```
[('4', 2762),
 ('7', 2787),
 ('0', 3127),
 ('1', 3667),
 ('8', 2639),
 ('9', 2521),
 ('5', 2653),
 ('6', 2713),
 ('3', 2814),
 ('2', 3294)]
```

Computation of π

To compute the value of π, you will generate the list of all pairs (x,y) of integers from 0 to K. Then you will compute the number of such pairs such that (2x+1)^2+(2y+1)^2 is less then (2*K)^2. The ratio between the number of such pairs and the number of total pairs is an approximation of π. For K=3000 you should obtain a value close to 3.14159.

In [13]:

```
K=1000
intUpToK = sc.parallelize(range(K))
pairs = intUpToK.cartesian(intUpToK)
nbTotal = pairs.count()

def isOk(v):
    x,y = v
    return (2*x+1)**2+(2*y+1)**2 <= 4*K*K

nbOk = pairs.filter(isOk).count()
print(4*float(nbOk)/nbTotal)
print(nbOk)
```

```
3.141676
785419
```

Readings files into RDD

In [14]:

```
import re
future_pattern = re.compile("""([^,"]+|"[^"]+")(?=,|$)""")

def parseCSV(line):
     return future_pattern.findall(line)


ratingsFile = sc.textFile("/home/jachiet/Downloads/ml-latest-small/ratings.csv")
moviesFile = sc.textFile("/home/jachiet/Downloads/ml-latest-small/movies.csv")
```

In [15]:

```
ratings = ratingsFile.map(parseCSV).filter(lambda x: x[0]!="userId")
movies = moviesFile.map(parseCSV).filter(lambda x:x[0]!="movieId")
movies.take(30)
```

Out[15]:

```
[['1', 'Toy Story (1995)', 'Adventure|Animation|Children|Comedy|Fantasy'],
 ['2', 'Jumanji (1995)', 'Adventure|Children|Fantasy'],
 ['3', 'Grumpier Old Men (1995)', 'Comedy|Romance'],
 ['4', 'Waiting to Exhale (1995)', 'Comedy|Drama|Romance'],
 ['5', 'Father of the Bride Part II (1995)', 'Comedy'],
 ['6', 'Heat (1995)', 'Action|Crime|Thriller'],
 ['7', 'Sabrina (1995)', 'Comedy|Romance'],
 ['8', 'Tom and Huck (1995)', 'Adventure|Children'],
 ['9', 'Sudden Death (1995)', 'Action'],
 ['10', 'GoldenEye (1995)', 'Action|Adventure|Thriller'],
 ['11', '"American President, The (1995)"', 'Comedy|Drama|Romance'],
 ['12', 'Dracula: Dead and Loving It (1995)', 'Comedy|Horror'],
 ['13', 'Balto (1995)', 'Adventure|Animation|Children'],
 ['14', 'Nixon (1995)', 'Drama'],
 ['15', 'Cutthroat Island (1995)', 'Action|Adventure|Romance'],
 ['16', 'Casino (1995)', 'Crime|Drama'],
 ['17', 'Sense and Sensibility (1995)', 'Drama|Romance'],
 ['18', 'Four Rooms (1995)', 'Comedy'],
 ['19', 'Ace Ventura: When Nature Calls (1995)', 'Comedy'],
 ['20', 'Money Train (1995)', 'Action|Comedy|Crime|Drama|Thriller'],
 ['21', 'Get Shorty (1995)', 'Comedy|Crime|Thriller'],
 ['22', 'Copycat (1995)', 'Crime|Drama|Horror|Mystery|Thriller'],
 ['23', 'Assassins (1995)', 'Action|Crime|Thriller'],
 ['24', 'Powder (1995)', 'Drama|Sci-Fi'],
 ['25', 'Leaving Las Vegas (1995)', 'Drama|Romance'],
 ['26', 'Othello (1995)', 'Drama'],
 ['27', 'Now and Then (1995)', 'Children|Drama'],
 ['28', 'Persuasion (1995)', 'Drama|Romance'],
 ['29',
  '"City of Lost Children, The (Cité des enfants perdus, La) (1995)"',
  'Adventure|Drama|Fantasy|Mystery|Sci-Fi'],
 ['30', 'Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)', 'Crime|Dra
ma']]
```

Compute the 10 best rated movies

Try the following functions to determine the average ratings:

1. sum(ratings)/numberOfRatings
2. sum(ratings)/(1+numberOfRatings)
3. sum(ratings)/max(20,numberOfRatings)

In [16]:

```
ratedMovie = ratings.map(lambda x: (x[1],(float(x[2]),1))).reduceByKey(lambda x,y: (x[0
]+y[0],x[1]+y[1]))
movieJoin = movies.map(lambda x: (x[0],x[1]))

ratedMovie3=ratedMovie.mapValues(lambda x:math.log(x[1])*(x[0])/(x[1]))
movieJoin.join(ratedMovie3).map(lambda x: x[1]).sortBy(lambda x: -x[1]).take(10)
```

Out[16]:

```
[('"Shawshank Redemption, The (1994)"', 25.506303124680446),
 ('Forrest Gump (1994)', 24.135559630846686),
 ('Pulp Fiction (1994)', 24.035971735394472),
 ('"Matrix, The (1999)"', 23.593497870526754),
 ('"Silence of the Lambs, The (1991)"', 23.43310709209536),
 ('Star Wars: Episode IV - A New Hope (1977)', 23.37860964684444),
 ('Fight Club (1999)', 23.00760161003686),
 ("Schindler's List (1993)", 22.78807638333873),
 ('Star Wars: Episode V - The Empire Strikes Back (1980)', 22.561506207236
782),
 ('"Godfather, The (1972)"', 22.549726244087907)]
```

In [17]:

```
ratedMovies = \
    ratings.map(lambda x: (x[1],(float(x[2]),1))) \
           .reduceByKey(lambda x,y: (x[0]+y[0],x[1]+y[1])) \
           .mapValues(lambda x: x[0]/x[1])


ratedMovies.take(3)
```

Out[17]:

```
[('1', 3.9209302325581397),
 ('50', 4.237745098039215),
 ('70', 3.5090909090909093)]
```

In [18]:

```
movieJoin = movies.map(lambda x: (x[0],x[1]))

moviesWithAvg = movieJoin.join(ratedMovies) \
                         .map(lambda x: x[1])

bestMovies = moviesWithAvg.sortBy(lambda x: x[1],ascending=False)
bestMovies.take(10)
```

Out[18]:

```
[('Lamerica (1994)', 5.0),
 ('What Happened Was... (1994)', 5.0),
 ('Denise Calls Up (1995)', 5.0),
 ('Lesson Faust (1994)', 5.0),
 ('"Sandpiper, The (1965)"', 5.0),
 ('My Man Godfrey (1957)', 5.0),
 ('Black Tar Heroin: The Dark End of the Street (2000)', 5.0),
 ('Slumber Party Massacre II (1987)', 5.0),
 ('Moscow Does Not Believe in Tears (Moskva slezam ne verit) (1979)', 5.
0),
 ('Cherish (2002)', 5.0)]
```

In [19]:

```
#with log metric

import math

ratedMovies = \
    ratings.map(lambda x: (x[1],(float(x[2]),1))) \
           .reduceByKey(lambda x,y: (x[0]+y[0],x[1]+y[1])) \
           .mapValues(lambda x: math.log(x[1])*x[0]/(x[1]))

movieJoin = movies.map(lambda x: (x[0],x[1]))

moviesWithAvg = movieJoin.join(ratedMovies) \
                         .map(lambda x: x[1])

bestMovies2 = moviesWithAvg.sortBy(lambda x: x[1],ascending=False)
bestMovies2.take(10)
```

Out[19]:

```
[('"Shawshank Redemption, The (1994)"', 25.506303124680446),
 ('Forrest Gump (1994)', 24.135559630846686),
 ('Pulp Fiction (1994)', 24.035971735394472),
 ('"Matrix, The (1999)"', 23.593497870526754),
 ('"Silence of the Lambs, The (1991)"', 23.43310709209536),
 ('Star Wars: Episode IV - A New Hope (1977)', 23.37860964684444),
 ('Fight Club (1999)', 23.00760161003686),
 ("Schindler's List (1993)", 22.78807638333873),
 ('Star Wars: Episode V - The Empire Strikes Back (1980)', 22.561506207236
782),
 ('"Godfather, The (1972)"', 22.549726244087907)]
```

In [20]:

```
# with one 0 vote

ratedMovie2=ratedMovie.mapValues(lambda x:(x[0])/((1+x[1])))
movieJoin.join(ratedMovie2).map(lambda x: x[1]).sortBy(lambda x: -x[1]).take(10)
```

Out[20]:

```
[('"Shawshank Redemption, The (1994)"', 4.415094339622642),
 ('"Godfather, The (1972)"', 4.266839378238342),
 ('"Streetcar Named Desire, A (1951)"', 4.261904761904762),
 ('Fight Club (1999)', 4.2534246575342465),
 ('"Godfather: Part II, The (1974)"', 4.226923076923077),
 ('Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1
964)',
  4.224489795918367),
 ('"Three Billboards Outside Ebbing, Missouri (2017)"', 4.22222222222222
2),
 ('"Usual Suspects, The (1995)"', 4.217073170731707),
 ('Goodfellas (1990)', 4.216535433070866),
 ('Star Wars: Episode IV - A New Hope (1977)', 4.214285714285714)]
```

In [21]:

```
ratedMovie1=ratedMovie.mapValues(lambda x:(x[0])/((1+x[1])))
movieJoin.join(ratedMovie1).map(lambda x: x[1]).filter(lambda x: x[1]>4).sortBy(lambda
x: -x[1]).take(10)
```

Out[21]:

```
[('"Shawshank Redemption, The (1994)"', 4.415094339622642),
 ('"Godfather, The (1972)"', 4.266839378238342),
 ('"Streetcar Named Desire, A (1951)"', 4.261904761904762),
 ('Fight Club (1999)', 4.2534246575342465),
 ('"Godfather: Part II, The (1974)"', 4.226923076923077),
 ('Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1
964)',
  4.224489795918367),
 ('"Three Billboards Outside Ebbing, Missouri (2017)"', 4.22222222222222
2),
 ('"Usual Suspects, The (1995)"', 4.217073170731707),
 ('Goodfellas (1990)', 4.216535433070866),
 ('Star Wars: Episode IV - A New Hope (1977)', 4.214285714285714)]
```

In [22]:

```
import math

ratedMovie1=ratedMovie.mapValues(lambda x:(x[0]/x[1])*math.log(x[1]))

movieJoin.join(ratedMovie1).map(lambda x: x[1]).filter(lambda x: x[1]>4).sortBy(lambda
x: -x[1]).take(10)
```

Out[22]:

```
[('"Shawshank Redemption, The (1994)"', 25.506303124680446),
 ('Forrest Gump (1994)', 24.135559630846686),
 ('Pulp Fiction (1994)', 24.035971735394476),
 ('"Matrix, The (1999)"', 23.593497870526754),
 ('"Silence of the Lambs, The (1991)"', 23.43310709209536),
 ('Star Wars: Episode IV - A New Hope (1977)', 23.37860964684444),
 ('Fight Club (1999)', 23.007601610036865),
 ("Schindler's List (1993)", 22.788076383338726),
 ('Star Wars: Episode V - The Empire Strikes Back (1980)', 22.561506207236
786),
 ('"Godfather, The (1972)"', 22.549726244087907)]
```

What are the names of movies seen by the userId=1?

In [23]:

```
moviesUser1 = ratings.filter(lambda x:  x[0]=="1").map(lambda x: (x[1],x[2])).join(movi
es).map(lambda x: (x[1][1],x[1][0])).sortBy(lambda x:x[0])
for m in moviesUser1.collect():
    print(m)
```

```
('"13th Warrior, The (1999)"', '4.0')
('"Abyss, The (1989)"', '4.0')
('"Adventures of Robin Hood, The (1938)"', '5.0')
('"American Tail, An (1986)"', '5.0')
('"Big Lebowski, The (1998)"', '5.0')
('"Black Cauldron, The (1985)"', '5.0')
('"Blues Brothers, The (1980)"', '5.0')
('"Clockwork Orange, A (1971)"', '5.0')
('"Dirty Dozen, The (1967)"', '5.0')
('"Few Good Men, A (1992)"', '4.0')
('"Fugitive, The (1993)"', '5.0')
('"Game, The (1997)"', '5.0')
('"Ghost and Mrs. Muir, The (1947)"', '4.0')
('"Ghost and the Darkness, The (1996)"', '5.0')
('"Good Morning, Vietnam (1987)"', '5.0')
('"Goonies, The (1985)"', '5.0')
('"Great Mouse Detective, The (1986)"', '5.0')
('"Green Mile, The (1999)"', '5.0')
('"Honey, I Shrunk the Kids (1989)"', '4.0')
('"Iron Giant, The (1999)"', '5.0')
('"Jungle Book, The (1967)"', '5.0')
('"Jungle Book, The (1994)"', '5.0')
('"Last of the Mohicans, The (1992)"', '3.0')
('"Lock, Stock & Two Smoking Barrels (1998)"', '5.0')
('"Longest Day, The (1962)"', '4.0')
('"Lord of the Rings, The (1978)"', '5.0')
('"Man with the Golden Gun, The (1974)"', '4.0')
('"Mask, The (1994)"', '4.0')
('"Matrix, The (1999)"', '5.0')
('"Messenger: The Story of Joan of Arc, The (1999)"', '5.0')
('"Mummy, The (1999)"', '2.0')
('"Negotiator, The (1998)"', '5.0')
('"NeverEnding Story, The (1984)"', '5.0')
('"Newton Boys, The (1998)"', '5.0')
('"Nosferatu (Nosferatu, eine Symphonie des Grauens) (1922)"', '4.0')
('"Princess Bride, The (1987)"', '5.0')
('"Quiet Man, The (1952)"', '5.0')
('"Rescuers, The (1977)"', '5.0')
('"Road Warrior, The (Mad Max 2) (1981)"', '5.0')
('"Rock, The (1996)"', '4.0')
('"Rocketeer, The (1991)"', '5.0')
('"Rocky Horror Picture Show, The (1975)"', '3.0')
('"Secret of NIMH, The (1982)"', '5.0')
('"Shining, The (1980)"', '3.0')
('"Silence of the Lambs, The (1991)"', '4.0')
('"South Park: Bigger, Longer and Uncut (1999)"', '5.0')
('"Sword in the Stone, The (1963)"', '5.0')
('"Talented Mr. Ripley, The (1999)"', '1.0')
('"Terminator, The (1984)"', '5.0')
('"Texas Chainsaw Massacre, The (1974)"', '5.0')
('"Thin Red Line, The (1998)"', '5.0')
('"Three Caballeros, The (1945)"', '5.0')
('"Three Musketeers, The (1993)"', '4.0')
('"Usual Suspects, The (1995)"', '5.0')
('"Wedding Singer, The (1998)"', '4.0')
('"Wizard of Oz, The (1939)"', '5.0')
('"Wolf Man, The (1941)"', '5.0')
('20 Dates (1998)', '4.0')
('Alice in Wonderland (1951)', '5.0')
('Alien (1979)', '4.0')
('All Quiet on the Western Front (1930)', '5.0')
```

```
('American Beauty (1999)', '5.0')
('American History X (1998)', '5.0')
('Apocalypse Now (1979)', '4.0')
('Austin Powers: International Man of Mystery (1997)', '5.0')
('Back to the Future (1985)', '5.0')
('Back to the Future Part III (1990)', '4.0')
('Bambi (1942)', '5.0')
('Basic Instinct (1992)', '5.0')
('Batman (1989)', '4.0')
('Batman Returns (1992)', '3.0')
('Bedknobs and Broomsticks (1971)', '5.0')
('Beetlejuice (1988)', '4.0')
('Being John Malkovich (1999)', '4.0')
('Best Men (1997)', '4.0')
('Big (1988)', '4.0')
('Big Trouble in Little China (1986)', '4.0')
('Billy Madison (1995)', '5.0')
('Blazing Saddles (1974)', '5.0')
('Blown Away (1994)', '3.0')
('Bottle Rocket (1996)', '5.0')
('Braveheart (1995)', '4.0')
('Canadian Bacon (1995)', '5.0')
("Charlotte's Web (1973)", '5.0')
('Citizen Kane (1941)', '5.0')
('Clear and Present Danger (1994)', '4.0')
('Clerks (1994)', '3.0')
('Con Air (1997)', '4.0')
('Conan the Barbarian (1982)', '5.0')
('Crocodile Dundee (1986)', '5.0')
('Dances with Wolves (1990)', '4.0')
('Dazed and Confused (1993)', '4.0')
('Desperado (1995)', '5.0')
('Dick Tracy (1990)', '4.0')
('Dogma (1999)', '5.0')
('Dr. No (1962)', '5.0')
('Dracula (1931)', '4.0')
('Duck Soup (1933)', '5.0')
('Dumb & Dumber (Dumb and Dumber) (1994)', '5.0')
('Dumbo (1941)', '5.0')
('E.T. the Extra-Terrestrial (1982)', '5.0')
('Easy Rider (1969)', '4.0')
('Ed Wood (1994)', '4.0')
('Edward Scissorhands (1990)', '5.0')
('Encino Man (1992)', '3.0')
('Enemy of the State (1998)', '5.0')
('Escape to Witch Mountain (1975)', '3.0')
('Excalibur (1981)', '5.0')
('Face/Off (1997)', '5.0')
('Fantasia (1940)', '5.0')
('Fargo (1996)', '5.0')
('Fight Club (1999)', '5.0')
('Flight of the Navigator (1986)', '4.0')
('Forrest Gump (1994)', '4.0')
('Frankenstein (1931)', '4.0')
('From Dusk Till Dawn (1996)', '3.0')
('From Russia with Love (1963)', '5.0')
('Full Metal Jacket (1987)', '5.0')
('Ghostbusters (a.k.a. Ghost Busters) (1984)', '5.0')
('Gladiator (2000)', '5.0')
('Go (1999)', '5.0')
('Goldfinger (1964)', '5.0')
```

```
('Goodfellas (1990)', '5.0')
('Grosse Pointe Blank (1997)', '4.0')
('Groundhog Day (1993)', '4.0')
('Grumpier Old Men (1995)', '4.0')
('Grumpy Old Men (1993)', '5.0')
("Gulliver's Travels (1939)", '5.0')
('Heat (1995)', '4.0')
('Henry V (1989)', '5.0')
('Highlander (1986)', '5.0')
('Hook (1991)', '4.0')
('Howard the Duck (1986)', '4.0')
('I Know What You Did Last Summer (1997)', '3.0')
('I Still Know What You Did Last Summer (1998)', '2.0')
('Independence Day (a.k.a. ID4) (1996)', '3.0')
('Indiana Jones and the Last Crusade (1989)', '5.0')
('Indiana Jones and the Temple of Doom (1984)', '5.0')
('JFK (1991)', '5.0')
('James and the Giant Peach (1996)', '5.0')
('Jurassic Park (1993)', '4.0')
('King Kong (1933)', '4.0')
('Kiss the Girls (1997)', '4.0')
('L.A. Confidential (1997)', '5.0')
('Labyrinth (1986)', '4.0')
('Ladyhawke (1985)', '4.0')
('Legend (1985)', '4.0')
('Lethal Weapon (1987)', '4.0')
('Live and Let Die (1973)', '5.0')
("Logan's Run (1976)", '3.0')
('M*A*S*H (a.k.a. MASH) (1970)', '5.0')
('Mad Max (1979)', '5.0')
("McHale's Navy (1997)", '3.0')
('Men in Black (a.k.a. MIB) (1997)', '3.0')
('Mission: Impossible (1996)', '3.0')
('Monty Python and the Holy Grail (1975)', '5.0')
("Monty Python's Life of Brian (1979)", '5.0')
('Mr. Smith Goes to Washington (1939)', '5.0')
('Mrs. Doubtfire (1993)', '3.0')
('Office Space (1999)', '5.0')
("Pete's Dragon (1977)", '3.0')
('Pink Floyd: The Wall (1982)', '5.0')
('Pinocchio (1940)', '5.0')
('Planet of the Apes (1968)', '5.0')
('Platoon (1986)', '4.0')
('Predator (1987)', '4.0')
('Psycho (1960)', '2.0')
('Psycho (1998)', '2.0')
('Pulp Fiction (1994)', '3.0')
('Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark)
(1981)', '5.0')
('Red Dawn (1984)', '5.0')
('Reservoir Dogs (1992)', '5.0')
('Return to Oz (1985)', '3.0')
('Road Trip (2000)', '4.0')
('Rob Roy (1995)', '5.0')
('Robin Hood (1973)', '5.0')
('RoboCop (1987)', '4.0')
('Rocky (1976)', '5.0')
('Romancing the Stone (1984)', '4.0')
('Run Lola Run (Lola rennt) (1998)', '5.0')
('Rush Hour (1998)', '4.0')
('Rushmore (1998)', '5.0')
```

```
('SLC Punk! (1998)', '5.0')
('Saving Private Ryan (1998)', '4.0')
("Schindler's List (1993)", '5.0')
('Scream 3 (2000)', '5.0')
('Seven (a.k.a. Se7en) (1995)', '5.0')
('Shaft (1971)', '5.0')
('Shaft (2000)', '4.0')
("She's the One (1996)", '4.0')
('Sister Act (1992)', '3.0')
('Sleeping Beauty (1959)', '4.0')
('Small Soldiers (1998)', '4.0')
('Sneakers (1992)', '3.0')
('So I Married an Axe Murderer (1993)', '4.0')
('Song of the South (1946)', '4.0')
('Space Jam (1996)', '3.0')
('Spaceballs (1987)', '5.0')
('Star Wars: Episode I - The Phantom Menace (1999)', '4.0')
('Star Wars: Episode IV - A New Hope (1977)', '5.0')
('Star Wars: Episode V - The Empire Strikes Back (1980)', '5.0')
('Star Wars: Episode VI - Return of the Jedi (1983)', '5.0')
('Stargate (1994)', '3.0')
('Starship Troopers (1997)', '3.0')
('Superman (1978)', '4.0')
('Superman II (1980)', '5.0')
('Swingers (1996)', '4.0')
('Teenage Mutant Ninja Turtles II: The Secret of the Ooze (1991)', '4.0')
('Teenage Mutant Ninja Turtles III (1993)', '4.0')
('That Thing You Do! (1996)', '4.0')
('Thunderball (1965)', '5.0')
('Tombstone (1993)', '5.0')
('Tommy Boy (1995)', '5.0')
('Total Recall (1990)', '4.0')
('Toy Story (1995)', '4.0')
('Toys (1992)', '2.0')
('Transformers: The Movie (1986)', '4.0')
('Tron (1982)', '4.0')
('Twister (1996)', '3.0')
('Very Bad Things (1998)', '5.0')
("Wayne's World (1992)", '5.0')
('Welcome to Woop-Woop (1997)', '4.0')
('What About Bob? (1991)', '4.0')
('Who Framed Roger Rabbit? (1988)', '5.0')
('Wild Things (1998)', '4.0')
('Willow (1988)', '4.0')
('Willy Wonka & the Chocolate Factory (1971)', '5.0')
('Winnie the Pooh and the Blustery Day (1968)', '5.0')
('X-Men (2000)', '5.0')
('Young Frankenstein (1974)', '5.0')
('Young Sherlock Holmes (1985)', '3.0')
('¡Three Amigos! (1986)', '4.0')
```

Compute the similarity coefficient between userId=1 and all other users. What is the similarity between userId=1 and userId=1? userId=2? userId=3? userId=4?

In [144]:

```
def avg(dic, items):
    s=0
    for i in items:
        s += dic[i]
    return s/len(items)

def safeDivide(x,y):
    if y==0:
        return 0
    return x/y


def simil(x,y):
    x = [e for e in x]
    y = [e for e in y]
    if x==y:
        return 0
    dX = dict(x)
    dY = dict(y)
    common = set(dX.keys()) & set(dY.keys())
    if len(common)<1:
        return 0
    avgX = avg(dX,common)
    avgY = avg(dY,common)
    varX = 0
    varY = 0
    res = 0
    for c in common:
        res += (dX[c]-avgX)*(dY[c]-avgY)
        varX += (dX[c]-avgX)**2
        varY += (dY[c]-avgY)**2
    if varX*varY == 0:
        return 0
    if res < 0:
        return 0
    #return res / (1+(varX*varY)**(0.5))
    return math.log(1+len(common))*res / ((varX*varY)**(0.5))

userRatings = ratings.map(lambda x: (x[0],(x[1],float(x[2])))).groupByKey()
user1 = userRatings.filter(lambda x:x[0]=='1').take(1)[0]
userSimil = userRatings.mapValues(lambda x: simil(x,user1[1]))

output = userSimil.filter(lambda x: x[0] in ['1','2','3','4']).collect() # (user, simi
l)
for (k,v) in output:
    print(str(k)+" -> "+str(v))
```

```
1 -> 0
4 -> 0.7962921953903238
2 -> 0
3 -> 0.16597864726681164
```

In [142]:

```
userSimil.map(lambda x:x[1]).reduce(lambda x,y: x+y) / userSimil.count()
```

Out[142]:

```
0.5171211853962832
```

Compute the 10 top movies recommended by collaborative filtering using pearson correlation for userId=1

In [147]:

```
userMovieRatings = ratings.map(lambda x: (x[0],(x[1],float(x[2])))) # (user, (movie,rating))
movieRatingSimil = userMovieRatings.join(userSimil).map(lambda x:x[1]) # (movie,rating),simil
moviePearsonWeight = movieRatingSimil.map(lambda x: (x[0][0],(x[0][1]*x[1],x[1]))) # movie, (rating*simil,simil)
```

In [163]:

```
moviePearson = moviePearsonWeight.reduceByKey(lambda x,y: (x[0]+y[0],x[1]+y[1])).mapValues(lambda x: safeDivide(x[0],0.5+x[1]))
moviePearson.sortBy(lambda x: -x[1]).take(20)
```

Out[163]:

```
[('318', 4.536635892289907),
 ('3030', 4.434424835574195),
 ('260', 4.431828042208414),
 ('1235', 4.429626031654389),
 ('1196', 4.425127195922486),
 ('3983', 4.406293660279343),
 ('527', 4.401320700492619),
 ('858', 4.397350914935063),
 ('1178', 4.3885014100181605),
 ('6442', 4.385843025893944),
 ('48516', 4.381036957849871),
 ('177593', 4.3803680316756655),
 ('1136', 4.369367624272602),
 ('912', 4.365948042565985),
 ('1208', 4.3450753526950185),
 ('750', 4.33420656292983),
 ('5618', 4.33370523670445),
 ('1217', 4.332198889360908),
 ('50', 4.326823154051749),
 ('1276', 4.323868264719015)]
```

In [160]:

```
res = movies.join(moviePearson).map(lambda x: x[1]).sortBy(lambda x: -x[1])
for i in res.take(10):
        print(str(i[1])+"\t"+str(i[0]))
```

```
23.475556149163417      Shawshank Redemption
22.57145093526896       Star Wars: Episode IV - A New Hope (1977)
22.09606349140326       Matrix
21.866642921345733      Forrest Gump (1994)
21.84896403920532       Star Wars: Episode V - The Empire Strikes Back (19
80)
21.417309716974163      Schindler's List (1993)
21.2499022130862        Fight Club (1999)
20.97659759622866       Pulp Fiction (1994)
20.9350672875493        Raiders of the Lost Ark (Indiana Jones and the Rai
ders of the Lost Ark) (1981)
20.772169377657566      Silence of the Lambs
```

Remove from the previous list the movies already rated (and thus seen) by userId=1

In [161]:

```
res2 = res.subtractByKey(moviesUser1.map(lambda x: (x,'#'))).sortBy(lambda x: (-x[1],x[
0]))
for i in res2.take(10):
        print(str(i[1])+"\t"+str(i[0]))
```

```
23.475556149163417      Shawshank Redemption
22.57145093526896       Star Wars: Episode IV - A New Hope (1977)
22.09606349140326       Matrix
21.866642921345733      Forrest Gump (1994)
21.84896403920532       Star Wars: Episode V - The Empire Strikes Back (19
80)
21.417309716974163      Schindler's List (1993)
21.2499022130862        Fight Club (1999)
20.97659759622866       Pulp Fiction (1994)
20.9350672875493        Raiders of the Lost Ark (Indiana Jones and the Rai
ders of the Lost Ark) (1981)
20.772169377657566      Silence of the Lambs
```

In [162]:

```
res2 = res.join(moviesUser1).sortBy(lambda x: -float(x[1][0]))
for i in res2.collect():
        print(str(i[1][0])+"\t"+str(i[1][1])+"\t"+str(i[0]))
```

```
22.57145093526896       5.0     Star Wars: Episode IV - A New Hope (1977)
22.09606349140326       5.0     Matrix
21.866642921345733      4.0     Forrest Gump (1994)
21.84896403920532       5.0     Star Wars: Episode V - The Empire Strikes
Back (1980)
21.417309716974163      5.0     Schindler's List (1993)
21.2499022130862        5.0     Fight Club (1999)
20.97659759622866       3.0     Pulp Fiction (1994)
20.9350672875493        5.0     Raiders of the Lost Ark (Indiana Jones and
the Raiders of the Lost Ark) (1981)
20.772169377657566      4.0     Silence of the Lambs
20.753782793532604      5.0     Star Wars: Episode VI - Return of the Jedi
(1983)
20.732622320665648      5.0     American Beauty (1999)
20.66617921688854       5.0     Usual Suspects
19.992327260337227      4.0     Braveheart (1995)
19.971943115919796      4.0     Saving Private Ryan (1998)
19.921027867222936      5.0     Seven (a.k.a. Se7en) (1995)
19.814582365898467      5.0     Monty Python and the Holy Grail (1975)
19.734517283099244      5.0     Fargo (1996)
19.5813158813971        5.0     Back to the Future (1985)
19.545677539334406      5.0     Princess Bride
19.319573928245074      5.0     Fugitive
18.806273697711045      5.0     American History X (1998)
18.715873678092297      4.0     Toy Story (1995)
18.624601694682728      5.0     Gladiator (2000)
18.592748905869428      5.0     Indiana Jones and the Last Crusade (1989)
18.5260176307388        4.0     Jurassic Park (1993)
18.51688239263711       4.0     Apocalypse Now (1979)
18.39923165569747       5.0     Reservoir Dogs (1992)
18.345141480680386      5.0     Green Mile
18.330881850743623      5.0     Goodfellas (1990)
18.217321207004932      4.0     Groundhog Day (1993)
18.08148830943172       5.0     Clockwork Orange
17.94117532110751       5.0     L.A. Confidential (1997)
17.732343690229456      5.0     Willy Wonka & the Chocolate Factory (1971)
17.50032574289912       5.0     Terminator
17.48711690840617       4.0     Alien (1979)
17.47189535515581       5.0     Office Space (1999)
17.43413852165866       5.0     Big Lebowski
17.245205102345352      5.0     E.T. the Extra-Terrestrial (1982)
17.240290342231035      5.0     Full Metal Jacket (1987)
17.189069994488086      5.0     Monty Python's Life of Brian (1979)
17.084523844107313      5.0     Ghostbusters (a.k.a. Ghost Busters) (1984)
17.000585956556904      4.0     Dances with Wolves (1990)
16.72548672379116       4.0     Being John Malkovich (1999)
16.559667140760222      5.0     X-Men (2000)
16.42107795867614       4.0     Batman (1989)
16.38690345146777       5.0     Citizen Kane (1941)
16.357943754821477      3.0     Shining
16.34613835097902       5.0     Blues Brothers
16.233429836919523      5.0     Run Lola Run (Lola rennt) (1998)
16.097471161719472      5.0     South Park: Bigger
15.859285887610264      5.0     Indiana Jones and the Temple of Doom (198
4)
15.85452092799789       5.0     Who Framed Roger Rabbit? (1988)
15.77844393169248       5.0     Austin Powers: International Man of Myster
y (1997)
15.77477489800321       5.0     Young Frankenstein (1974)
15.672955249081001      5.0     Lock
15.658993773665769      5.0     Wizard of Oz
```

```
15.601305206489993      4.0     Total Recall (1990)
15.545058238674546      4.0     Big (1988)
15.512891524006916      4.0     Heat (1995)
15.481515267644092      3.0     Men in Black (a.k.a. MIB) (1997)
15.404858989643383      3.0     Mission: Impossible (1996)
15.32677097289503       4.0     Platoon (1986)
15.288477352959195      3.0     Clerks (1994)
15.275554288192597      3.0     Independence Day (a.k.a. ID4) (1996)
15.257609572355772      5.0     Edward Scissorhands (1990)
15.210320556613544      4.0     Rock
15.052847903338318      5.0     Rocky (1976)
14.945728519922215      5.0     Dogma (1999)
14.838747261721783      5.0     Blazing Saddles (1974)
14.829751170056346      2.0     Psycho (1960)
14.72991282594486       4.0     Lethal Weapon (1987)
14.686408497756782      4.0     Clear and Present Danger (1994)
14.516336878229891      5.0     Jungle Book
14.516336878229891      5.0     Jungle Book
14.46866998990897       4.0     Beetlejuice (1988)
14.444240038292175      5.0     Game
14.395961642806816      3.0     Mrs. Doubtfire (1993)
14.360379954821406      4.0     Back to the Future Part III (1990)
14.281738464090665      4.0     Star Wars: Episode I - The Phantom Menace
(1999)
14.264896899388939      5.0     Rushmore (1998)
14.16515207474294       4.0     Labyrinth (1986)
14.094149867812483      5.0     Highlander (1986)
14.067745917168931      5.0     Planet of the Apes (1968)
14.066375970233658      4.0     Few Good Men
13.994174818261982      4.0     Mask
13.940659132253213      5.0     Dumb & Dumber (Dumb and Dumber) (1994)
13.926857428009434      4.0     Dazed and Confused (1993)
13.902627985646516      5.0     Goonies
13.853450893868999      5.0     Wayne's World (1992)
13.852382481377495      5.0     M*A*S*H (a.k.a. MASH) (1970)
13.715675936218764      3.0     Stargate (1994)
13.69564973596506       5.0     Tombstone (1993)
13.644767186751215      5.0     Goldfinger (1964)
13.604741651498 5.0     Fantasia (1940)
13.600803475618015      4.0     Superman (1978)
13.589717808586252      5.0     Spaceballs (1987)
13.582631915104695      4.0     Ed Wood (1994)
13.517526817051262      4.0     Grosse Pointe Blank (1997)
13.468729621845144      5.0     Iron Giant
13.414312472843326      5.0     Tommy Boy (1995)
13.364297028670714      5.0     Pinocchio (1940)
13.338096131701281      5.0     James and the Giant Peach (1996)
13.338067539878805      4.0     Wedding Singer
13.194275246103528      5.0     Good Morning
13.144077790087891      5.0     Face/Off (1997)
13.00530874835503       4.0     Abyss
12.941470358906876      4.0     RoboCop (1987)
12.865475609496087      5.0     Go (1999)
12.847603739196556      5.0     Desperado (1995)
12.831928893347328      5.0     Enemy of the State (1998)
12.808541471192425      4.0     Predator (1987)
12.692839212774315      4.0     Romancing the Stone (1984)
12.645884771357757      3.0     Twister (1996)
12.57576972473941       3.0     Rocky Horror Picture Show
12.474899628870626      5.0     Road Warrior
12.45192727559246       3.0     Sneakers (1992)
```

12.378900808597734	4.0	Swingers (1996)
12.371798159082013	5.0	From Russia with Love (1963)
12.358162233200977	3.0	Last of the Mohicans
12.353399225960201	5.0	Rob Roy (1995)
12.319012659114565	5.0	Billy Madison (1995)
12.302198114673661	4.0	So I Married an Axe Murderer (1993)
12.287964290406592	3.0	Starship Troopers (1997)
12.280769425151663	5.0	Alice in Wonderland (1951)
12.233662473076352	4.0	Willow (1988)
12.179779718168968	5.0	Dr. No (1962)
12.170688735565184	5.0	Robin Hood (1973)
12.049892716971753	5.0	Dumbo (1941)
12.003242092284307	5.0	Dirty Dozen
11.970154460195406	4.0	Tron (1982)
11.940337664302035	4.0	Hook (1991)
11.931718231624686	5.0	Bambi (1942)
11.90868373189522	5.0	Crocodile Dundee (1986)
11.847212495666131	5.0	NeverEnding Story
11.755631004420012	3.0	Batman Returns (1992)
11.743847918462128	4.0	Big Trouble in Little China (1986)
11.56767271668213	5.0	Duck Soup (1933)
11.560306402822839	5.0	Live and Let Die (1973)
11.52955027596778	5.0	Bottle Rocket (1996)
11.496721227171742	4.0	Grumpier Old Men (1995)
11.472921415813806	5.0	Grumpy Old Men (1993)
11.460846083192338	4.0	Three Musketeers
11.44083616259382	5.0	JFK (1991)
11.430913107163068	2.0	Mummy
11.28934247388463	5.0	Rocketeer
11.26497677907009	5.0	Mad Max (1979)
11.229047147177441	5.0	Sword in the Stone
11.219681169697724	5.0	Basic Instinct (1992)
11.20179032831612	4.0	What About Bob? (1991)
11.176785035910328	4.0	Honey
11.019230526877283	4.0	Con Air (1997)
11.009969860037039	5.0	Pink Floyd: The Wall (1982)
10.995333566456015	5.0	Superman II (1980)
10.990140280174014	5.0	Jungle Book
10.990140280174014	5.0	Jungle Book
10.972560006518359	5.0	Thin Red Line
10.91143014847191	4.0	Rush Hour (1998)
10.887326458985164	3.0	From Dusk Till Dawn (1996)
10.867430921463262	4.0	That Thing You Do! (1996)
10.773092707585512	5.0	Excalibur (1981)
10.643771263369013	5.0	Charlotte's Web (1973)
10.52978933830124	4.0	Flight of the Navigator (1986)
10.46017225724893	1.0	Talented Mr. Ripley
10.41459193196214	4.0	Easy Rider (1969)
10.331886496724463	4.0	Road Trip (2000)
10.189845406147379	5.0	Negotiator
10.10914125238206	5.0	Bedknobs and Broomsticks (1971)
10.045122868217836	5.0	Thunderball (1965)
10.031774113970792	4.0	King Kong (1933)
9.967078376775763	5.0	Secret of NIMH
9.627595054842919	5.0	Henry V (1989)
9.612880795311344	5.0	Very Bad Things (1998)
9.574991583207359	4.0	Ladyhawke (1985)
9.556013502174144	4.0	¡Three Amigos! (1986)
9.496558144353452	5.0	American Tail
9.443749251088297	5.0	Conan the Barbarian (1982)
9.339316638989118	5.0	Mr. Smith Goes to Washington (1939)

9.327137967333552	4.0	Wild Things (1998)
9.193282420819534	4.0	Dick Tracy (1990)
9.170806481825316	4.0	Man with the Golden Gun
9.159341289928417	3.0	Sister Act (1992)
8.961366234670683	4.0	Sleeping Beauty (1959)
8.908575844497749	5.0	Winnie the Pooh and the Blustery Day (196 8)
8.896381211008324	5.0	Red Dawn (1984)
8.829746406459819	3.0	Space Jam (1996)
8.673425572858568	3.0	Young Sherlock Holmes (1985)
8.588900731198393	4.0	Longest Day
8.531536804670427	5.0	SLC Punk! (1998)
8.516482858451214	5.0	Messenger: The Story of Joan of Arc
8.515575010043596	3.0	Logan's Run (1976)
8.445059317958556	4.0	Nosferatu (Nosferatu
8.402908142279324	5.0	Lord of the Rings
8.184698547300528	5.0	Rescuers
8.135740129613529	5.0	Ghost and the Darkness
8.099111047668293	4.0	Legend (1985)
7.977434412836571	3.0	Pete's Dragon (1977)
7.699198976165826	5.0	Scream 3 (2000)
7.645139542532953	4.0	13th Warrior
7.410632107801193	5.0	All Quiet on the Western Front (1930)
7.388082318845258	5.0	Black Cauldron
7.331199338867387	4.0	Small Soldiers (1998)
7.169836773535172	5.0	Adventures of Robin Hood
7.0064631940405455	2.0	Toys (1992)
7.001973343762652	4.0	Teenage Mutant Ninja Turtles II: The Secre t of the Ooze (1991)
6.997453962390998	4.0	Kiss the Girls (1997)
6.968683739712381	3.0	I Know What You Did Last Summer (1997)
6.637189662152352	5.0	Great Mouse Detective
6.611006250085478	4.0	Transformers: The Movie (1986)
6.58105241218076	4.0	Ghost and Mrs. Muir
6.5218772158356915	3.0	Encino Man (1992)
6.462719921496879	4.0	She's the One (1996)
6.191548802253656	3.0	Return to Oz (1985)
6.114415858340263	4.0	Shaft (2000)
5.9183436022731595	3.0	Blown Away (1994)
5.907801101270121	3.0	Escape to Witch Mountain (1975)
5.864030849936576	5.0	Quiet Man
5.8373792559534365	5.0	Canadian Bacon (1995)
5.622834911247529	4.0	Dracula (1931)
5.606487056269216	4.0	Frankenstein (1931)
5.569720960615982	4.0	Honey
5.366597539732538	4.0	Teenage Mutant Ninja Turtles III (1993)
5.362014854187082	5.0	Texas Chainsaw Massacre
5.334362649443451	4.0	Howard the Duck (1986)
5.319223308057846	2.0	I Still Know What You Did Last Summer (199 8)
5.111450066767953	4.0	Song of the South (1946)
5.068910598867106	5.0	Three Caballeros
4.83107458654831	2.0	Mummy
4.616664594331109	5.0	Shaft (1971)
4.407829652992515	4.0	20 Dates (1998)
4.113548891374922	2.0	Psycho (1998)
4.100364220626212	3.0	McHale's Navy (1997)
3.5825903285268432	5.0	Wolf Man
3.31276082651236	4.0	Three Musketeers
3.1754635241856777	4.0	Three Musketeers
2.9189747653141804	5.0	Newton Boys

```
2.7486682554468014      5.0     Texas Chainsaw Massacre
1.7289419164528208      4.0     Welcome to Woop-Woop (1997)
1.6695539642413681      2.0     Mummy
1.5003706664941987      5.0     Fugitive
0.9045076583874632      4.0     Three Musketeers
0.4769283544537032      4.0     Honey
0.1851173006592013      5.0     Gulliver's Travels (1939)
0.0     4.0     Best Men (1997)
```

In [ ]: